FOREWORD

Consciousness is a scientific problem that is unlike any other. Our own consciousness, as Descartes noted, is the most indubitable feature of our existence. It is the most precious one, as well: consciousness is life itself, and for most people having their bodies kept alive in a vegetative state is no better than dying. The major religions are defined by their theories of consciousness: whether a person's essence consists of his consciousness (his soul) or his body; how that consciousness ultimately fares as the result of its choices in life (whether it goes to a special place, or melds into a global mind); and whether the world contains forms of pure, disembodied consciousness in the form of gods, demons, angels, and spirits. And the conviction that other people can suffer and flourish as each of us does is the essence of empathy and the foundation of morality.

This encyclopedia is thus a celebration of a momentous recent development in the history of science. Consciousness has not only become a respectable scientific topic, but it is one that has seen tremendous progress. And it enjoys a surprising degree of consensus for such an elusive subject. My hunch is that a majority of consciousness researchers would assent to the following propositions. Human consciousness is entirely a product of the physiological activity of the brain. Consciousness does not depend on language. Nor is it the same thing as self-conscious rumination. The vast majority of information-processing taking place in the brain is unconscious. The parts that are conscious have no privileged access to the unconscious parts, and are more likely to confabulate about how they work than to report those workings accurately. Consciousness is not a unitary process that is located in a single part of the brain or that has complete control of behavior. The neural correlates of consciousness are defined both over space (in terms of connections to the frontal lobes and other cortical areas) and over time (in terms of oscillatory patterns in cortico-thalamic loops). Consciousness is intimately tied to– perhaps the same thing as—attention, imagery, and working memory. The cognitive representations that can become conscious tend to be found at intermediate levels of processing hierarchies, corresponding neither to information in the sensorium nor to the most abstract categories and goals. Consciousness plays a critical role binding together distributed activity over the brain that corresponds to a single entity in the world, and its main computational function is to serve as a ‘‘blackboard’’ on which the output of one cognitive process can be made available to many others.

If I am correct that this is a rough emerging consensus among consciousness researchers, it would represent an extraordinary degree of progress. The modern science of consciousness is barely two decades old. It has progressed in the teeth of objections from behaviorist philosophers and psychologists that the very idea of consciousness is a category mistake, and has had to distance itself from a horde of obfuscators, mystics, kibitzers, and kooks.

For all that, it is possible that our knowledge about consciousness will never be complete. Consciousness researchers do not agree that we will inevitably have an intellectually satisfying explanation of the ‘‘hard problem’’ of consciousness. This term, from the philosopher David Chalmers, is, of course, something of an in-joke. There is nothing easy about the so-called ‘‘Easy Problem’’ of distinguishing conscious from unconscious mental computation, identifying its correlates in the brain, and explaining why it evolved.

But when it comes to the Hard Problem—why the conscious portion of brain activity consists of first-person, subjective experience rather than pure information processing of the kind that takes place in thermostats and calculators–no one knows what a solution might look like, or even whether it is a genuine problem in the first place.

I am among those who suspect that the hard problem is a genuine problem but one that we will never solve. It is an idea that goes back at least to Hume, and has also been embraced by the biologist Gunther Stent, the linguist Noam Chomsky, and most extensively, the philosopher Colin McGinn. The theory holds that our vertigo when pondering the Hard Problem is itself a quirk of our brains. The brain is a product of evolution, and just as animal brains have their limitations, we have ours. Our brains can't hold a hundred numbers in memory, can't visualize 7-dimensional space, and perhaps can't intuitively grasp why neural information-processing observed from the outside should give rise to subjective experience on the inside.

I have found that most scientists hate this theory (which McGinn calls "cognitive closure"), acknowledging that that the Hard Problem remains unsolved for now, but assuming that it will eventually succumb to research that chips away at the Easy Problem. I have also found that most scientists don't understand the theory. I have been told that the cognitive closure theory tries to put the study of consciousness off-limits to science, a bizarre misinterpretation (the theory encourages the study of consciousness, while offering an empirical prediction of how it will turn out). I have been told that the theory is unfalsifiable. In fact the theory could be demolished if a genius came up with a flabbergasting new idea which made it all clear to us (in contrast, it is the conviction that some day in the future, we don't know when, we don't how, the problem will succumb, that is unfalsifiable). I have read that the theory is an endorsement of mystery over science, whereas in fact it seeks to explain of the phenomenon of mystery—what it is about the human brain that makes it susceptible to feelings of mystery. Moreover, it pinpoints the aspect of cognitive processing that is responsible for our mystification: the fact that our explanations exploit the combinatorial powers of the brain, which explains complex phenomena in terms of rule-governed interactions among simpler elements, and thus is frustrated when it runs across problems that have a holistic flavor, such as the hard problem of consciousness. And I have been told that research that characterizes consciousness (for example, by cross-modal similarity judgments) solves the hard problem, whereas such research just addresses the easy problem, documenting patterns of representation and information processing.

Whether or not my prediction withstands attempts at empirical falsification, it would matter little to the actual science of consciousness. The infuriating thing about the hard problem is that nothing hangs on it scientifically. As soon as you start to study it, you are studying one of its objective, third-person manifestations, which as far as we know can always be explained as information processing in a physical system. If we ever could trace all the neurocomputational steps from perception through reasoning and emotion to behavior, the only thing left missing by the lack of a theory of the hard problem of consciousness would be an understanding of the hard problem of consciousness itself. That is why, for all the unsatisfied intellectual curiosity we might have as to why subjective experience exists, the absence of an explanation takes nothing away from the exhilarating ideas and findings presented in this encyclopedia.

Steven Pinker
Harvard University

FOREWORD BIOGRAPHY

Steven Pinker is Harvard College Professor and Johnstone Family Professor of Psychology at Harvard University. He has also taught at Stanford and MIT. His research on visual cognition and the psychology of language has won prizes from the National Academy of Sciences, the Royal Institution of Great Britain, and the American Psychological Association. He has also received six honorary doctorates, several teaching awards, and numerous prizes for his books *The Language Instinct*, *How the Mind Works*, and *The Blank Slate*. He is currently Honorary President of the Canadian Psychological Association and Chair on the Usage Panel of the *American Heritage Dictionary*, and writes frequently for *The New Republic*, *The New York Times*, and other publications. He has been named Humanist of the Year, and is listed in *Foreign Policy* and *Prospect* magazine's "The World's Top 100 Public Intellectuals" and in *Time* magazine's "The 100 Most Influential People in the World Today." His latest book is *The Stuff of Thought: Language as a Window into Human Nature*.

PREFACE

The study of consciousness, like psychology, has a long past and a short history, but the history part is shorter and the past much longer than psychology's. Topics now included in the study of consciousness have been matters of fundamental existential human concern since prehistoric times. The earliest ceremonial burial sites, which suggest a belief in a sentient essence independent of the body, may be 70,000 years old. Only very recently has the approach to these issues been substantially different from those of the previous millennia. The difference between now and then is an empirical approach to consciousness and mental activity. This difference comes from the change in worldview that led to what we now call science and was a long time coming to the study of the mind.

The application of an empirical criterion for evidence brought consciousness into the domain of "normal science." Not everything from the older approach gets into normal science, however, only the few parts that follow the rules. The rules include no supernatural explanations and no untestable ones, of which the supernatural is a subset. The anti-supernatural rule constrains us to physical explanations, which are for the most part biological. The result is that some questions unthinkable in the prescientific tradition, such as neural correlates of consciousness, become important. Another previously unthinkable idea is that thought is a product of neural activity. (Imagine that! Reason founded on neurons?) Still another unthinkable idea is that some means is needed to maintain the coherence and continuity of experience. If the self is a spiritual, supernatural entity, continuity is in its nature. If it is a product of brain activity we need a neurophysiological explanation of how continuity and unity can come from the massively parallel operation of the brain. Articles by Schmidt, Vallacher, Bermudez, Olson, and Gallagher (Perception: The Binding Problem and the Coherence of Perception; Self: The Unity of Self, Self-Consistency; Self: Body Awareness and Self-Awareness; Self: Personal Identity and Consciousness of Time and the Time of Consciousness respectively) cover theories of the self. These are just a few of the previously unthinkable explanations needed in a science of consciousness.

One of the developments that puts the study of consciousness on a solid empirical footing is the array of neural imaging tools that let us see relations between awareness and brain activity. The images themselves are exciting because they give a glimpse of what the brain is doing when we are consciously seeing, remembering, thinking our own thoughts and so on. We can see changes in brain activity as the focus of attention switches from one image to another, for example. In addition to the scientific value of such findings, there is what could be called rhetorical value: the techniques drive home the point that what is going on in the mind is going on in the brain.

Articles by Haynes, Kouider, and Smythes, Goodale, and Sterzer (The Neural Basis of Perceptual Awareness; Neurobiological Theories of Consciousness; The Neurochemistry of Consciousness; Perception, Action, and Consciousness and Bistable Perception and Consciousness) and others in this collection report some of the scientific progress made through these tools. Several generalizations can be drawn from these contributions. One is that perception, and presumably other conscious processes, has a division of labor among many special-purpose analyzers. Some of these operate on sensory input and others on very high-level analysis such as face recognition. Their results are not by themselves conscious but if coordinated with other special-purpose analyzers can result in a conscious perception. The evidence that

sophisticated representations can be created unconsciously would support theories that postulate an active unconscious.

Control and coordination is one of the many processes subsumed under the rubric of attention. Change blindness and inattentional blindness illustrate the central importance of attention to consciousness. Change blindness is found when a scene or part of a scene is changed while the observer is not attending. The changes in the scene can literally be not seen at all if not attended. Rensink (Attention: Change Blindness and Inattentional Blindness) notes that such a gap in perception would have been classed as a momentary aberration before it had theoretical significance. Change blindness indicates that at a phenomenal level our impression of seeing the whole scene before us is not correct. The entirety of the scene may be available, but only what is attended will be perceived. Inattentional blindness is similar to change blindness, but it is not noticing something that is present but unattended in the visual field. These ''blindnesses'' both suggest that attention is a gateway to awareness.

Unattended objects may be analyzed to some depth outside of attention, as when emotionally laden words are seen in intentional blindness experiments. These words had to have been processed to the level of meaning without attention, or else they would not have been recognized. This finding is one of many that show analysis progressing quite far without attention or awareness. Breitmeyer (Perception: Unconscious Influences on Perceptual Interpretation) discusses in depth the relations between conscious and unconscious processes in perception. Snodgrass (Perception: Subliminal and Implicit) analyzes effects of stimuli that cannot be consciously represented, and covers many of the methodological problems in this area.

Processing of unattended material brings us to the larger question of what kind of processing can take place unconsciously. Consciousness is the tip of the iceberg of mental activity, the rest being unconscious, but the nature of the unconscious activity is a question. Some theorists assume that unconscious thinking is essentially like conscious thinking, just not conscious. Others, like Freud, theorized that unconscious mental activity goes on but it has different characteristics than conscious thinking has. It's an associative ''primary process'' rather than the rational thought of the conscious, waking mind. The articles by Kihlstrom (Unconscious Cognition) and Macmillan (Psychodynamic Theories of the Unconscious) discuss the active unconscious. Still others do not require an active unconscious but only action patterns or habits that can be automatically called into action. Approaches like this are covered in several articles, including those by Aarts (Habit, Action, and Consciousness) and Dijksterhuis (Unconscious Goals and Motivation). Schneider (Automaticity and Consciousness) contrasts automatic processing, which is unconscious, fast, and virtually effortless with controlled processing, which has properties that might classify it as conscious. Any or all of these versions of unconscious activity may take place in some situations.

Implicit cognition does much of our daily mental work, just as automatic processing spares us the effort of controlling skilled actions. Cleeremans (Implicit Learning and Implicit Memory) describes a wide range of studies that show how implicit memory can develop and influence performance, usually for the better, without conscious recall of the learning event or events. Bar-Anan and Nosek (Implicit Social Cognition) describe the Implicit Attitude Test (IAT) for measuring implicit attitudes that can influence opinions and social behavior but may be in contrast to admitted, explicit attitudes. Vallacher (Self: The Unity of Self, Self-Consistency) describes how the IAT is used to assess implicit self-esteem. Self-esteem measured by overt techniques such as questionnaires may differ from implicit measures. The result can be dysfunctional uncertainty about self-concept and conflicting and erratic behavior in different situations.

As Steven Pinker discusses in his foreword, scientific models of functions that have a conscious consequence run into what the philosopher David Chalmers termed the ''hard'' problem. For example, a complete model of color vision may predict such things as the outcome of color mixtures very well, but where in the explanation is the experience of the resulting color? We would like to have it explain not only the result, that a paper reflecting a certain spectral distribution is called ''red,'' but also why it has the qualitative look it has, the ''redness'' of red. There is a disjoint here. Scientific models have not explained subjective experience, and there seems no way to do it with the concepts they employ. Seager (History of Philosophical Theories of Consciousness) and Rowlands (The MindBody Problem) refer to the

problem of explaining experience with a scientific model as intractable. We seem to be attempting to connect two things, the wavelength distribution, which is a physical quantity, and the experience, which is known only to the observer. The models don't have a meaning for the expression, ''It looks red to me.'' It is though we have returned to the dualistic universe of material and spiritual (supernatural) substances. One can't hook a neuron to a feeling. They are different kinds of things.

The scientific enterprise goes about its business of explaining psychological phenomena without a worry about the intractability of the connection between explanation and experience. It is a problem seemingly at the center of the discipline that stands like an elephant in the room, yet science goes on as if it were not an issue. Perhaps it is enough to work only on the ''easy'' problems and ignore that old hard one. However, there are reasons to ask for some account of the experiential side of the process. The first is mentioned by Rowlands. It is unsatisfactory to leave this account out of the story. The perception of red is not explained to our satisfaction until the perception itself is explained – wasn't that the point, after all?

Another reason to have an answer about subjective states comes from the other hard problem, volition. How does the decision to act result in action? Again we seem to have two different things: the decision to act, which is an idea, and the action, some sort of motion of a group of muscles. How can an idea move a muscle? Here we do not have a passive sensation of red, we have an idea that seems actually to do something. It is as though the action is an effect of the idea and clear evidence that a connection is made between the idea and the material world. Articles by Pockett (Brain Basis of Voluntary Control), Hommel (Conscious and Unconscious Control of Spatial Action), Goodale (Perception, Action, and Consciousness), Jeannerod (Neuroscience of Volition and Action), and Custers (Memory: Procedural Memory, Skill, Perceptual-Motor Learning, and Awareness) constitute together a reference volume on volition and action. These accounts show that action is much more complex than the simple formula ''idea causes action.'' It is possible that our clear impression of a causal relation comes from an overly simplistic description.

Pinker leans toward one resolution of the hard problem in his forward. This is that there is a resolution, but our simian minds, evolved for other things, are simply not capable of finding it, or of understanding it if it were presented to us. I would strive for the negative capability that the poet John Keats admired. That is the ability to resist the urge to choose one option or another and remain in uncertainty. I should add that I have a suspicion that the problem may be formulated wrongly. The history of science shows many cases in which when a problem has gone unsolved for many years, and a number of very skilled problem-solvers have failed with it over decades or centuries, the problem has been improperly formulated, often because a wrong theory, implicit or explicit, has been applied.

Our language betrays an implicit dualism in many locutions: a thought becomes ''available'' to consciousness (as if consciousness were outside the system); a certain neural process ''reports'' to consciousness; attention brings something ''into'' consciousness. These metaphors can constrain in subtle ways. In the dictionary ''mental'' and ''physical'' are given as antonyms, and they lurk in my personal lexicon as such. Again, language tips us off to the implicit metaphor. I don't think I'm alone in having this problem. Francis Crick called the idea that consciousness is a product of brain events ''the astonishing hypothesis.'' It is astonishing because we accept it as a scientific matter but some small part of us still finds it an affront to common sense. Clearly, the old way of thinking is very deep and still may haunt us, so to speak. Perhaps a generation that grows up with results of the sort reported in this handbook will have an informed intuition that will not find the hypothesis astounding. Perhaps a member of that generation will come up with a formulation that will lead to a scientific answer.

As a final note, consciousness research has developed subdisciplines remarkably quickly for such a young field. These include analysis of attention, memory, the self, unconscious cognition, emotion, social perception, volition, psychopathology, and more. Just scan the table of contents of this encyclopedia to see the richness and variety of this work. It is as though researchers in many areas suddenly discovered that awareness was an important aspect of the topic they studied. There is more than just growth of interest in consciousness here. Awareness has become a central question in many areas. In order to understand perception, emotion, memory, volition, and many other

human functions we need to consider the role of awareness. The result is a transformation that ranks with the Copernican revolutions that have come at critical times in the history of science.

This encyclopedia depended on the good work of a number of people. I thank Nikki Levy, the Publisher, for conceiving of this project and getting it through the inevitable rough spots gracefully and effectively. Maria Turnock, Development Editor, for Major Reference Works, and Edward Taylor, as Project Manager, deserve thanks for the daily operation of this project and dealing with hundreds of academics. The Editorial Assistants that have tirelessly worked as support on this project are Victoria Sayce, Caroline Phipps and Milo Perkins. Finally, the editors, Bernard Baars, Mahzarin R. Banaji, Bruce Bridgeman, Shaun Gallagher, Geraint Rees, Jonathan Schooler, and Daniel Wegner, are responsible for assembling a stellar group of authors and obtaining the superb set of articles you see here.

William P. Banks

EDITOR-IN-CHIEF

William P. Banks received his BA, cum laude, from St. John's College, Annapolis, Maryland in 1964, and his MA in 1966 and PhD in 1968 from Johns Hopkins University. He took a post-doctoral fellowship at the Institute of Human Learning at the University of California, Berkeley, for the academic year 1968–1969.

He has been on the faculty at Pomona College since 1969, and was promoted to Professor in 1979, and in 2007 became the Edwin F. and Margaret Haynes Professor of Psychology. He was head of the psychology department 1974–1990 and again 2001–2004. His courses have included perception and cognition, consciousness, the psychology of evil, and a currently ongoing high-impact aerobics class that students have taken for PE credit every semester since 1985. He has twice received the Pomona college award for distinguished teaching.

Among some eighty publications of his are four co-authored books. His publications cover a variety of topics, including cognition of magnitude, Gestalt factors in perception, the psychophysics of alcohol intoxication, signal-detection theory and memory, the recovered memory debate, and of course consciousness. Support for his research has come from the National Institutes of Mental Health, the National Science Foundation, the Irvine Foundation, the Fetzer Foundation, and the National Institute on Alcohol Abuse and Alcoholism.

Between 1986 and 1992 he was a Project Director and co-PI under two Sloan New Liberal Arts (NLA) grants. The motivation of the NLA project was to counter what appeared to be a growing neo-Luddism among liberal arts students. His main contribution to the project was to develop a problem-solving program that gave students the opportunity to tackle real-world problems. Teams of students spent an academic year working on paid contracts with government and business. Thirty-five contracts were completed with clients as diverse as city governments, McDonnell Douglas, and NASA.

He has been Psychology Editor for the International Encyclopedia of Science and Technology since 1988. From 1984 to 1987 he was Associate Editor for the Journal of Experimental Psychology: Human Perception & Performance. He has been a member of the panel on Memory and Cognitive Processes,

National Science Foundation; the Air Force Office of Scientific Research, Cognitive Science Panel; the National Science Foundation Graduate Fellowship Selection Panel; and the American Psychological Foundation Teaching Award Committee. He is a Fellow of the Association for Psychological Science.

In 1990 he and Bernard Baars founded the journal *Consciousness and Cognition*. The journal has grown from 377 pages in the first volume, published in 1992, to 1400 pages in the most recent, volume 17. He began serving as Editor-in-Chief in 2005.

EDITORIAL BOARD

Bernard J. Baars is former Senior Research Fellow in Theoretical Neurobiology at the Neurosciences Institute in San Diego (www.nsi.edu), Adjunct Professor at the Institute for Intelligent Systems at the University of Memphis, and Distinguished Consulting Faculty member at Saybrook Graduate School in San Francisco. His PhD is in Cognitive Psychology from UCLA. He is interested in human language, the brain basis of consciousness, volition, and a variety of related topics including the history of scientific studies of consciousness and neuroethics. Baars pioneered a cognitive theory of consciousness called Global Workspace Theory, which is widely cited in philosophical and scientific sources. Together with William P. Banks, Baars has edited the journal Consciousness & Cognition for more than fifteen years. (From Academic Press/ Elsevier). He has written an introductory text for cognitive neuroscience, called Cognition, Brain & Consciousness: An Introduction to Cognitive Neuroscience (Baars & Gage, Eds. San Diego, Calif.: Elsevier/ Academic Press, 2007). Baars was founding President of the Association for the Scientific Study of Consciousness and has given presentations internationally (see www.nsi.edu/users/baars and http://bernardbaars.pbwiki.com, and http://en.wikipedia.org/wiki/Bernard_Baars).

Mahzarin R. Banaji was born and raised in India, in the town of Secunderabad, where she attended St. Ann's High School. Her B.A. is from Nizam College and her M.A. in Psychology from Osmania University in Hyderabad . She received her Ph.D. from Ohio State University (1986), was a postdoctoral fellow at the University of Washington, and taught at Yale University from 1986 until 2001 where she was Reuben Post Halleck Professor of Psychology. In 2002 she moved to Harvard University as Richard Clarke Cabot Professor of Social Ethics in the Department of Psychology and Carol K. Pforzheimer Professor at the Radcliffe Institute for Advanced Study.

Banaji is a Fellow of the American Association for the Advancement of Science, the American Psychological Association (Divisions 1, 3, 8 and 9), and the American Psychological Society. She served as Secretary of the APS, on the Board of Scientific Affairs of the APA, and on the Executive Committee of the Society of Experimental Social Psychology. She was elected fellow of the Society for Experimental Psychologists in 2005. Banaji has served as Associate Editor of Psychological Review and of the Journal of Experimental Social Psychology and is currently Co-Editor of Essays in Social Psychology. She serves on the editorial board of several journals, among them Psychological Science, Psychological Review, Journal of Personality and Social Psychology, and The DuBois Review. Her research has been funded by the National Science Foundation, the National Institute of Mental Health, and the Third Millennium Foundation.

Banaji was Director of Undergraduate Studies at Yale for several years, chaired APS's Task force on Dissemination of Psychological Science, and served on APA's Committee on the Conduct of Internet Research. Among her awards, she has received Yale's Lex Hixon Prize for Teaching Excellence, a James McKeen Cattell Fund Award, and fellowships from the Guggenheim Foundation and the Rockefeller Foundation. In 2000, her work with R. Bhaskar received the Gordon Allport Prize for Intergroup Relations. With Anthony Greenwald and Brian Nosek, she maintains an educational website that has accumulated over 3 million completed tasks measuring automatic attitudes and beliefs involving self, other individuals, and social groups. It can be reached at www.implicit.harvard.edu, and details of the research may be found at www.people.fas.harvard.edu/~banaji.

Banaji studies human thinking and feeling as it unfolds in social context. Her focus is primarily on mental systems that operate in implicit or unconscious mode. In particular, she is interested in the unconscious nature of assessments of self and other humans that reflect feelings and knowledge (often unintended) about their social group membership (e.g., age, race/ethnicity, gender, class). From such study of attitudes and beliefs of adults and children, she asks about the social consequences of unintended thought and feeling. Her work relies on cognitive/affective behavioral measures and neuroimaging (fMRI) with which she explores the implications of her work for theories of individual responsibility and social justice.

Bruce Bridgeman is a professor of psychology and psychobiology a the University of California, Santa Cruz. He holds a BA in psychology from Cornell University and a PhD in physiological psychology from Stanford University. He was a post-doctoral fellow at the Free University of Berlin with a Humboldt Foundation fellowship, and at the UC Berkeley School of Optometry with a NIH fellowship. His research has been supported by the National Institutes of Health, the National Science Foundation, the Air Force Office of Scientific Research, and NASA. For six years he was a Guest Professor at the Max-Planck Institute for Psychological Research, Munich, and has served on the advisory committees of two Max-Planck Institutes. He has been an invited faculty member at summer schools in Germany and Holland, and held research appointments at the Universities of Bielefeld, Germany, and Padua, Italy. His research concentrates on spatial aspects of vision - how information is coded and processed in the brain during activities such as pattern recognition and visually guided action. Professor Bridgeman is the author of *The Biology of Behavior and Mind* (Wiley) and *Psychology and Evolution: The Origins of Mind* (Sage), and more than 120 scientific journal publications, as well as book chapters and international conference contributions. His most recent academic honor is the Chemers award for excellence in research.

Shaun Gallagher is Professor and Chair of the Philosophy Department at the University of Central Florida and Research Professor of Philosophy and Cognitive Science at the University of Hertfordshire. He received his Ph.D. in philosophy from Bryn Mawr College, an MA in economics from the State University of New York, and an MA in philosophy from Villanova University. He has been an invited Visiting Professor at the Ecole Normale Supériure, Lyon (2007), occasional invited Visiting Professor at the University of Copenhagen (2004–06), and Visiting Scientist at the Medical Research Council's Cognition and Brain

Sciences Unit at Cambridge University (1994). His research interests include phenomenology and philosophy of mind, cognitive sciences, hermeneutics, theories of the self and personal identity. His books include *How the Body Shapes the Mind* (Oxford University Press, 2005); *The Inordinance of Time* (Northwestern, 1998); *Hermeneutics and Education* (SUNY, 1992) and, co-authored with Dan Zahavi, *The Phenomenological Mind: Introduction to Philosophy of Mind and the Cognitive Sciences* (Routledge, 2007). He is co-editor of the interdisciplinary journal *Phenomenology and the Cognitive Sciences*, and a recent co-editor of *Does Consciousness Cause Behavior? An Investigation of the Nature of Volition* (MIT Press, 2006).

Geraint Rees is a Wellcome Senior Clinical Fellow and Professor of Cognitive Neuroscience at University College London. Trained as a neurologist and experimental psychologist at University College London and the California Institute of Technology, his group uses functional neuroimaging to study the neural basis of perceptual awareness. He is an Associate Editor of *Brain* and has published extensively on the neural basis of human consciousness.

Jonathan W. Schooler Professor of psychology at the University of California at Santa Barbara, Jonathan W. Schooler pursues research on consciousness, memory, the relationship between language and thought, problem-solving, and decision-making. A cum laude graduate of New York's Hamilton college, Dr. Schooler earned a Ph.D. in psychology at the Univeresity of Washington in 1987. He joined the faculty of the University of Pittsburgh as assistant professor, and in 2004 accepted and held the position of full professor and Canada Research Chair in Social Cognitive Science until 2007 when he accepted his current position.

A fellow of the Association for Psychological Science, Dr. Schooler has been the recipient of three Akumal Scholar Awards from the Positive Psychology Network, an Osher Fellowship given by the Exploratorium Science Museum, and a Lilly Foundation Teaching Fellowship. His work has been

supported, among others, by the National Institute of Mental Health, the Office of Educational Research and Canada's Social Sciences and Humanities Research Council. He is currently on the editorial boards of *Consciousness and Cognition* and *Social Cognitive and Affective Neuroscience*. Dr. Schooler is the author or co-author of more than one hundred papers published in scientific journals and the editor (with J.C. Cohen) of *Scientific Approaches to Consciousness*.

Daniel M. Wegner is Professor of Psychology at Harvard University. A Ph.D. of Michigan State University (1974), he has held professorships at Trinity University in Texas and at the University of Virginia. Wegner has published over a hundred articles on the role of thought in self-control and in social life, pioneering the study of thought suppression (why we have trouble keeping unwanted thoughts out of mind), transactive memory (how we remember things cooperatively with others), and apparent mental causation (what gives us the sense that we are consciously causing our actions). Among Wegner's books are *The Illusion of Conscious Will* (2002) and *White Bears and Other Unwanted Thoughts* (1989). His research has been funded by the National Institute of Mental Health and by the National Science Foundation, and he has been a Fellow of the Center for Advanced Study in the Behavioral Sciences. A Fellow of the American Psychological Association and of the American Association for the Advancement of Science, he has served as associate editor of *Psychological Review*, and on the Board of Reviewing Editors for *Science*. His recent research focuses on how people perceive the minds of others.

PERMISSION ACKNOWLEDGMENTS

The following material is reproduced with kind permission of Nature Publishing Group

Figure 3 of General Anesthesia
Figure 3 of Perception, Action, and Consciousness
Figure 1 of Sleep: Implications for Theories of Dreaming and Consciousness

http://www.nature.com/nature

The following material is reproduced with kind permission of Oxford University Press

Figure 5 (e) of General Anesthesia
Figure 1 of Hypnosis and Suggestion

http://www.oup.co.uk

The following material is reproduced with kind permission of American Association for the Advancement of Science
Figure 4 of The Control of Mnemonic Awareness
Figure 5 of The Control of Mnemonic Awareness

http://www.sciencemag.org

Aesthetics and the Experience of Beauty

W Hirstein, Elmhurst College, Elmhurst, IL, USA
M Campbell, National University, La Jolla, CA, USA

Glossary

Neuroaesthetics – A research paradigm whose proponents attempt to learn more about the human aesthetic experience by studying the brain, in particular the perceptual system.

Introduction

What is required for aesthetic experience to take place? Who or what is capable of having aesthetic experience and of judging something to be beautiful? A wild wolf surveying the canyon floor for small moving prey is in an exquisitely sensitive perceptual state, but he is not having an aesthetic experience. Is aesthetic experience solely the province of humans, then? Is there no animal model of our appreciation of art? Or what about the nest of the bowerbird, discussed by Ernst Gombrich in his study of the psychology of decorative art? Gombrich compares the sorts of repetitive patterns and color arrays with no underlying meaning beyond their visual impact found in decorative art with products found in the animal world. He points to the example of the nest of the bowerbird, known for creating highly decorated nests made complex (and beautiful?) with shells, brightly colored objects such as berries, feathers, flowers, and even found items of human manufacture such as plastic or glass. Not only does the male bird spend many hours building, creating, and crafting the nest, but he will also replace moved or disturbed elements of a seemingly intentional design. Moreover, each bower built is completely original and reflects decisions made by its creator on the type, color, and structural array of objects collected for decorating the nest. Should we say that these birds are artists? Should we attribute the capacity of aesthetic judgment to the female bowerbirds who select their mate on the basis of the brilliance of his bower?

In the human realm, our aesthetic drive seems to predate history at the very least. Paleolithic people imprinted cave walls with figures, symbols, and what seem to be projected forms of internal images. It is not known whether the underground paintings at Chauvet cave or Lascaux, France, are truly works of art, products of an 'aesthetic impulse,' or just primitive attempts at magical wish-fulfillment, but they clearly arouse aesthetic responses in modern-day viewers who marvel at their beauty and elegance. Principled reflection on the concept of the aesthetic response and on its relation to cognition probably starts with the ancient Greek philosophers Plato and Aristotle, who each had something to say about aesthetics (from the Greek, *aisthanomai* and *aesthetikos*, perception or apprehension through the senses). Plato holds beauty itself in the highest regard, as an eternal aspect akin to virtue and truth. However, Plato has something quite different to say about the experience of beauty as effected through perceptual activity, and we get a sense of his displeasure by noting his censure of those who would attempt to play with our sensory intake and manipulate the quality of perceptual experience – artists – who deal in images, representations, and analogies, but not, as Plato would claim, unfiltered truth. Unlike Plato, Aristotle sees a place for the arts and the development of aesthetics as worthwhile imitations of life and its passions: The arts can provide an occasion for *katharsis*, a purging of emotions through vicarious experience.

The field of aesthetics comes into its own in the eighteenth century when Alexander Baumgarten, a German rationalist philosopher, resurrected the term 'aesthetics' from the earlier Greek usage to serve as the label for a philosophical discipline dealing with knowledge from sensory perception.

Another eighteenth-century thinker, Immanuel Kant, realized that sensory experience, in all its complex and concept-defying richness, requires its own methods of discovery and analysis. Turning his thinking to aesthetics, Kant distinguishes three types of aesthetic, or reflective, judgments based on feelings of pleasure or displeasure: judgments of the agreeable, judgments of the beautiful (or of taste), and judgments of the sublime. In deeming something agreeable, one is merely attributing to it a subjectively determined pleasure-producing quality. We do not expect others to find it so; we would call something agreeable 'beautiful' only in this limited and nonbinding sense. Notions of the sublime go beyond the concept of beauty: The sublime is awe-inspiring, fearsome, overwhelming. Turning to judgments of beauty, Kant mentions four aspects that are essential to such judgments: (1) disinterested pleasure (we find pleasure in the object because it is beautiful, not the other way around); (2) universal validity (but this universality is not based on any kind of rigid conceptual categorization); (3) something he calls 'purposeless purposiveness' (the result of 'free play' of the imagination); and (4) necessity (a subjective attribution of beauty exemplifies how the object ought to be judged).

Aesthetic Perception

Aesthetic judgments, although grounded in sensory cognition, differ in the role they play in human experience. Beyond enabling the attribution of properties such as shape, color, odor, texture, and so forth to particular things, aesthetic judgments allow us to ascribe meaning and value to experiences. Not only is a sample of Chateau Lafite wine pungently fragrant, purplish red, tart, and cool, a sip of it delivers a highly complex blend of aromas and flavors that can be named only by comparison to mundane substances: oak, cherry, smoke, chocolate. The experience of the connoisseur supervenes on the base elements of taste and smell, but there is more: He transforms these elements into something new through the prism of his cognitive stock – his knowledge that the wine is rare, outrageously expensive, and coveted by experts – along with his memories of having drunk other rare wines and his feelings of anticipation fueled by imagining the pleasure to come.

Qualities in nonhuman creatures that attract mates (or in some cases frighten away predators) may be compared with those qualities which, at first glance, comprise human beauty or attractiveness. But again, at least in the case of humans, the aesthetics of beauty require more than recognition of forms and sounds and perception of colors and odors. The appreciation of beauty involves conscious activity in which areas of the brain are engaged in the production of a complex mental state – a hybrid of sensory and cognitive components. Recent research on the facial characteristics we consider beautiful has found that symmetric faces are judged to be more beautiful than asymmetric ones. Of course, perfect faces can be boring, and introducing asymmetries or flaws can improve on perfection. We call a single dark mole on one side of a woman's mouth a 'beauty mark,' but correct the asymmetry by placing a matching mole on the opposite side, and we see the pair as a sort of disfigurement. The single mole reorganizes our perception of her face by accentuating certain features and drawing attention away from others. Our understanding of how we read emotions from faces is also progressing nicely, and much of this information is relevant to the artistic expression of emotion through the depiction of faces, such as the Mona Lisa's intriguing smile. By steadily understanding the processes behind our perception of emotion and attractiveness in faces, we can both come to understand those judgments and teach the artist the dimensions along which faces vary. Many visual artists, particularly painters of the Impressionist and post-Impressionist periods of the late eighteenth and early twentieth centuries, are interested in understanding and exploiting the workings of the eye and brain in perceiving light and constructing colors.

Aesthetics and Neurophysiology

The early findings of the Gestalt psychologists demonstrate that our visual systems will complete gaps in lines and corners, fill in colors, and so on in order to produce a more unified and coherent perceptual end-product. We have reached the point where we can provide neuroscientific

explanations of these Gestalt phenomena, something that seems to invite the possibility of our gaining a neuroscientific understanding of more complicated perceptual events such as those involving the perception and aesthetic appreciation of works of art. The new science of neuroaesthetics investigates perception and the way the brain generates aesthetic responses. How are forms and patterns processed? Are certain forms 'preferred' over others by the perceptual system?

Neuroaesthetics has produced two early theories of how the brain generates aesthetic responses. Semir Zeki has developed a theory of aesthetics that appeals to neurophysiological structures and events in the human brain, and he goes some distance beyond well-established claims that they are essential in expounding or expanding traditional ideas about painting as an art. Zeki asserts that artists, because they study the brain and the mechanisms of perception and the causal avenues for producing aesthetic responses, are themselves neurologists. Where a painter differs from a scientist who studies visual perception is in methodology, not in subject matter, according to Zeki. Viewing the work of visual artists from Zeki's perspective provides a deeper insight into why an artist does what he does in constructing a painting, why he chooses the exact shapes, colors, and specific juxtapositions of painterly elements such as line, shading, perspective, and so on. The artist is exploring, exploiting, challenging, and manipulating the brain's ability to absorb and interpret sensory data in novel, unexpected, and exciting ways. Therein lies the basis of the aesthetic response; not only does the spectator see shapes, lines, and colors of various tints and shades when he or she looks at a painting, but there is also the essential component of any successful work of art that is the production of a type of mental state that outstrips ordinary perception. Great, or even good, art will arouse subtle, or perhaps strong, emotions. It will jostle and arouse memory and effect numerous associations with other mental states. It can induce a sense of joy or of infinite sorrow; it may even lift the spectator to heights of spiritual or existential transcendence. And the artist accomplishes all this without electrodes, probes, or chemicals. The artist, in Zeki's view, is a virtuoso scientist using instruments much simpler and less invasive, but requiring great ingenuity and immense perspicacity in their application.

Zeki emphasizes the analogy between the way that our perceptual systems function to extract the more permanent patterns in the flow of energy reaching them with the way that artists often strive to capture the essence of objects they depict. Ramachandran argues further that the representational artist is presenting us with a sort of caricature of the represented object, in which certain perceived features have been exaggerated while other features have been deemphasized. A skilled artist can evoke a particular person or place with a few deftly drawn lines, using the art of caricature to capture a sort of 'formal essence' of the represented person or thing. What we are able to see through such caricatures depends, in scientifically describable ways, on the structure of our visual system and its connections to the different memory systems. Much of visual art can be captured by rules of form, many of which are derived from the findings of the Gestalt psychologists, such as symmetry, balance, and grouping of forms. Artists combine these basic form-primitives in new and evocative ways, according to Ramachandran.

Consciousness and Aesthetic Experience

What is the connection between consciousness and aesthetic experience? Are aesthetic experiences necessarily conscious, for instance? While it may be no easier to show the necessity of consciousness to aesthetic experience than to any other sort of experience or mental function, aesthetic experiences are often included in lists of paradigm conscious experiences. It is also interesting to note that several of the classic puzzles that theorists of consciousness are tasked to solve either make reference to aesthetic reactions themselves, or to their essential ingredients, such as the ability to perceive color, as in Frank Jackson's 'Mary case.' Mary is a future neuroscientist who knows about all of the physical processes involved in human color vision, but has never seen any colors other than black and white (she is kept in a special room, only given certain foods, etc.). When Mary is

finally allowed to see colors, she seems to learn something new, that is, what it is like to consciously experience red. But she already knew about all the physical events involved, and so there appears to be a problem for theorists who claim that the mind is physical, in that the conscious experience of red seems not to be among the physical events that we agreed initially that Mary knew. Another classic puzzle, Dennett's 'Chase and Sanborn problem,' involves something close to aesthetic perception – in this case, coffee tasting. Chase and Sanborn have been brewing their coffee for decades, but neither of them likes the taste any longer. Chase says that his taste preferences have not changed over time, but that the way the coffee tastes to him has changed. Sanborn says that the coffee tastes the same to him, but his preferences have changed so that he no longer enjoys that flavor. Dennett claims that no physicalist theory can capture the difference between Chase and Sanborn, because he believes that the taste and the preference for the taste cannot be teased apart. Delving further into the distinction between perception and preference, and whether and how it might be made, will also reveal information of interest about art. Indeed, to pose a final neuroaesthetic theory of the experience of art, the scientist may have to wait until these more basic ontological questions are answered.

But must aesthetic experiences be conscious? Could there be a zombie artist, for instance, a being who looks and acts just like a normal person, creating and appreciating works of art but not capable of consciousness? But, is not the whole point of creating works of art to produce those special conscious experiences? What is the point of his making these objects, we want to know. We might look more closely into our normal aesthetic experience for unconscious events that are at least important to our experiences, if not outright components of it. One candidate for an unconscious aesthetic response is the sort of case where, for example, one realizes some time after seeing a film an interesting subplot of which he had not been conscious aware on first seeing the film. In more complicated artworks, especially of the narrative variety such as the novel, how the author handles the main plot and its relations to various subplots is an important part of how we aesthetically evaluate such works. In cases like the film viewer who only later fully appreciates the work, the brain seems to be continuing the attempt to understand the artwork all on its own. Certainly many artists have reported being inspired by dreams, although it is debatable whether dreams themselves are a type of conscious state.

Consciousness relates to aesthetic experience at two basic points, which then can permute and recombine to produce more complex aesthetic experiences. The first point of contact concerns the artist's act of creation, including the intentions specifying how the artwork will be made and what the end result will (ideally) be like. The second point of contact concerns the conscious perception of the finished artwork by a human perceiver. One of the first recombinations of these basic ingredients occurs in the brain of this viewer when she considers the intentions of the artist. There is also the fact that many artists attempt to emulate the mind of a viewer, or often more specifically, viewers of different sorts: critics, peers, average people, etc. So the viewer is modeling the intentions of the artist, while the artist is modeling the understanding of the viewer. This sort of activity falls under the category of 'theory of mind' or 'mindreading'; in this case intentions and perceptual states are modeled. Another kind of reading we do of other people occurs when we watch their intentional activities while employing mirror neurons to understand their actions. Rizzollati, who is the first to attribute the familiar imitative behavior we observe not just in monkeys but in humans as well to these specialized 'mirror neurons,' notes that mirror neurons are at their most effective when the observer is face-to-face with the observed. Stafford suggests that this fact explains why so many visual artists focus on the face as the center of a work, sometimes presenting the viewer with nothing but a face. Here the artist is playing up the natural emphasis on frontality and the intuitive recognition that some of our most deep-seated reflections about who people are, about our own identity, arise from contemplating the face of the other, or of the face in the mirror.

Several neurological patients have recently been described who have become obsessed with art after brain lesions, or repeated epileptic seizures, very often in the temporal lobes. These obsessions typically focus on a particular medium, often painting

or music. The obsessions can compel the patient to spend every waking moment creating and performing art, and some of the patients have become quite accomplished and are even able to sell paintings for high prices, or impress music critics with their composition or performances. Apparently, these patients find thinking about and creating works of art extremely rewarding, so much so that they sometimes forget to eat while in the throes of creation. This suggests an explanation for aesthetic responses that relates them to consciousness in general. There is little point in assembling the expensive and complicated machinery needed to achieve consciousness if we do not trouble ourselves to use it. Perhaps because conscious thinking requires effort, there needs to be a reward for engaging in conscious thought. Nature's way of enticing us toward conscious activity, according to this view, is to make conscious experience itself rewarding. Contemplation of the contents of consciousness is not, as we might have thought, basically neutral, with any emotional response being traceable to the particular contents of that mental state and not to consciousness itself, but rather pleasurable in itself. Sometimes, naturally, the particular contents of consciousness are unpleasant enough to cancel out this basically positive reaction, as usually happens in the case of pain experience, for instance. One perhaps sees a similar 'enticement' mechanism at work in the case of sex; nature entices us to use our sexual organs so that the race proliferates by making their use highly rewarding. It seems clear that conscious sensations play a vital role in this case.

Conclusion

Once art is brought into the realm of science, several classical claims about it can be seen to be empirically testable. In his landmark work, *Art and Experience*, for example, John Dewey argues that aesthetic experience requires dissolution of self and object as separate existences. This stands in contrast with scientific observation, he notes, which calls for as much separation from subjectivity as possible in perceptual judgments. Empirical observation for the purposes of science idealizes a form of perception cleansed of the observer's expectations, personal desires, and preexisting beliefs. Any significant level of projection or interpretation skewed toward personal preferences, tastes, emotions, or remembrances is considered as tainting, or even invalidating, scientific observations. On the contrary, aesthetic experience depends, both ontologically and epistemologically, on such subjective contributions. Conscious or not, these sorts of differences should be quite measurable as differences in brain activity.

The hypothesis that specialized brain activity can be directly correlated with specific sorts of aesthetic responses received some confirmation by a recent experiment conducted by Kawabata and Zeki using functional magnetic resonance imaging (fMRI) to locate and quantify what they term 'neural correlates of beauty.' Prior to being scanned, their subjects classified paintings of different genres (abstract, still life, portrait, and landscape) as beautiful, neutral, or ugly. Subjects then viewed the same paintings while in the fMRI scanner. Distinct, specialized areas of the brain were found to be activated when perceiving different genres of painting. More interestingly, independent of the genre of the perceived painting, the orbitofrontal cortex was found to be differentially activated when the subject was perceiving beautiful or ugly stimuli, while the motor cortex was found to be activated differentially during perception of paintings as beautiful or ugly. Activity in the orbitofrontal cortex increased during the perception of beautiful stimuli and decreased in the case of ugly stimuli. The reverse was seen to be true of the motor cortex. Stimuli perceived as ugly produced the most activity while stimuli judged to be beautiful the least. That finding perhaps helps explain why aesthetic perception can sometimes have a curiously paralyzing effect on us, as Stendahl noted. Typically conscious experience induces one to act, whereas aesthetic experiences 'stop us in our tracks'; we end bodily motion to engage in deeper contact with the artwork. Aesthetic appreciation of this sort is not goal-oriented. If any action is motivated by aesthetic appreciation, it is to move closer to the art object, or at least to prolong the experience.

One often-heard objection, especially from the art establishment, is that neuroaesthetics is 'reductionist,' in the sense that its proponents intend to

claim that the aesthetic experiences we are aware of are somehow unreal or unimportant, whereas the underlying brain events, as described by brain scientists, contain all the reality there is to them. The critics are forgetting, however, that science need not work this way. There is no reason in principle why understanding the phenomena underlying something should make our existing ways of understanding that thing go away. This only tends to happen when serious flaws are discovered in the existing ways of thinking and speaking, and this is not in general what is being proposed by the neuroaestheticians. This sort of noneliminative reduction can serve to ground aesthetic experience more firmly in scientific perceptives, making it more solid and real rather than less.

Suggested Readings

Aristotle (1984) In: Barnes J (ed.) Poetics, The Complete Works of Aristotle. Princeton, NJ: Princeton University Press.
Collingwood RG (1995) Principles of Art. Oxford: Clarendon Press.
Dewey J (1934) Art as Experience. New York: Perigee Books.
Gombrich E (1956) Art and Illusion: A Study in the Psychology of Pictorial Representation. New York and London: Andrew William Mellon Lectures in the Fine Arts, 5; Bollingen Series, 35: 5.
Gombrich E (1979) The Sense of Order, A Study in the Psychology of Decorative Art. Oxford: The Wrightsman Lectures.
Goodman N (1976) Languages of Art: An Approach to a Theory of Symbols. New York: Hackett.
Jackson F (1982) Epiphenomenal qualia. Philosophical Quarterly 32: 127–136.
Jackson F (1986) What Mary didn't know. The Journal of Philosophy LXXXIII 5: 291–295.
Kawabata H and Semir Z (2004) Neural correlates of beauty. Journal of Neurophysiology 91: 1699–1705.
Kant I (1790) In: Nicholas W (ed.) The Critique of Judgment. Oxford: Oxford University Press.
Onians J (2007) Neuroarthistory: From Aristotle and Pliny to Baxandall and Zeki. New Haven and London: Yale University Press.
Plato (2002) In: Cooper JM (ed.) Complete Works. New York: Hackett.
Ramachandran VS (2004) A Brief Tour of Human Consciousness. New York: Pi Press.
Rizzolatti G and Craighero L (2004) The mirror neuron system. Annual Reviews of Neuroscience 27: 169–192.
Sacks O (2007) Musicophilia: Tales of Music and the Brain. New York: Knopf.
Stafford BM (2007) Echo Objects: The Cognitive Work of Images. Chicago, IL: Chicago University Press.
Wollheim R (1980) Art and Its Objects. Cambridge: Cambridge University Press.
Zeki S (2000) Inner Vision: An Exploration of Art and the Brain. Oxford: Oxford University Press.

Biographical Sketch

William Hirstein is professor and chair of the philosophy department at Elmhurst College, in Elmhurst, Illinois, USA. He received his PhD from the University of California, Davis, in 1994. His graduate and postdoctoral studies were conducted under the supervision of John Searle, V.S. Ramachandran, and Patricia Churchland. He is the author of several books, including *On the Churchlands* (Wadsworth, 2004), and *Brain Fiction: Self-Deception and the Riddle of Confabulation* (MIT, 2005).

Melinda Campbell is associate faculty in arts and humanities at National University and also teaches courses in philosophy and women's studies at San Diego Mesa College. She is an artist and the author of several articles on the metaphysics of color. She studied with Richard Wollheim and received her PhD from the University of California, Davis, in 1993.

Altered and Exceptional States of Consciousness

A Revonsuo, University of Skövde, Skövde, Sweden; University of Turku, Turku, Finland

Glossary

Hypnagogic state – The transitional state between wakefulness and sleep when falling asleep.
Hypnopompic state – The transitional state between sleep and wakefulness when waking up from sleep.
Lucid dream – A dream during which the dreamer realizes that the experience is a dream.
Muscular atonia – The lack of muscle tension in voluntary muscles, causing temporary inability to move.
NREM sleep – Non-REM sleep, stages of sleep without rapid eye movements and less frequent dreaming, or no dreaming at all.
Psychedelic drugs – Chemical substances that affect the central nervous system and cause radical changes in overall mental function, such as hallucinations and delusions.
REM sleep – Rapid eye movement sleep, a stage of sleep when the eyes move rapidly behind the closed eyelids, and when the subject is frequently dreaming.

Introduction

Consciousness, the stream of subjective experiences, manifests itself in many different forms, such as waking perception, dreaming, and mental imagery. Under some circumstances our subjective consciousness becomes so different from our usual experience that the state of our consciousness can be said to be qualitatively altered. The notion of altered state of consciousness (ACS) presupposes that there is some definable normal or baseline state of consciousness that is temporarily lost or somehow transformed in ACSs. What exactly is the normal state of consciousness and what counts as an altered state is difficult to define precisely and therefore continues to be a matter of debate. In the following, different definitions of altered state are first considered and then the major phenomenological features of the most significant ASCs are reviewed.

Some altered states are called exceptional or higher states of consciousness. The notion of a 'higher' state of consciousness suggests that different ASCs can be ordered in a hierarchical manner such that some altered states are lower or deficient in comparison with the 'normal' state of consciousness (say, hallucinations induced by high fever or the delirium caused by heavy drinking), whereas others, the higher states, are in some sense better, more desirable, or perhaps spiritually more advanced than the normal state. The so-called higher states are usually experienced as positive and personally significant. Many of them are closely related to religious and mystical experiences and thus are among the most intriguing ASCs. The major phenomenological features of exceptional and higher states will be reviewed and possible explanations for them explored.

In the scientific study of consciousness, ASCs may shed light on aspects of consciousness that remain otherwise hidden. Therefore, they constitute a unique source of empirical evidence for theories and models of consciousness. They may also reveal the underlying brain mechanisms of some aspects of consciousness that would be difficult or impossible to study by focusing only on the normal state of consciousness.

What Is an ASC?

All definitions of ASC take for granted that there is a normal baseline state of consciousness, and that any ASC only temporarily deviates from this

normal state in some crucial manner. Hence, an ASC is a temporary, reversible state that typically lasts from a few minutes to at most a few hours. Permanent, irreversible changes in conscious experience, as in psychiatric or neurological disease, are usually not counted as ASCs.

In an ASC, what exactly is the nature of the deviation from the normal state? Can it be precisely defined so that a distinguishing feature (or a set of such features) can be pointed out that separates all the ASCs from the typical normal state of consciousness? Attempts to define the concept of ASC and the empirical criteria that separate ASCs from the normal state have been put forward, but it remains unclear if any of these definitions is successful in drawing a clear line between normal and ASCs.

Typically, the 'normal,' unaltered state of consciousness (henceforth NSC) is taken to be the state where we are awake, alert, perceptually aware of our own self and of the physical and the social environment, and capable of rational thought and smooth behavioral interactions with the environment. Our thought processes, sensory and perceptual representations, emotional experiences and reactions, as well as current beliefs reflect our current situation in a relatively realistic manner.

When in the NSC our thoughts and beliefs are not delusional, our sensations and perceptions are not hallucinatory, our emotional experiences and expressions are not overly exaggerated, and our voluntary mental activities and external behavior is under control and not entirely inappropriate for the physical and the social situation in question. If any of these features are however present in our experience or behavior, they may be signs of an ongoing ASC.

Still, this list of potential signs of an ASC is hardly a good definition of the concept. We should expect the definition of the concept of ASC to draw a clear line between the paradigmatic NSC and all the different ASCs. Thus, the ideal definition should focus on the common core of all the different types of ASCs, and also connect the notion of an ASC to an overall conceptual or theoretical framework in the science of consciousness. In the following sections we will consider different attempts to define the common core of all ASCs.

ASCs as Defined by an Overall Changed Pattern of Experience

One way to define an ASC is to say that in an ASC, the overall pattern of subjective experience is significantly different from the baseline NSC. The idea behind this definition is that the change that has happened in consciousness is global in nature and therefore affects several different dimensions of experience, cognition, or behavior. The dimensions of experience where the changes take place are many and varied, for example, attention, perception, mental imagery, inner speech, memory processes, thought processes, meaningfulness of experience, time experience, emotional feelings and expressions, level of arousal, self-control, suggestibility, body image, and personal identity.

While such a list is surely illustrative of aspects of experience that may be altered in ASCs, it still fails to offer a crisp general definition of the concept of ASC. Clearly, different ASCs may involve more or less global changes in subjective experience (i.e., involve changes in only a few or almost all of the above-mentioned dimensions). Thus, it remains unclear how covering or drastic the changes should be to count as an ASC. The borderline between NSC and ASCs thus remains rather vague. A related problem is that these same dimensions of consciousness vary widely also within the boundaries of the baseline NSC. Watching a film in a movie theater, giving a lecture, taking a shower, playing football, driving a car, singing in a choir, diving into a swimming pool, walking in the woods, making love, joking with friends, sitting in a dentist's chair while one's tooth is being drilled – all these experiences usually take place in the NSC, yet we may say that the overall pattern of subjective experience (and cognition and behavior) surely is drastically different when we move from one situation to another in the above list. The richness and sheer variety of the contents of experience within the scope of the NSC makes it difficult to draw a sharp line between the NSC and ASCs only on the basis of changes in the patterns of experience.

Furthermore, in some ASCs the contents or overall patterns of experience are not much different from the NSC. For instance, a dream-related

ASC called false awakening is known to occur during sleep, often in the morning at the time when one should get up. This dream experience may be almost indistinguishable from the corresponding events in the NSC. In false awakening, the sleeper dreams about waking up, getting out of bed, and engaging in normal morning routines, only to wake up again after awhile to discover that he or she is still in bed and that the first experience was but a realistic dream of getting up. During the false awakening the sensations, percepts, emotions, thought processes, etc. may be almost identical to a real awakening – no wonder then that the sleeper does not realize he or she has not woken up at all!

Thus, the overall patterns of experience change radically within the NSC, whereas in some ASCs, the overall pattern of experience remains identical to a corresponding NSC. Therefore, the definition of ASC as an altered pattern of experience may not be able to demarcate the NSC from ASCs in a reliable manner.

ASCs as Defined by the Recognition of a Change in the Overall Contents or Patterns of Experience

Perhaps a better definition could be reached by adding a further condition: the subject having the ASC must feel or recognize that his or her experience is remarkably different from the normal state. This definition adds the requirement that an ASC must be reflectively recognized or recognizable by the subject having it. Clearly, this is not true about many ASCs during the time they take place. For example, during dream experiences, we are usually oblivious to the fact that we are dreaming, and it is extremely difficult to arrive at the conclusion that 'this is a dream' while we continue to dream. Also, sometimes in the NSC we may mistakenly believe we are in an altered state, if something utterly unexpected or shocking suddenly happens so that we have difficulties in believing what we see before our very eyes, but might for a moment think that we must be dreaming.

However, if we allow that the recognition of the ASC as an ASC may take place also afterward when the altered experience is already long gone, then we will be able to correctly classify most ASCs, such as dream experiences, as ASCs. Still, it seems we do the classification not on the basis of the changed pattern of experiences as such – in some mundane dreams and especially in false awakenings, the experiences themselves are identical to waking experiences in NSC – but rather on the basis of the observation that we must have been asleep and therefore only dreaming. Here, the crucial difference between NSC and ASC is not in the content of experience, after all, but rather in the relation between the experience and the real world that the experience represents.

ASCs Defined as the Changed Relation between the Content of Consciousness and the Real World

This leads us to a third potential definition ASCs. Perhaps the core of all ASCs is not the change in the overall patterns of experience (which of course may also happen) but rather the fact that while in an ASC, the content of our experience relates differently to the real world. According to this definition, the causal or representational relation between contents of experience and their typical sources breaks down in such a way that (at least some) contents of experience in consciousness carry false information about (some aspects) of the world, or of ourselves. The contents of experience, in other words, in some way misrepresent external reality or the self.

This definition can easily handle all cases of hallucinatory sensations or percepts. Positive hallucinations are experiences that have no corresponding stimulus as cause in the external world. Our experience thus misrepresents such causes as being out there, but the pattern of experience need not be out of the ordinary in any way. Negative hallucinations, conversely, are omissions of stimuli from experience that would become experienced in the NSC but are not accessible to consciousness because of the ASC. For example in hypnosis or deep meditative states, subjects may not see or hear some stimuli at all if the direction of attention and the contents of consciousness are strongly dominated by other things. Also altered time experiences are covered by this definition: in an

ASC, our conscious experience misrepresents objective time so that it seems to flow either more slowly or more quickly than it really does. During meditation or hypnosis, an hour may pass in a time that feels like a few minutes only.

This definition also seems to include delusions, changes in memory functioning and rational thought. Delusions are false beliefs held in the face of contrary evidence or despite knowledge of the actual state of affairs. In dreams we are often deluded into believing things we would never accept during wakefulness. Psychedelic drugs inducing ASCs may have similar effects. Delusions thus misrepresent the world by interpreting some aspects of the world or ourselves as being in some sense something quite different than what we would take them to be in the NSC. An ASC may involve both retrograde and anterograde amnesia. During the ASC we may not be able to access memories we would have no difficulties in retrieving in the NSC. Moreover, the events that happen during the ASC may not be normally encoded into memory and thus our memories of the experience may be vague and quickly forgotten. When we dream, we usually cannot recall our true circumstances in the real world: that we went to bed, what happened during the day, or the place where we are sleeping, although we immediately remember such things if we wake up for a moment, even in the middle of the night in total darkness. During dreaming memory encoding does not work normally: we tend to forget our dream experiences more quickly and more easily than corresponding waking experiences. Some subjects show spontaneous amnesia about events during hypnosis. Furthermore, false or confabulated memories may be spontaneously formulated, so that during an ASC we seem to remember something that in actual fact has never happened and that we would never take seriously in our NSC. Thus, in these cases our memory (or our conscious access to our memory) misrepresents our past experiences in one way or another, and our critical thinking is not working properly to recognize the fact that we are deluded, amnesic, or confabulatory.

Still, also the definition of ASCs as involving misrepresentations has its difficulties. Trivial illusions and misperceptions sometimes occur in the NSC, thus not all types of misrepresentations count as criteria for an ASC. Furthermore, the notion of misrepresentation of reality may not apply so easily to changes in attention or emotional experience and expression. But if we take emotional states as normally being caused by some external circumstances and/or internal beliefs, then having strong emotions such as fear or ecstasy without any appropriate external or internal cause may be comparable to a kind of misrepresentation of our true situation.

Definition of ASC

Perhaps the most workable definition of an ASC could be arrived at by combining the above ideas: An ASC is any temporary, reversible state of consciousness in which the relationship between at least some of the patterns of experience, and their typical, appropriate causes has been changed so that some patterns of experience tend to occur without their appropriate causes, or some patterns of experience do not occur despite the presence of their appropriate causes, or both. An individual should be able to recognize either during the ASC or after it that an ASC is or was occurring, that the pattern of experience was different and its connection to the external world was not equally accurate as in the normal waking state.

An ASC may be brought about by multiple different circumstances such as perfectly normal changes in brain physiology and neural activity (sleep and dreaming), chemically induced changes in brain physiology (alcohol, psychedelic drugs), special patterns of stimulation in the physical and/or social environment (hypnotic inductions and suggestions, sensory deprivation, strong rhythmic or repetitive patterns of stimulation) or in internal control of attention (meditation). In the following we will take a closer look at the major types of ASCs.

Typical ASCs

Sleep-Related ASCs

While the most typical NSC is an alert state of wakefulness, the most typical ASCs are associated with sleep. When we fall asleep, the patterns of experience and the connection between experience and the external world change suddenly and

drastically. At the borderline between wakefulness and sleep, we may experience an ASC that involves a mixture of perception and dream images. This transitional state from wakefulness to sleep is called hypnagogia (literally, leading to sleep) and the internally generated images in this state are called hypnagogic hallucinations. Conversely, hypnopompic hallucinations (meaning literally: leading out of sleep) occur in the transitional state from sleep back to wakefulness. Hypnagogic and hypnopompic hallucinations are regarded as intrusions of rapid eye movement (REM) sleep imagery into the borderline between wakefulness and sleep. During these hallucinations some degree of perceptual or bodily awareness remains, but other aspects of experience consist of hallucinatory images. Most typical are visual hallucinations of various kinds, from simple geometric forms to objects, faces, animate characters, or entire landscapes. Also auditory phenomena are common: noises, sounds, music, or human voices. Other sensory modalities may also be involved, such as bodily feelings of various kinds or tactile sensations. Hypnagogic hallucinations give way to brief sleep-onset dreams where sensorimotor events from the previous days are briefly revisualized. If you have studied intensely the whole day you may see similar figures or words dancing in front of your eyes at sleep onset, or if you have played sports or computer games, you may see brief excerpts of the game. Also other kinds of brief dreams with complex content may be experienced at sleep onset.

Another ASC that often takes place in the hypnagogic or hypnopompic state is sleep paralysis. It is a mixture of wakefulness and REM-sleep-related muscular atonia: the subject feels awake, but cannot move any part of his or her body. This may be accompanied by difficulties in breathing or by the feeling that something heavy is pressing against one's chest. Sometimes this is perceived to be an evil character that is sitting on the chest (this is called 'the old hag experience'). Sleep paralysis is also often associated with the sense of an evil presence or the strong feeling that there is another person or being present close by, observing the subject and having some sort of evil intentions toward him or her.

Thus, sleep paralysis may be a frightening experience emotionally. It has been speculated that the true origin of many experiences taken to be paranormal in the vernacular (ghosts, apparitions, UFO abductions) is to be found in sleep paralysis and hypnagogic hallucinations. Such phenomena typically happen during the night when the subject is lying in bed in a dark room and likely in a transitional state between sleep and wakefulness. As the experiential content of these ASCs may be intense, realistic, and extremely frightening, the subject who has never heard about sleep paralysis or hypnagogic hallucinations may interpret the experiences as representing real but paranormal events in the world.

During sleep, subjective experiences of some kind occur most of the time, constituting the contents of further sleep-related ASCs. About 85% of REM sleep awakenings and about 25%–50% of non-REM (NREM) awakenings lead to reports of subjective experience. Subjective experiences during sleep can be divided into two categories: sleep mentation and dreaming. The difference between these two is in the complexity of experience. Typical sleep mentations consist of a single image that occurs in a single sensory modality and remains static or repeats itself in the same form. An image of a visual object, a word or sentence or sound heard repetitively, or a thought that runs through the mind again and again are common types of sleep mentation. By contrast, dreaming involves complex, organized, and animated imagery in multiple sensory modalities that shows progression and change through time. Thus, dreams depict a sensory–perceptual world with objects and characters, and simulate events that take place in such a world.

Systematic research on the content of dreams has shown that all of our sensory modalities may be involved in dreams, vision and auditory experiences being the most common however, that most dreams have a central character or a dream self, who is a representation of the dreamer in the dream; that most dreams also include other human (or animal) characters; and that social interaction and communication between dream characters is common. Negative events and emotions are more common in dreams than positive ones, and some activities we often engage in the real world are almost totally absent from the dream world, such as reading, writing, typing, working

with computer, or calculating. Dreams often include bizarre contents and events that would be impossible or highly unlikely in the real world, but we suffer from disorientation, amnesia, and a lack of critical thinking in dreams, and therefore we are only rarely able to recognize the peculiarity of the dream events while they go on, or able to recognize the dream for what it is – a full-scale hallucination or simulation of a world.

Lucid dreaming is the name for the special ASC in which we become reflectively aware of the fact that our ongoing experience is a dream while the dreaming continues. Thus, lucidity renders the dream into a pseudohallucination: a hallucination whose subject is fully aware of the hallucinatory nature of the current experience. Lucid dreaming can be considered a higher state of consciousness and thus will be handled in more detail with other higher states below.

Further sleep-related ASCs include bad dreams or extremely unpleasant dreams that do not wake the dreamer up, nightmares or extremely unpleasant dreams that wake the dreamer up, and sleepwalking or nocturnal wandering which involve a mixture of NREM sleep and wakefulness: the sleepwalker's eyes are open and he or she usually pursues some unreasonable goal without realizing that he or she is still sleeping and that the goal does not make sense.

All in all, sleep-related ASCs are perhaps the most common types of ASCs. Dreaming and sleep mentation are states that occur basically every night for every one of us even if we cannot recall the experiences we had after we wake up. While they are rightly classified as ASCs, there is nothing pathological about them: they are not abnormal phenomena, but perfectly normal concomitants of physiological sleep.

Hypnosis as an ASC

Hypnosis occurs in a situation where a hypnotist gives first a hypnotic induction to a subject, telling the subject to intensely focus on something (such as the hypnotist's voice, a light), relax, and let one's eyes close. After the induction to hypnosis, the hypnotist gives more specific suggestions about specific changes in experience for the subject. The suggestions typically concern changes in how one feels one's own body or one's own actions (e.g., feelings of heaviness in one arm, inability to open one's eyes, numbness or painlessness of a part of the body) or changes in sensation, perception, memory, or thinking (e.g., seeing or hearing something that is not there, not remembering something one normally remembers, not being able to think straight, or believing uncritically some statement and acting as if it were true).

Different people respond differently to hypnotic inductions and suggestions. Some experience almost nothing at all, whereas others report that they really experienced all the things that the hypnotist suggested to them. Among hypnosis researchers, there is a long-lasting controversy whether hypnosis involves an ASC or whether hypnosis is just a peculiar social situation where subjects behave according to their expectations and play along with the rules (just like in any other social situation) but do not enter any kind of ASC in the process. This controversy has been difficult to resolve empirically as there have been no universally accepted definitions or measurable criteria for what would count as an ASC. One possibility is that only a small proportion of subjects truly enter an ASC after getting the hypnotic induction. These very highly hypnotizable people, also called hypnotic virtuosos, may, due to the suggestions given to them by the hypnotist, experience vivid hallucinations in different sensory modalities, amnesia such that they suddenly cannot remember something that they obviously do remember outside hypnosis. They may also forget totally what has happened during the hypnosis session, and have an altered sense of time, thinking that they were under hypnosis only a few minutes when an hour has passed.

Hypnotic suggestibility or the ability to experience the suggested changes is normally distributed in the population, so that most people are moderately suggestible, whereas a small proportion of people is not suggestible at all and feels no changes in experience, and an equally small proportion is highly suggestible and feels many kinds of changes in experience if such changes are suggested to them. Thus, most people, when hypnotized, experience at most only similar things as one is expected to experience with guided mental imagery in a relaxed state. What happens in hypnosis

for most people, especially the low and moderately hypnotizable ones, does not necessarily involve any ASC, but only mental imagery, expectations, and playing voluntarily along with the hypnotists' suggestions.

Exceptional or Higher States of Consciousness

Exceptional states, also called higher states of consciousness, are considered deeply meaningful, satisfying, and desirable, but also difficult to reach or maintain. They go beyond the NSC in the sense that in them subjective experience reaches extreme attentional, emotional, or cognitive levels.

An exceptional or higher state characterized by changes in attention involves total absorption with the object in the narrow focus of attention (one-pointedness of mind), or alternatively the widening of attention to simultaneously cover the entire sensory–perceptual field (full awareness or mindfulness). Higher attentional states are often characterized by the absence of reflective thoughts, especially negative ones, and a deep inner peace or calmness of mind, which can be experienced as highly satisfying or even blissful.

Higher emotional states typically involve strong positive feelings of well-being, contentment, loving-kindness, compassion, joy, elation, or bliss. The quality of inner emotional experience is thus characterized by happiness: the presence of intense positive affect and the absence of negative affect.

Higher cognitive states involve feelings of deep understanding, sudden revelation or insight into the nature of things, glimpses of higher knowledge about the order of the universe, or feelings of being directly connected or absorbed into the cosmos or with higher spiritual realms or beings, such as god. In these states, one seems to get in touch with deeply meaningful information about the nature of reality or have direct knowledge of it. However, it is unclear whether such information or knowledge is actually possessed, or is it only a feeling of deep insight without any actual informational content. In any case this knowledge is often impossible to express precisely in words or it is easily lost once one returns to the ordinary state and level of consciousness. Even when not lost but recalled, it may lose its original significance and seem a trivial platitude when reconsidered in the NSC. This is especially true of the deep insights and supposed truths obtained in drug states, like under the influence of LSD.

The attentional, emotional, and cognitive components of exceptional or higher states of consciousness may appear separately or in various combinations in different ASCs. Next, we will review some paradigm examples of exceptional or higher states.

Meditation

Meditation as such is not a higher or ASC, but rather a set of various techniques and practices that aim at controlling and altering consciousness. Thus, meditation may lead to an altered state, and sometimes this is the explicit goal of meditation. There are far too many radically different meditation techniques and traditions to cover in this context; thus, only some of the most central principles and techniques can be mentioned here.

In one way or another, different meditation techniques involve deliberate control or manipulation of attention. In concentrative types of meditation, the scope of attention is kept narrow and highly selective, as only a particular content of consciousness (an object, a mental image, a word or a sentence, a repetitive action such as breathing) is fixed into the focus of attention for prolonged durations whereas everything else, all distractions, are driven out of consciousness. In mindfulness types of meditation, by contrast, the scope of attention is widened to encompass all available sensations, percepts, emotions, bodily feelings, etc., to be vividly aware of all of them in as much detail and intensity as possible.

In Buddhist and Yoga literature, the term 'samadhi' refers to a higher state of meditative consciousness in which perfect concentration is reached and where the distinction between the object of meditation and the subject who meditates totally disappears. This state is characterized by mental one-pointedness and a merging together of the object and the subject. According to some traditions, the systematic practice of meditation to reach samadhi states can lead to progressively higher mystical states of consciousness, such as

nirvana or enlightenment, which will be discussed below in connection with mystical states.

Some forms of meditation combine the attentional and the emotional components of higher states and deliberately focus on diminishing negative emotions or strengthening positive emotions, such as loving-kindness or compassion. Thus, such meditative practices may lead to higher states of consciousness with intense positive emotional experiences. There is evidence from recent brain imaging studies that Buddhist monks, who have practiced this type of meditation for decades, actually do reach a state of consciousness unreachable by beginners or laymen, evidenced by the strong changes in brain activity that correlates with their meditative state.

Optimal Experience and Flow

Optimal experience is a higher emotional (and attentional) state of consciousness which characterizes the best moments of our lives, moments when we feel deep enjoyment, exhilaration, or happiness and forget about everything else. Empirical studies have revealed that people reach this kind of 'flow' state of mind when there is a certain kind of order in consciousness: attention is firmly focused on reaching a meaningful, challenging goal; we are intrinsically motivated to reach the goal for its own sake, and our skills and resources are just sufficient to reach the goal. When we struggle toward such goals, we momentarily forget about everything else, including the sense of time and our own selves; we become fully immersed into the actions necessary to reach the goal. During such moments, experience simply flows onward, we feel in control of the situation, and our minds are free of worries. Experience becomes one with the actions leading to the meaningful goal, and everything else disappears from consciousness.

In some ways, intense flow experiences are similar to meditative samadhi states where self-awareness disappears and experience becomes one with the focus of attention. Flow states may emerge in almost any kind of activity, such as listening or performing music, engaging in games or sports, hiking in nature, immersing oneself in a conversation with an interesting person, reading a good book, sailing, and so on.

Runner's High

A higher state of consciousness that sometimes occurs during endurance running is known as runner's high. It has to some extent similar phenomenological features as flow and samadhi experiences. This is understandable because, like some forms of meditation, endurance running is associated with highly regular, long-lasting rhythmic patterns of action and breathing. And like typical flow-producing activities, it is challenging but not anxiety arousing, and it involves physical activity where awareness and action can become merged together. In runner's high, reflective or analytical thoughts disappear and subjective experience becomes immersed in the here and now. Intense feelings of pure happiness, timelessness, unity with nature, inner harmony, boundless energy, and floating may emerge. At the same time, there is reduced awareness of one's surroundings and reduced sensitivity to bodily discomfort or pain. A similar state may emerge also in connection of other types of endurance training.

Lucid Dreaming

Lucid dreaming is primarily a higher cognitive state of consciousness. Sleep laboratory studies have confirmed that lucid dreaming takes place during REM sleep. The defining feature of lucidity is the cognitive realization or reflective consciousness of the fact that 'This is a dream!' When this realization takes place, the dream changes from an ordinary one to a lucid dream, and lucidity lasts as long as the dreamer is aware of the fact that he or she is dreaming. Lucidity is like an awakening within the dream, possessing the revelatory knowledge that the whole world around me right now is unreal or hallucinatory, none of the objects or persons around me really exist, they are mere inventions of my dreaming mind.

Although the defining feature of this higher state is cognitive in that it constitutes of possessing knowledge that one ordinarily does not have during dreaming (lucid dreams have also been called 'dreams of knowledge'), there are also attentional and emotional components to this state. Once lucidity ensues, the dreamer can deliberately pay attention to features of the dream world, make deliberate plans of action and carry them out within the dream, or explicitly recall the facts of waking life from long-term memory.

The ability to carry out deliberate and even preplanned actions was the key to the laboratory studies in which it was shown that lucidity occurs during continuous REM sleep; no disruption of sleep or brief awakening is involved. Highly trained lucid dreamers are able to give preplanned eye-movement signals in the dream when lucid. The eye-movement recordings show that these incontestable objective signs of lucidity are clearly recognizable and that the EEG at the same time was typical of uninterrupted REM sleep.

The heightened attentional state in lucid dreaming is often accompanied by a heightened realism of the dream world where the sensory–perceptual features seem almost unnaturally vivid, clear, radiating, and beautiful. Emotionally, lucid dreaming is often characterized by a positive tone, a feeling of full control, freedom and well-being, sometimes even elation.

Although many people may have been briefly lucid during dreaming, for most people lucidity happens only very rarely, if ever. In dream samples, lucidity occurs on average only in one or a few dream reports out of a hundred. Only about 20% of people report having lucid dreams at least once per month. However, lucidity is a learnable skill and various training programs and tips exist that increase the probability of becoming lucid. These include frequent reality testing while awake (asking yourself frequently during the day: Is this a dream?), paying attention to impossible oddities and bizarre features of the dream world which reveal that it must be a dream, and reminding oneself before going to sleep of the intention to become lucid. Also technical devices exist that give signals (a flashing red light, for example) to the dreamer in REM sleep that are supposed to be perceived within the dream without waking up the dreamer. The signal is expected to intrude into the dream and be noticed by the dreamer (why is the world suddenly flashing in red?). This is supposed to immediately lead to the realization by the dreamer that 'This is a dream!'

Exceptional and Mystical Experiences

Out-of-body experiences (OBEs)

OBE is an experience where the subject has a visual perspective or spatial location, which seems to be outside the subject's physical body. The thinking, acting, and perceiving subject or self seems to have left its physical body behind, and may see its body from the outside, usually from above. The subject often feels that the perceptual environment seen in this state is identical with the actual environment. The subject may feel that although the physical body has been left behind, he or she still possesses some kind of ghostly body. In the old parapsychological literature this is known as the 'astral body.' In some cases the subject has no clear body image at all, but constitutes a vague cloud or only a formless point of view.

About 15%–20% of people report having experienced an OBE. In most cases, OBEs occur when the person is lying down but apparently in the waking state rather than sleeping. OBEs may occur at any time and under any circumstances, however, also during intense physical and mental activity, and sometimes in response to life-threatening situations. An OBE usually lasts for a few seconds to a minute.

OBEs often have features that are similar to other higher and mystical states of consciousness. The subject may have the impression of being able to see distant events, or to be able to travel at will to any place. A sense of freedom and control reminiscent of lucid dreaming may occur, as well as feelings of exhilaration or elation, resembling mystical experiences.

People typically interpret OBEs as evidence that something – a spirit or a soul – actually did leave the body during the experience. There is little objective evidence that this would ever have been the case. Experiments where the OBE subject's task has been to retrieve some otherwise inaccessible information from the world (e.g., a number written on a piece of paper and placed out of ordinary sight) while out of the body have not produced any convincing results.

Cognitive and neuropsychological theories try to explain the phenomenon by referring to hallucinatory dissociations between visual perspective and body image. In recent studies, OBEs and other similar distortions of body image and visuospatial perspective have been correlated with and induced by activity in particular cortical areas (the temporoparietal junction). Thus, one explanation for

OBEs may be a temporary failure to bind the body image and the visuospatial representation of the world coherently together in the temporoparietal cortex.

Near-death experiences (NDEs)

Near-death experiences occur when a person's life is physically threatened (e.g., cardiac arrest, drowning), when the person perceives that death is imminent even without or before any fatal physical damage (e.g., falling from a height), and sometimes in connection of nonlife threatening events (e.g., general anesthesia). The core features of typical NDEs are, in the order in which they are usually experienced: (1) peacefulness and weightlessness, (2) an OBE, (3) a dark tunnel into which the subject is drawn and through which the subject feels moving, (4) seeing a brilliant light at the end of the tunnel, and (5) entering the light or another world at the end of the tunnel. This last stage may be associated with meeting dead relatives, religious figures, and with a review of one's life. Also, at this stage the experience becomes very difficult to describe, reminiscent of other mystical experiences.

The estimations of the incidence of NDE in people who have become near to death vary from 10% to 50%. Among people who have experienced NDE, most have reported only the first stage of feeling peacefulness (60%). Only 10% report proceeding through to the stage of entering the light or an otherworldly realm. The core content of NDEs is remarkably similar across cultures, times, and different study populations, although only few subjects have experienced all the typical features of NDE. Age, sex, personality, or religious beliefs do not separate people who have had NDEs from those who have not, even though being equally close to death.

Explanations of NDE can be roughly divided to supernatural (dualistic) and natural (physiological, psychological and neurocognitive). According to the first type of explanation, which can also be called the afterlife hypothesis, what happens in an NDE is that a nonmaterial soul or self is detached from the body, it travels through the tunnel into another spiritual realm where it is met with deceased relatives, and angelic or godlike being or beings radiating unconditional love. There, the person's life is reviewed like a film and some sort of self-judgment takes place, as well as a decision whether to go back to earthly life. After returning, the subjects themselves often feel profoundly transformed and regard the afterlife hypothesis as a self-evident explanation to their experience.

According to the naturalistic explanations, also called the dying brain hypothesis, changes in physiological processes and brain function can account for NDEs. First, the feeling of peacefulness, positive emotion and bliss could be brought about by increased endorphin release in the brain under stress. Endorphins may also trigger abnormal or seizure-like activity in the temporal lobe. Epileptic seizures and direct stimulation of the temporal lobe (or the temporoparietal junction) may induce a variety of anomalous experiences, such as OBEs, distortions of body image, realistic memory images, and feelings of the sense of presence of some other conscious being. Anoxia (lack of oxygen) of the brain might lead to the release of cortical inhibition that is known to induce visual hallucinations in other conditions (e.g., drugs, neurological damage of visual pathways). Tunnels are one of four most common types of visual forms typically experienced when visual hallucinations are induced by drugs, seizures, or other causes.

It goes without saying that the 'afterlife hypothesis' is impossible to integrate together with the current worldview of science. The naturalistic explanations are able to account for many of the core features of NDEs, but only by using speculative and indirect evidence, leaving many open questions. There is no direct evidence that during NDEs the hypothesized physiological or neurocognitive mechanisms would actually be at work and would therefore correlate with specific aspects of the experience or cause them. Thus, we have no direct evidence during NDEs of increased endorphin levels, cerebral anoxia, or seizure-like cortical activity in the occipital, parietal, and temporal lobes. This is not to say that such things do not occur, only that it is extremely difficult to get direct measures of them while a person is having an NDE. Another feature that may be difficult to explain by referring to abnormal or pathological brain activity is the well-organized nature and relative universality and uniformity of NDE. Hallucinations induced by epileptic seizures in

the temporal lobe, drug states, or dreamlike states have enormous variability of experiential content both within and between subjects. A uniform and seemingly well-organized experience such as NDE would seem to be based on some mechanism that is widely shared and activated in a roughly similar manner and order in different people, rather than by a variety of processes running wild in a brain under high metabolic stress, very low arousal (unawareness and unresponsiveness to the external world), and burdened by pathological electrophysiological seizures.

Thus, for the time being we have to admit that there is no explanatory model that would satisfactorily predict the existence or occurrence of NDEs and that would account for the remarkably universal and seemingly well-organized phenomenological features of NDE.

Mystical experiences

Mystical experiences are perhaps the 'highest' of all the higher states of consciousness. They involve many similar features as some of the other higher states, but in an extreme form. Also the effects of mystical states on the subsequent life of the person are often deep and long-lasting. Such experiences, even if relatively brief, are vividly recalled for years and they may be regarded as among the most significant moments of life. William James, the father of American psychology, regarded mystical states as closely related to personal religious experience.

Mystical states are difficult to describe in words or communicate to other people. James took this feature, ineffability, as one of the defining features of mystical states. Mystical states involve both emotional and cognitive components. Emotionally, mystical states are intensely positive, involving overwhelming feelings of peace, calmness, harmony, joy, love, elation, awe, or bliss. Cognitively, mystical states seem to communicate highly significant information for the subject about the true nature of the world, revealing the underlying, hidden order of the universe and its guiding principles. Perceptually, mystical states may involve unusual visions or other forms of imagery, or seeing the ordinary perceptual world as unusually bright, clear, radiant, and beautiful. Mystical experiences are characterized by a sense of heightened reality and significance, and the sense of time may be distorted. The experiences are usually brief, from a few seconds to one hour at most, but their aftereffects may last throughout life. They happen unexpectedly and suddenly and cannot be summoned by will, although certain practices (such as yoga or meditation) or drugs (hallucinogens) enhance the likelihood of their occurrence.

Cosmic consciousness is a term introduced by Canadian psychiatrist R.M. Bucke in early 1900s to describe a paradigmatic mystical experience (quoted by William James in *Varieties of Religious Experience*):

> The prime characteristic of cosmic consciousness is a consciousness of the cosmos, that is, of the life and order of the universe. Along with the consciousness of the cosmos there occurs an intellectual enlightenment which alone would place the individual on a new plane of existence – would make him almost a member of a new species. To this is added a state of moral exaltation, an indescribable feeling of elevation, elation, and joyousness, and a quickening of the moral sense, which is fully as striking, and more important than is the enhanced intellectual power. With these come what may be called a sense of immortality, a consciousness of eternal life, not a conviction that he shall have this, but the consciousness that he has it already.

Cosmic consciousness entails a widening of consciousness to encompass the entire universe and its deeper working principles. Although such insights are experienced as being absolute truths by the subject of the experience, outsiders may remain doubtful, and rightfully so. The conviction of the subject and the felt authority of the experience are no guarantee of that the insights gained during the mystical state of consciousness carry any truth or validity in the objective sense.

Enlightenment is an ultimate form of mystical (and religious) experience, and an ultimate or highest conceivable state of consciousness, usually associated with Eastern religions such as Buddhism. Enlightenment is an experience where one reaches, through meditative practices, complete and total understanding of the nature of reality, and of the nature of oneself in relation to reality. In those traditions, the terms 'bodhi' and 'budh' refer to awakening, wisdom, and brightness; thus 'Buddha' literally means 'the awakened one.' Enlightenment thus is a mystical experience that

awakens ordinary consciousness into seeing the true nature of reality and thereby transforms consciousness into a qualitatively different, higher form that transcends normal consciousness, perhaps in a somewhat similar manner as becoming lucid reveals the true nature of the dream world to the dreamer whose conscious state thereby transcends the ordinary dreaming mind.

According to Buddhist thought, enlightenment entails the cessation of all selfish desires and all clinging to material possessions, sensory pleasures, human relationships, and other external passing things. The true nature of everything is seen to consist in impermanence and emptiness; thus even one's own self is seen to be a mere illusion. The meditative state samadhi, discussed above, involves the mystical union of subject and object, or disappearance of self, an important step toward full enlightenment. These revelatory insights and experiences are supposed to bring about an absolute emotional calmness, peace of mind, cessation of suffering, and deep compassion and unconditional love for all the unenlightened beings who continue to suffer. It is unclear, though, whether enlightenment once achieved subsequently persists, or can be lost and again regained later.

Summary

ASCs can be defined as temporary states of consciousness when the relationship between inner conscious experience and the world outside of consciousness has been changed. Thus, in altered states, the patterns of thought and the sensory–perceptual and emotional experiences in some way represent the world or ourselves in a way that does not match with the actual state of affairs. Often, though not always, ASCs also involve highly unusual patterns of thought and experience. Furthermore, we usually are able to recognize when we are in an ASC and when not, or at least after returning to the normal state we can infer that we were in an altered state. Exceptional or higher states are a variety of altered states that involve very positive, desirable and insightful experiences that are felt to be personally deeply meaningful, sometimes leading to profound and long-lasting transformations of personal beliefs and experiences afterward.

Altered and exceptional states of consciousness reveal the richness of different forms and varieties of our subjective existence. Any theory of consciousness should be able to explain not only the typical features and mechanisms of normal waking consciousness, such as sensory and perceptual experiences in response to physical stimuli, but also the features and mechanisms of altered states. This may be a challenge to the scientific study of consciousness, because many altered states are difficult or impossible to control experimentally, they are highly subjective in that their occurrence or content is impossible to verify by outsiders, and sometimes their precise experiential nature is impossible to describe verbally. Nonetheless, there is growing evidence from neurocognitive studies of ASCs such as dreaming, OBEs, and hypnotic hallucinations that ASCs are real in the sense that they have specific, objectively measurable neural correlates and mechanisms in the brain. Brain stimulation studies of the temporal lobe have furthermore established that OBEs and even mystical experiences can be triggered by simply stimulating the brain in the appropriate location. There is thus some hope that even the most mysterious of ASCs are not entirely beyond the reach of scientific experimentation.

See also: The Neurochemistry of Consciousness; Psychoactive Drugs and Alterations to Consciousness; Religious Experience: Psychology and Neurology.

Suggested Readings

Blackmore SJ (1992) Beyond the Body. An Investigation of Out-of-the-Body Experiences. Chicago: Academy Chicago Publishers.

Blackmore SJ (1993) Dying to Live: Near Death Experiences. Buffalo, NY: Prometheus Books.

Bünning S and Blanke O (2005) The out-of-body experience: Precipitating factors and neural correlates. Progress in Brain Research 150: 331–350.

Cheyne JA, Rueffer SD, and Newby-Clark IR (1999) Hypnagogic and hypnopompic hallucinations during sleep paralysis: Neurological and cultural construction of the night-mare. Consciousness and Cognition 8(3): 319–337.

Farthing WG (1992) The Psychology of Consciousness. New York: Prentice Hall.

Green C and McCreery C (1994) Lucid Dreaming. The Paradox of Consciousness During Sleep. London: Routledge.

Hobson JA (2001) The Dream Drugstore. Chemically Altered States of Consciousness. Cambridge, MA: MIT Press.
James W (1902) The Varieties of Religious Experience. New York: Longman, Green.
Kallio S and Revonsuo A (2003) Hypnotic phenomena and altered states of consciousness: A multilevel framework of description and explanation. Contemporary Hypnosis 20(3): 111–164.
LaBerge S (1985) Lucid Dreaming. New York: Ballantine.
Mahowald MW and Schenck CH (1992) Dissociated states of wakefulness and sleep. Neurology 42(supplement 6): 44–52.
Mavromatis A (1987) Hypnagogia. The Unique State of Consciousness between Wakefulness and Sleep. London: Routledge.
Strauch I and Meier B (1996) In Search of Dreams. Results of Experimental Dream Research. New York: SUNY Press.
Tart CT (ed.) (1990) Altered States of Consciousness. New York: Harper Collins.
Vaitl D, Birbaumer N, Gruzelier J, et al. (2005) Psychobiology of altered states of consciousness. Psychological Bulletin 131(1): 98–127.

Biographical Sketch

Antti Revonsuo is a professor of psychology at the University of Turku, Finland, and a visiting professor of cognitive neuroscience at the University of Skövde, Sweden. He is an associate editor of the journal *Consciousness and Cognition*, the director of the consciousness research group at the Center for Cognitive Neuroscience, University of Turku, and a founding member of the faculty of an undergraduate degree program on consciousness studies at the University of Skövde, Sweden. He has done research on consciousness since the early 1990s, focusing mostly on dreaming, hypnosis, visual consciousness and its neural correlates, and theories of consciousness. His major publications include a new theory of the function of dreaming, the threat simulation theory, published in *Behavioral and Brain Sciences* in 2000, a theory of hypnosis as an altered state of consciousness (together with S. Kallio) in *Contemporary Hypnosis* in 2003, and an overall theoretical framework for the explanation of consciousness, published in the book *Inner Presence: Consciousness as a Biological Phenomenon* (MIT Press, 2006).

Animal Consciousness

D B Edelman, The Neurosciences Institute, San Diego, CA, USA

Glossary

Accurate report – A first-person account, through language or other types of voluntary response, of what an individual is experiencing. Accurate report has provided a critical means of investigating conscious states in humans. In the case of nonhuman animals, certain motor channels other than language might also be used to gauge the ability to make higher-order discriminations that might bear on a given animal's conscious states.

Behaviorism – A school of psychology that emerged in the nineteenth century and was developed and famously elaborated upon by John Watson and B.F. Skinner in the early part of the twentieth century. Behaviorism takes the position that everything an animal does can be reduced to behaviors, and, in fact, the only legitimate object of psychological study is behavior. In the behaviorist framework, no recourse to internal, or mental, states is considered necessary or even legitimate.

Cephalopods – The class of marine mollusks that includes octopuses, squid, cuttlefish, and nautiloids. Distinguishing features of the cephalopods include highly prehensile tentacles (eight in the case of octopuses; ten in the case of cuttlefish and squid); very large eyes (octopuses, cuttlefish, and squid have single compartment, camera eyes with lenses that closely resemble those of vertebrates); chromatophore organs in their skin, which allow adaptive camouflage; and, in many species, ink sacs for defense or escape.

Commentary key – A behavioral index that enables an animal to indicate a choice of stimuli in a given task and at the same report the nature and/or contents of its conscious experience during the performance of that task. The 'commentary key' paradigm was used effectively by Alan Cowey and Petra Stoerig to show that rhesus macaques with lesions to one-half of the striate cortex (V1) behaved very much like blindsighted humans in their responses to – and likely experiences of – stimuli presented in intact and occluded visual hemifields.

Decussation – The X-shaped crossing of nerve fibers to opposite sides of the brain or spinal cord. In mammals with stereoscopic vision, the optic tract of each eye is partially decussated to both sides of the brain. In many species of birds, the optic tract of each eye is almost completely decussated to one side of the brain.

Dynamic core – The neural basis of consciousness in mammals as proposed by Gerald Edelman and Giulio Tononi. The populations of neurons in the thalamus and the cortex that function as coherent groups, through richly degenerate – or reentrant – connections, which are proposed to underlie conscious states. The reentrant connectivity of the thalamocortical system is considered by Edelman and Tononi to be uniquely capable of generating the complex dynamic properties of conscious states.

Higher-order consciousness – An explicit awareness of self. A frame of reference that seemingly encompasses past, present, and future scenes. Higher-order conscious states reify the contents of primary consciousness as objects.

Metacognition – Thinking about thinking. Awareness of one's own thoughts. The ability to monitor one's own cognitive states. Metacognition involves both conscious and unconscious knowledge.

Phylogeny – The evolutionary history of a group (e.g., phyla, genus, species) of organisms and/or their constituent characters or traits.
Primary (sensory) consciousness – Awareness of a multimodal scene. Unification of percepts into episodic scenes.
Telencephalon – The anterior forebrain in birds and mammals. The cerebral hemispheres and parts of the hypothalamus are derived from the telencephalon during brain development.
Thalamocortical – The densely reentrant circuitry that connects the thalamus and cortex in mammals and possibly in birds. Complex dynamic neural activity in the thalamocortical system is believed by Edelman and Tononi to underlie conscious states.

Introduction

There is much in our daily experience, as well as numerous anecdotal reports and indirect evidence from cognitive studies, that suggests that human beings are not the only animals capable of conscious states. We are, however, the only animals naturally capable of relating our conscious experiences via linguistic reports, and therefore the only animals for which there is an increasingly robust body of evidence for conscious states in a variety of forms. In the absence of accurate linguistic report, other means must be sought for establishing that certain nonhuman animals are conscious.

Defining Consciousness Across Phylogeny

Critical to the investigation of animal consciousness is both an adequate definition of consciousness as it might manifest itself in nonhuman species and the recognition that consciousness is a process that is instantiated within biological substrates, that is, nervous systems, and as such, has therefore been subject to evolutionary forces.

Sensory- versus Higher-Order Consciousness: A Critical Distinction

First, a necessary distinction must be drawn between sensory, or primary, consciousness – a capacity that might be common across a wide swath of phyla – and higher-order consciousness, which may be unique to *Homo sapiens* and perhaps a small group of intensively trained nonhuman animals, including some great apes and certain birds such as parrots, which are capable of sophisticated linguistic performance, and, in the case of the latter, both vocal learning and, apparently, accurate linguistic report. A variety of claims have been made that certain animals are self-aware since Gordon Gallup's demonstration in 1970 that chimpanzees that had been anesthetized and marked with two red dots – a control dot on an inaccessible region of the animals' bodies and a red dot on the forehead – would, when awakened, behave in a manner consistent with self-recognition if confronted with their reflection in a mirror (the control dot was ignored, indicating that sensory cues other than visual recognition of the dot on the forehead were not involved in the animals' reactions). Although elaborations of Gallup's 'mirror test' have been performed on a number of different species with varying degrees of success (all great apes, dolphins, killer whales, Asian elephants, and certain birds have been reported to show signs of self-recognition, but, interestingly, certain old world monkeys have not), the results of this experimental methodology remain difficult to interpret and somewhat controversial. Among the more trenchant criticisms of the mirror test is the contention that in animals in which vision is not as well elaborated as other sensory modalities (e.g., dogs, which may often rely more heavily on olfaction), visual self-recognition might not be expected, as it is in animals with acute stereoscopic color vision, such as humans. Given the controversial nature of the mirror test as well as the very nascent state of affairs in the development of a rigorous framework for the scientific study of animal consciousness, perhaps it is best to begin by investigating sensory consciousness in nonhuman species.

Sensory consciousness may be defined as the construction of a unified perceptual 'scene,' that is, the integration of input from disparate sensory

modalities into a seamless whole. Higher-order consciousness, or metacognition, is the awareness of a self that is embedded within a unified perceptual scene; it is the awareness of being aware. Arguably, higher-order consciousness is contingent upon language, which may be unique to *Homo sapiens* and perhaps some laboratory-trained animals. Sensory consciousness is contingent upon an ongoing link between perception, across disparate sensory modalities, and memory; in this regard, many mammals and some birds may be said to have sensory consciousness. For the purposes of the present discussion, references to animal consciousness will be ostensibly in regard to sensory consciousness.

Consciousness as a Product of Natural Selection

More than 125 years ago, Charles Darwin cast the emergence of the human mind clearly and unambiguously in the context of an evolutionary continuum, with roots planted deeply within nonhuman animal lineages. To Darwin, the human mind was embedded in the natural world, not something separate and apart from all other organic structures, and therefore a product of natural selection. Indeed, Darwin famously took Alfred Russell Wallace, the codiscoverer of natural selection, to task for Wallace's insistence that the human mind resided outside of the evolutionary continuum. Unlike Darwin's theory of natural selection, which took its place as the founding principle of modern biology with the new evolutionary synthesis of the 1940s, the notion of the animal mind, either as a product of natural selection or otherwise, languished throughout much of the twentieth century, particularly after John Watson, the founder of the American behaviorist school, dismissed it as irrelevant to the task of understanding animal – and ultimately human – behavior. The ascendancy of the radical behaviorism represented by Watson's turn away from internal mental states and B.F. Skinner's stringent behaviorist experimental paradigm influenced generations of academic psychologists and philosophers, and effectively quashed the rigorous study of animal consciousness for much of the twentieth century by erecting a sort of early functionalist perspective in which external, or behavioral, manifestations of the reactions of animal nervous systems occupied central importance to the exclusion of all else. For much of the twentieth century, few dared to challenge this orthodoxy (one early exception was the biologist Donald Griffin, who held that there was an obvious continuum between the cognitive capabilities of nonhuman animals and the human mind). Curiously, although behaviorism has largely been repudiated, many anecdotal observations of the cognitive performance of animals are often dismissed by modern behavioral scientists professing caution, but sounding very much like latter day behaviorists in the tone of their methodological rhetoric.

With the emergence of evolutionary psychology, it has largely been accepted that a variety of cognitive capacities are the products of natural selection. Given such recent attempts to naturalize cognition generally, perhaps it is not unreasonable to finally embed the process of consciousness in nature as well.

A Methodological Framework for the Study of Animal Consciousness

A number of correlates of consciousness have been identified in humans, including reentrant thalamocortical signaling, a characteristic electroencephalography (EEG) signature comprising fast, irregular, low amplitude electrical activity ranging from 12 to 70 Hz, and widespread brain activity correlated with conscious contents. Using the human case as a benchmark and reference, it is now possible to search for such correlates of consciousness in nonhuman species.

Human Consciousness Studies as Benchmarks for Investigating Animal Consciousness

Critical to addressing the question of animal consciousness is the establishment of benchmarks based on the only reliable, reproducible source for data: studies of awake, behaving humans. Such data come from humans reporting their conscious states in a wide variety of experimental paradigms, often involving noninvasive imaging techniques.

Among the most definitive examples are binocular rivalry tasks performed by humans during magnetoencephalography (MEG), in which unmistakable changes in brain activity were observed when consciousness of an object was reported. These studies generally implicated the thalamocortical system in the generation of conscious states. In addition, evidence from strokes and destruction of brain regions has indicated that structures such as the thalamocortical circuit and the mesencephalic reticular formation are critical neural substrates for consciousness. The identification of such an anatomical benchmark, together with neurophysiological recordings during a task requiring rich discriminatory behavior, might suggest the presence of conscious states in primates and other mammals, even in the absence of something like accurate report. Indeed, a binocular rivalry paradigm has been employed in which monkeys were presented with rivalrous and nonrivalrous stimuli, and, during rivalry, the majority of neurons (>90%) recorded in the inferior temporal cortex (IT) fired in response to the perceived stimulus; the monkeys were trained to report the stimuli they perceived by pulling a lever. The IT neurons were not responsive to the nonperceived stimulus. The results of this study suggest that the monkeys were probably conscious of certain visual stimuli.

A theoretical proposal to account for the properties and evolution of consciousness has been elaborated over several decades by Gerald Edelman and his colleagues. Central to this proposal is the idea that consciousness in mammalian species emerged when reentrant connectivity evolved between brain areas for perception and those involved in memory. Evidence for such reentry in humans has been demonstrated using MEG. Also notable is the distinction this proposal makes between two varieties of consciousness: primary consciousness, in which percepts are merged into episodic scenes, each of which is of a piece; and higher-order consciousness, which involves self-awareness, the ability to reconstruct past scenes and formulate future scenes, and in humans the ability to represent both the internal state and the external world symbolically through language or other means.

Types of Evidence for Animal Consciousness

In the absence of accurate verbal reports, how can we assess the conscious states of different animals? Although human language is by far the most elaborate mode of communication known, other sensorimotor channels are sufficiently complex that they might be exploited as means of gauging what a given animal is or is not aware of. Indeed, the 'commentary key' paradigm laid out by Alan Cowey and Petra Stoerig in their study of blindsight in rhesus macaques is one example of how certain sensorimotor channels might be exploited as forms of accurate behavioral report in certain nonlinguistic species. Cowey and Stoerig devised an ingenious experiment in which rhesus macaques with lesions to one-half of the striate cortex (V1) were trained to touch the area of a video display where stimuli would appear. These monkeys were able to learn to detect stimuli presented in their occluded visual hemifields, that is, those hemifields rendered 'cortically blind' by lesions in the contralateral half of V1. They could even make discriminations across stimuli presented in their occluded hemifields. However, the monkeys could not distinguish between stimuli presented in their occluded hemifields and blank regions of the video display (containing no stimuli) presented in their unaffected hemifields. The foregoing result could be interpreted as a report by the monkeys that they were not aware of the difference between the two hemifields.

The extension of Cowey and Stoerig's 'commentary key' paradigm to a variety of sensorimotor channels might provide a means of investigating consciousness in a number of nonhuman species. Moreover, the human benchmark could provide sufficient neuroanatomical and neurophysiological neural correlates for conscious states that might also be investigated in species far from the human, or even mammalian, lineages.

Proceeding from the three principal properties of human consciousness mentioned earlier, it is clear that evidence for consciousness in nonhuman species might be gathered from three areas of study: neuroanatomy, neurophysiology, and behavior.

Evidence for Consciousness in Nonhuman Species

The broadly conserved architecture of the mammalian brain as well as ample evidence of complex behavioral repertoires are at least consistent with the possibility of conscious states in many mammals. But the recent revision of avian neuroanatomy supports a substantial structural homology to the mammalian brain. Moreover, reports of feats of avian memory, deception, vocal learning and reproduction, the capacity of some species to employ lexical terms in meaningful ways, and finally the ability of certain birds to make higher-order discriminations all suggest the elaboration of a nervous system as sophisticated as those of most mammals. In fact, the foregoing discussion would suggest that consciousness emerged in birds, probably independently of mammals.

Birds and Mammals: Evolutionary Homology Lost and Found

Birds exhibit a broad behavioral repertoire, which includes simple nest building in swallows, manufacture and use of tools by New Caledonian crows, food caching by scrub jays and other birds in perhaps a hundred or more sites (which certainly involved complex spatial memory capacity), seasonal migrations of thousands of miles, sophisticated song learning and production, stunningly accurate mimicry of natural and artificial sounds, and, in some birds, word comprehension, production, and naming. This last series of behaviors is based on the capacity for vocal learning, a capacity shared by perhaps seven animal groups: humans, cetaceans, bats, parrots, songbirds, hummingbirds, elephants, and perhaps even some rodents.

Vocal learning, which is richly manifest in a number of forms in birds, might allow us the possibility to test for avian conscious states. In particular, song production in songbirds might act as a channel for something like accurate report, given an appropriately designed experiment. The studies of blindsight in lesioned monkeys by Cowey and Stoerig suggest a possible model for an experiment in which one-half of a bird's visual cortex is lesioned, and the bird is then queried as to whether it 'sees' a particular stimulus in the affected visual hemifield. Its response to a given stimulus, or lack thereof, would be through a vocalization previously entrained to that stimulus. Lesioning the ectostriatum (the avian brain area analogous to V1 in the mammalian tectofugal pathway) might disrupt early visual processing and therefore might lead to a form of avian blindsight. But it should be noted as well that the optic tracts of many birds are nearly completely decussated (>99%) to opposite hemispheres of the brain, and therefore in many birds would preclude the same type of hemifield arrangements found in monkeys and humans.

Another approach might be an electrophysiological paradigm based on the finding by Nikos Logothetis that the majority of neurons in a behaving monkey's IT fired in response to the percept (as defined by a behavioral report), while neurons in V1 responded to the signal. Such an experiment in birds would of course be contingent on identifying neurons functionally analogous to those found in the mammalian IT and striate (or visual) cortex.

Some of the difficulties inherent in the experimental approaches suggested above might be obviated if the bird being tested is an African grey parrot capable of reproducing human speech. Irene Pepperberg and her colleagues have shown that African grey parrots are capable of performing naming tasks after acquiring vocabularies roughly equivalent to those of some of the chimpanzees employed in language-training projects over the past four decades. Notably, these chimpanzees had learned a large number of lexical terms after many years of rigorous training and reinforcement. By far, the most impressive nonhuman symbolic capacity was shown in captive bonobos by Sue Savage-Rumbaugh. One individual, Kanzi, acquired several hundred lexical terms and responded to requests or queries through the use of a keyboard containing arbitrary symbols.

Although the concept of accurate report in humans or something analogous in monkeys (e.g., the commentary key paradigm developed by Cowey and Stoerig and the studies of the IT and striate cortex by Logothetis) was not explicitly invoked in the studies by Pepperberg and her colleagues of naming and categorization in African

grey parrots, by naming objects, these animals appeared to be producing accurate reports of the discriminations they were capable of making. Moreover, Alex, the principal subject of many of Pepperberg's experiments, seemed to be able to make a judgment to the effect that, "I know that something in this perceptual scene has changed, and here is what has changed," when presented with an altered array of objects, suggesting the ability to make discriminations about discriminations. Although such a capability may in fact indicate the presence of higher-order consciousness, and indeed it has been suggested by others that such metacognition necessarily underlies animal consciousness, Pepperberg's more modest conclusion – that parrots may possess primary or 'perceptual' consciousness – for now seems to be a more readily justifiable conclusion.

The overall organization of the vertebrate central nervous system is part of a highly conserved body plan that emerged with the earliest chordates more than 500 million years ago. Identifying homologous neural structures is often not difficult because developmentally, many brain structures in lower vertebrates (i.e., reptiles and lizards), birds, and mammals can be traced to common origins in specific embryological tissues. In fact, there is substantial evidence from comparative embryological studies that suggests that much of the neuronal properties and circuitry that underlie the mammalian cortex were established within nucleated or clustered arrangements long before the evolutionary appearance of the distinct six-layered mammalian cortical mantle. In addition, the somatomotor circuitry of the avian dorsal pallium may be homologous to the mammalian basal ganglia–cortico–thalamic loop. Yet there remains little agreement as to which avian neural structure is the homologue of the mammalian isocortex. Indeed, among the most distinctive gross anatomical properties of the avian brains is the division of the telencephalon into structures resembling nuclei that lack the laminated cortical mantle characteristic of mammalian brains. The avian optic tectum and cerebellum are more elaborated than their mammalian homologues. More basal regions, such as the hypothalamus and preoptic area, are relatively easy to recognize.

Mid-brain regions such as the amygdala and hippocampus cannot be identified easily through structural similarities with mammalian homologues. In such cases, deeper homologies must be sought through investigation of the functional properties of neuronal populations within particular regions. For example, the claim that the avian anterior forebrain pathway is functionally analogous to the mammalian corticobasal ganglia–thalamocortical loop is supported both by observations that the medial nucleus of the avian dorsolateral thalamus (DLM) contains inhibitory as well as excitatory pathways and that neurons in the DLM exhibit functional properties similar to those of thalamocortical neurons. But the use of electrophysiological techniques to establish functional homology, while potentially fruitful, sometimes does not yield results easily. Electrophysiological studies have established common properties of mammalian thalamocortical neurons, such as low-threshold calcium (Ca^{2+}) spikes and slow oscillations, but such studies have yet been adequately reproduced in birds. In instances where neurophysiological means are intractable or unavailable, molecular or histological techniques may be critical in the identification of certain regions of the avian brain. Molecular markers such as neurotransmitters, neuropeptides, and receptors that are specific to certain cells known to reside in particular regions of the mammalian brain can be targeted in avian brain regions by immunohistological techniques, as has been shown by Erich Jarvis and his colleagues. Gene expression patterns too can be compared in developing bird and mammalian brains. Such a comparison of homeotic genes involved in early brain development in chick and mouse embryos has indicated clear structural homology between parts of avian telencephalon and mammalian cortex. The evolutionary conservation of ancient vertebrate neurochemical systems has also been useful in the identification of avian brain regions homologous to known mammalian neural structures. In one instance, the identity of certain avian brain structures as components of the limbic system was established through the use of an antibody to the steroid metabolizing enzyme estrogen synthetase (aromatase), which in mammals is known to be specific to certain cells in the limbic system. Other molecular techniques, such as cloning and *in situ* hybridization, have been useful in the establishment of avian homologues of certain mammalian neural structures. For example, large numbers of glutamate

receptors expressed in the mammalian CNS were successfully cloned from regions of songbird brain, and *in situ* hybridizations in mouse and zebra finch brains using probes based on these clones showed patterns of glutamate receptor expression that were highly conserved in cerebellum, midbrain, thalamus, and basal ganglia. Interestingly, on the basis of these findings, Kazuhiro Wada and his colleagues argue that avian and mammalian brains have become functionally similar through evolutionary convergence. Gross differences in topological organization, they point out, have led others to relegate the avian brain mistakenly to a lower functional status.

Lower vertebrates, birds, and mammals have all been shown to possess certain homologous excitatory and inhibitory systems, including the serotonergic and GABAergic systems, and, in the case of birds and mammals, the dopaminergic system. Although neural structures have been shown to be highly conserved across the vertebrates, gross anatomical and histological data alone have been insufficient to establish this conservation. Where structural homologues have been difficult to establish through anatomy and histology, a variety of different techniques have been applied. For example, parsimony-based phylogenetic analyses of comparative embryological data, lesioning and behavioral response experiments, and comparative studies of gene expression patterns during development have yielded the identity of structures such as the amygdala and hippocampus in widely different nonmammalian vertebrates, including lampreys, goldfish, and birds. The extremely ancient origin of the hippocampus was established through comparative histological studies, which showed that the primordium that gives rise to the hippocampus in mammals is also present in the developing lamprey. Finally, comparisons of gene expression data have shown that the organization of the vertebrate brain was largely established in the earliest vertebrate ancestor.

An increasingly sophisticated armamentarium of molecular techniques is enabling researchers to forge links between avian neuroanatomy and well-characterized regions of the mammalian brain. Such an approach may be instrumental in identifying structures that support conscious states in nonmammalian vertebrates in the same way as the thalamocortical circuit in mammals.

Although it is uncertain when consciousness appeared among the vertebrates, it seems likely that it emerged independently at least twice, sometime after the divergence, 300 million years ago, of the anapsid and synapsid reptilian lines that led to birds and mammals, respectively. Although it is impossible to say whether animals in either of the reptilian lineages preceding birds and mammals were conscious, it seems conceivable that consciousness could have emerged separately and independently in the two different vertebrate lines – a possibility that may find support as homologous structures or similar analogous arrangements in birds are discovered.

The characterization of avian brain structures that are analogous or homologous to the mammalian cortex and thalamus is an important first step in investigating consciousness in birds. The key to identifying conscious states in birds, however, will be the discovery of neurophysiological signatures that resemble those of the mammalian conscious state. One such signature would be patterns of electrical activity that, as Anil Seth and his colleagues suggest, reflect "...widespread, relatively fast, low-amplitude interactions...driven by current tasks and conditions." Similarities between the waking EEG patterns of birds and mammals, as well as slow wave electrical activity recorded during sleep in birds (albeit embedded within an overall EEG pattern that is distinctly different than that of mammals), are at least suggestive of generally analogous avian brain functioning that might support conscious states.

Clearly, birds possess a number of the necessary substrates and conditions for primary consciousness. It remains for us to design experiments, based on the human benchmark and perhaps be informed by such work as that of Cowey, Stoerig, and others, that will make a convincing case that birds are indeed capable of conscious states.

Cephalopods: An Alien Intelligence and a Challenge for Consciousness Research

In invertebrate species, in which the organization of the nervous system diverges markedly from those of vertebrates such as birds and mammals, the case for consciousness becomes understandably quite tenuous. Of the many extant invertebrate

groups, spiders and cephalopod mollusks would appear to have the most sophisticated behavioral capabilities and, on this basis, might be appropriate candidates to test for properties associated with conscious states. Exhaustive scientific accounts of cephalopod behavior suggest that some of these animals – particularly the octopus – exhibit a measure of behavioral sophistication and plasticity that is in many ways on par with certain higher vertebrates.

Although exhaustive evidence has been presented previously by Jennifer Mather to make the case for sensory, or primary, consciousness in certain cephalopods, the overwhelming majority of this evidence has been of a behavioral nature. Behavioral or anatomical evidence suggesting the possibility of conscious states in these animals has not been forthcoming largely because much of the anatomy and physiology relevant to this problem has not yet been worked out. Since it is likely that a key substrate of mammalian conscious states is the reentrant circuitry between the cortex and the thalamus as suggested by Gerald Edelman and Giulio Tononi, it might not be unreasonable to suggest that a similar functional loop between analogous structures for relaying sensory input and those for different kinds of memory is also a critical contingency of consciousness in certain nonmammalian species. Although the technical hurdles represented by the task of demonstrating the functional neuroanatomical and neurophysiological substrates of consciousness in cephalopods are formidable, there is sufficient evidence of sophisticated sensory and cognitive capabilities to suggest that a search for conscious states is warranted.

Over many decades, a number of behavioral studies have probed the cognitive capacities of cephalopods. Among the coleoid cephalopods, it is the rich cognitive capacity of octopuses that has been most thoroughly documented. For example, the pioneering work of J.Z. Young, as well as that of Martin Wells and Stuart Sutherland, has amply demonstrated the ability of the octopus to make sophisticated discriminations between different objects based on size, shape, and intensity. Perhaps suggestive of more general properties of learning in complex nervous systems, Sutherland noted that octopuses classify differently shaped objects in the same manner as vertebrates such as goldfish and rats. More recent work on octopus learning and problem solving, such as the studies of Graziano Fiorito and his colleagues, have involved such tasks as finding the correct path to a reward in a plexiglas maze or retrieving an object from a clear bottle sealed with a plug.

The most impressive capabilities of cephalopods have generally reflected an extremely plastic behavioral repertoire and, on the basis of reports of observational learning by Fiorito and Scotto, highly developed attentional and memory capacities. Researchers, including Veronique Agin, Raymond Chichery, and Graziano Fiorito, have documented evidence of distinct capacities for short-term and long-term memory in both the octopus and the cuttlefish. For example, in experiments in which an octopus was exposed to a maze containing obstacles that were changed ad libidum by the researcher, the animal could remember these changes and maneuver around shifting obstacles accordingly. Notably, the authors of this study, Tohru Moriyama and Yukio-Pegio Gunji, asserted that the octopus appeared to first consider the layout of the maze before proceeding. The formidable quality of octopus memory was also shown beautifully by the work of Fiorito and Scotto, who demonstrated the ability of these animals to solve problems through observational learning, and not merely through mimicry.

Perhaps the most common task employed by researchers to assess the cognitive capacity of octopuses involves the presentation of a prey species, such as a crab, enclosed in a plugged jar. Commonly, as Fiorito and his colleagues have reported, the octopus's first reaction to the jar is an immediate attack, as if it is confronting freely moving prey. In short order, after repeated exposure to the crab-containing jar, the octopus learns to pull the plug and retrieve the crab. Prior exposure to the empty jar does not significantly reduce the time it takes the octopus to retrieve the crab, which might suggest that perhaps there is no transitive component in octopus learning. Moreover, it is notable that, even as the octopus appears to solve the prey-in-a-jar problem, it nevertheless shows strong, species-specific, predatory behavior – the attack – which persists over repeated trials. Such invariance in the octopus's behavioral repertoire, even as it solves the prey-in-a-jar problem, suggests the absence of

the kind of inhibitory pathways that might be activated in a mammal's brain as it becomes familiar with a particular task. Although such lack of inhibition does not necessarily bear on the possible presence or absence of conscious states in octopuses, it does suggest a nervous system markedly different – and perhaps more primitive – functionally than that of a mammal.

The question of what the necessary conditions for cephalopod consciousness might be has not yet been addressed explicitly. It is conceivable that, if cephalopod neuroanatomy becomes sufficiently well characterized to identify functional analogues of mammalian visual cortex, it might be possible to design and perform experiments similar to those involving the 'commentary key' paradigm employed by Cowey and Stoerig on Rhesus macaques. The cephalopod chromatophore – the organ system comprising pigment saccules surrounded by radial muscles in the skin – is employed by cephalopods for what researchers Roger Hanlon and John Messenger have referred to as 'adaptive coloration' (i.e., camouflage, mimicry, sexual display), but could perhaps just as easily be co-opted as a channel for accurate report in such experiments. Perhaps an even more accessible type of psychophysical experiment, such as the presentation of a serial stream of stimuli with an embedded stimulus that can be temporally shifted in or out of perception (the so-called attentional blink), could be designed around the cephalopod chromatophore as a channel for reporting what part of the stimulus stream the animal is, or is not, aware of. One formidable challenge in the design of such an experiment has to do with the necessity of presenting streams of stimuli to animals by video to allow for the kind of rapid switching of stimulus type that would otherwise be impossible in a live presentation. Most cephalopods, including octopuses, are color blind, as John Messenger, Nadav Shashar, Thomas Cronin, and others have noted, and there is increasing evidence that many of these animals actually perceive polarized light and are able to make subtle discriminations on the basis of this capacity. Unfortunately, many video monitors currently available (e.g., LCD and CRT monitors) employ fluorescent backlighting and polarizing filters, both of which might create artifacts that could render video streams of stimuli effectively unintelligible to most cephalopods. While not an insurmountable technical problem by any means, a solution would have to be found before proceeding with a program of 'cephalopod psychophysics.'

Of all the cephalopods, the octopus has extraordinarily complex sensory receptors coupled to a nervous system that, normalized to body weight, rivals those of some vertebrates (smaller than those of birds and mammals, but larger than those of fish and reptiles, as noted by Roger Hanlon, John Messenger, Binyamin Hochner, Tal Shomrat, and Graziano Fiorito, among others). The brain of an adult octopus may contain between 170 million and 500 million cells, the vast majority of which are neurons. In contrast to the avian brain, which, despite early misprisions of nomenclature, closely resembles the mammalian brain, the organization of the cephalopod nervous system is utterly alien, and presents a profound set of problems for identifying necessary structural correlates of systems underlying consciousness. In this regard, the search for a functional cephalopod analogue of the reentrant thalamocortical circuitry found in mammals will present a major challenge. Where would one begin to search for such circuitry? Biochemical and molecular techniques that have been tremendously useful in identifying functional vertebrate neuroanatomies might be deployed to effectively address this question; certainly, at the level of constituent neurons and receptor cells, the similarities between the cephalopod nervous system and its vertebrate counterpart are clear. Indeed, the resemblance between giant squid axons and synapses and those of vertebrate neurons was an early watershed in modern neurophysiology and allowed many of the most important investigations of neuronal physiology that followed, including those by Alan Hodgkin, Andrew Huxley, Bernard Katz, Ricardo Miledi, and John Zachary (J.Z.) Young.

Notwithstanding the profound gaps in our understanding of the functional circuitry of the cephalopod nervous system, some functional properties of neuroanatomy have been established over several decades. In the brain of at least one genus of squid, *Loligo*, Linda Maddock and J.Z. Young reported that at least 30 distinct nucleus-like lobes have been identified so far. The largest of these lobes, the optic lobes (which contain as many

as 65 million neurons in some species of octopus), handle visual processing and memory establishment as well as some higher motor control. Although the optic lobes are not organized like the laminar sheets of the mammalian cortex, it has been suggested by Hanlon and Messenger that these lobes are functionally analogous to the vertebrate forebrain. Starting with the early work of J.Z. Young, the vertical, superior frontal, and inferior frontal lobes were also found to play important roles in the establishment of memories. In experiments by Fiorito and Chichery in which the vertical lobe of *Octopus vulgaris* was lesioned, the ability of animals to learn visual discriminations was severely impaired, whereas the consolidation of long-term memories remained unaffected. In another study by Wells and Young, when the median inferior frontal lobe was removed, learning was compromised due to memory impairment. If the faculty of recall was, indeed, what was affected in the foregoing studies, this would suggest that some regions of the octopus frontal and vertical lobes are functionally similar to regions of mammalian cortex.

Carla Perrone-Capano and her colleagues have noted that considerably more genetic expression occurs in the squid nervous system than in other tissues or organs – an observation that is consistent with the notion that selection for highly complex nervous systems is phylogenetically ancient. It is therefore plausible that complex brains capable of rich behavioral repertoires began evolving in two very different lineages between 530 and 540 million years ago.

Perhaps the most daunting quality of the cephalopod nervous system is its radically different overall organization. The peculiar parallel distributed nature of the octopus locomotor system is perhaps the signal exemplar of the alien character of the cephalopod nervous system. J.Z. Young has noted that the density of neurons located in the tentacles of the octopus, taken together, exceeds the total number of neurons in the central fused ganglia of the brain itself. Recent work by German Sumbre, Yoram Gutfreund, Tamar Flash, Graziano Fiorito, and Binyamin Hochner has shown that a detached octopus arm could be made to flail in a realistic manner when stimulated with short electrical pulses. Although the presence of semiautonomous motor programs is not unique to this species, or even to cephalopods in general, there is ample evidence for the existence of central pattern generators (CPGs) in many invertebrates and vertebrates alike; in the case of vertebrates, CPGs seem to be located in the spinal cord, that is, within the central nervous system itself. Thus, it is not possible to produce the full suite of coordinated movements of locomotion in a detached vertebrate limb. This distinction makes the sophistication of the semiautonomous neural networks in octopus tentacles and their local motor programs all the more striking. Given this observation, notions of embodiment and bodily representation that might bear on the assessment of consciousness might necessarily be different for octopus than those set forth by cognitive scientists and philosophers when considering the vertebrate case.

The identification of higher levels of neural organization in cephalopods poses even more profound challenges. Such functional concepts as cell assemblies, modules, cortical columns, thalamocortical loops, cytochrome oxidase-labeled blobs (first characterized in the cortex by Margaret Wong-Riley), and neuronal groups have been copiously and variously defined in mammals as groups of cells that share similar structure and/or function and populate in large numbers certain defined regions of the brain (i.e., cortex, nuclei, and ganglia). Cortical columns, as described more than 50 years ago by Vernon Mountcastle, might be the smallest functional modules of the mammalian neocortex. Esther Leise's notion of the neural module comprises large concentrations of neuropil in invertebrates, and is perhaps functionally similar to the so-called minicolumn concept elaborated upon by Mountcastle and others. Although there is no experimental validation of such an invertebrate neural module, the concept might nevertheless provide a useful roadmap in the search for functional elements in the cephalopod brain. Certainly, the discovery of such neural modules might indicate an organization of brain circuitry that has functional properties similar to the mammalian cortex. At present, though, there is little insight into the organization of functional maps (perhaps akin to the topographic representations found in the mammalian cerebral cortex) in the cephalopod nervous system.

Since cephalopod brains have been shown to contain many of the same major neurotransmitters that are found in the brains of mammals, it is conceivable that, as is now the case with functional avian neuroanatomy, it will eventually be possible to use immunohistochemistry and genetic manipulation to characterize those areas of the cephalopod brain that are analogous in function to neural regions showing correlated activity during waking, conscious behavior in mammals.

Given the success of the cephalopod radiation and its occupation of so many marine niches, it is not surprising that there is substantial interspecies variation in sensorimotor capacities and behavioral repertoires. The depth and breadth of known ecological adaptations, as well as the emergence of large numbers of cells and cell-types and dense enough connectivity in certain species, at the very least suggests entirely plausible and sufficient preconditions for the appearance of the rich discriminatory capacities necessary for consciousness to emerge in some cephalopod species.

Although cephalopod neurophysiology has only recently come into its own, perhaps the most suggestive finding in favor of precursor states of consciousness in at least some members of Cephalopoda is the demonstration of EEG patterns, including event-related potentials, that, as researchers Theodore Bullock and Bernd Budelmann have noted, look quite similar to those in awake, conscious vertebrates. This has been demonstrated in at least one other invertebrate, the fruit fly, by Bruno Van Swinderen and Ralph Greenspan. An obvious prerequisite to identifying cephalopod EEG patterns that reflect the signature of fast irregular activity, similar to that observed in human conscious states, will be to determine precisely where to record from. On the basis of earlier studies, it is possible that the optic, vertical, and superior lobes of the octopus brain are relevant candidates and that they may function in a manner analogous to mammalian cortex. In any event, they appear to be among the substrates of octopus learning and memory. Indeed, recent studies of slice preparations of octopus vertical lobe by Hochner, Fiorito, and colleagues have identified long-term potentiation (LTP) of glutamatergic synaptic field potentials similar to that found in vertebrates. Most intriguing is the finding by Hochner, Shomrat, and Fiorito that, unlike the vertebrate form of neural plasticity, LTP in the octopus vertical lobe is NMDA independent, suggesting an entirely convergent evolution of this form of neural plasticity.

Although we are far from making a strong case for sensory consciousness in cephalopods such as the octopus, present evidence from behavior, neuroanatomy, and neurophysiology does not preclude this possibility. Moreover, and tantalizingly, a variety of techniques are approaching a stage of maturity that holds the promise of substantive investigations of the necessary conditions for consciousness in these animals so phylogenetically distant from us.

Conclusion: Complex Nervous Systems, Evolutionary Convergence, and Universal Properties of Consciousness

If consciousness did indeed arise in radically different nervous systems of animals such as mammals and birds on the one hand and cephalopods on the other, then the possibility that certain properties of consciousness are universal becomes highly likely. What might the implications of this be? A very important implication is that the elaboration of a capacity for higher-order discrimination, such as that which might underlie the process of consciousness, can arise in different kinds of neural substrates so long as the same functional interactions are in play within these substrates. In effect, the organization of structures, either over the course of development or in the sense of overall gross anatomy, may actually be secondary to issues of functional identity, that is, such elements of conscious states as the reentrant circuitry proposed by Edelman and Tononi, linking areas specialized for perception and those specialized for memory, which seems to be critical for mammalian consciousness (i.e., thalamocortical loops), could well have been instantiated a number of times in tissues with very different embryological – and evolutionary – histories. Finally, if consciousness indeed appeared independently at least three times during animal evolution, it seems likely that increasingly complex ecologies have

tended to drive the nervous systems of highly motile animals – major structural differences notwithstanding – inexorably toward the capacity for higher-order discriminations and, ultimately, conscious experience.

Acknowledgments

The work of the author was supported by the Neurosciences Research Foundation.

See also: Ethical Implications: Pain, Coma, and Related Disorders.

Suggested Readings

Aboitiz F (1999) Comparative development of the mammalian isocortex and the reptilian dorsal ventricular ridge: Evolutionary considerations. Cerebral Cortex 9: 783–791.
Agin V, Dickel L, Chichery R, and Chichery MP (1998) Evidence for a specific short-term memory in the cuttlefish, Sepia. Behavior Processes 43: 329–334.
Ayala-Guerrero F (1989) Sleep patterns in the parakeet Melopsittacus undulates. Physiology and Behavior 46(5): 787–791.
Bachmann T, Breitmeyer B, and Ögmen H (2007) Experimental Phenomena of Consciousness. Oxford: Oxford University Press.
Boycott BB and Young JZ (1955) A memory system in Octopus vulgaris Lamarck. Proceedings of the Royal Society of London. Series B: Biological Sciences 143(913): 449–480.
Bullock TH (2003) Have brain dynamics evolved. Should we look for unique dynamics in the sapient species? Neural Computation 15: 2013–2027.
Bullock TH and Budelmann BU (1991) Sensory evoked potentials in unanesthetized unrestrained cuttlefish: A new preparation for brain physiology in cephalopods. Journal of Comparative Physiology 168(1): 141–150.
Buxhoeveden DP and Casanova MF (2002) The minicolumn hypothesis in neuroscience. Brain 125: 935–951.
Cheney DL and Seyfarth RM (1990) How Monkeys See the World: Inside the Mind of Another Species. Chicago: University of Chicago Press.
Cowey A and Stoerig P (1995) Blindsight in monkeys. Nature 373(6511): 247–249.
Darwin C (2004 [1871]) The Descent of Man, and Selection in Relation to Sex. London: Penguin Classics.
Darwin C (1898 [1872]) The Expression of the Emotions in Man and Animals. New York: D. Appleton and Company.
Edelman GM (1987) Neural Darwinism. New York: Basic Books.
Edelman GM (1988) Topobiology: An Introduction to Molecular Embryology. New York: Basic Books.
Edelman GM (1989) The Remembered Present. New York: Basic Books.
Edelman GM (2004) Wider than the Sky: The Phenomenal Gift of Consciousness. New Haven: Yale University Press.
Edelman DB, Baars BJ, and Seth AK (2005) Identifying hallmarks of consciousness in non-mammalian species. Consciousness and Cognition 14(1): 169–187.
Fiorito G and Chichery R (1995) Lesions of the vertical lobe impair visual discrimination learning by observation in Octopus vulgaris. Neuroscience Letters 192: 117–120.
Fiorito G, Biederman GB, Davey VA, and Gherardi F (1998) The role of stimulus preexposure in problem solving by Octopus vulgaris. Animal Cognition 1: 107–112.
Fiorito GP and Scotto P (1992) Observational learning in Octopus vulgaris. Science 256: 545–574.
Gallup GG, Jr (1970) Chimpanzees: Self recognition. Science 167(3914): 86–87.
Gardner RA and Gardner BT (1969) Teaching sign language to a chimpanzee. Science 165(894): 664–672.
Grin DR (1976) The Question of Animal Awareness: Evolutionary Continuity of Mental Experience. New York: The Rockefeller University Press.
Grin DR and Speck GB (2004) New evidence of animal consciousness. Animal Cognition 7(1): 5–18.
Hanlon RT and Messenger JB (2002) Cephalopod Behavior. Cambridge: Cambridge University Press.
Hochner B, Brown ER, Langella M, Shomrat T, and Fiorito G (2003) A learning and memory area in the octopus brain manifests a vertebrate-like long-term potentiation. Journal of Neurophysiology 90(5): 3547–3554.
Hochner B, Shomrat T, and Fiorito G (2006) The octopus: A model for a comparative analysis of the evolution of learning and memory mechanisms. Biological Bulletin 210: 308–317.
Jarvis ED, Ribeiro S, Da Silva ML, Ventura D, Vielliard J, and Mello CV (2000) Behaviorally driven gene expression reveals song nuclei in hummingbird brain. Nature 406: 628–632.
Kardong KV (1995) Vertebrates: Comparative Anatomy, Function, and Evolution. Dubuque, IA: W.C. Brown.
Karten HJ (1997) Evolutionary developmental biology meets the brain: The origins of mammalian cortex. Proceedings of the National Academy of Sciences of the United States of America 94: 2800–2804.
Karten HJ, Hodos W, Nauta WJ, and Revsin AM (1973) Neural connections of the visual wulst'' of the avian telencephalon: Experimental studies in the pigeon (Columbia livia) and owl (Speotyto cunicularia). Journal of Comparative Neurology 150(3): 253–278.
Leise EM (1990) Modular construction of nervous systems: A basic principle of design for invertebrates and vertebrates. Brain Research. Brain Research Reviews 15(1): 1–23.
Logothetis NK (1998) Single units and conscious vision. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 353: 1801–1818.
Luo M and Perkel DJ (2002) Intrinsic and synaptic properties of neurons in an avian thalamic nucleus during song learning. Journal of Neurophysiology 88: 1903–1914.
Luo M, Ding L, and Perkel DJ (2001) An avian basal ganglia pathway essential for vocal learning forms a closed topographic loop. Journal of Neuroscience 21: 6836–6845.

Mather J (2008) Cephalopod consciousness: Behavioral evidence. Consciousness and Cognition 17(1): 37–48.
Medina L and Reiner A (2000) Do birds possess homologues of mammalian primary visual, somatosensory and motor cortices. Trends in Neuroscience 23(1): 1–12.
Merker B (2005) The liabilities of mobility: A selection pressure for the transition to consciousness in animal evolution. Consciousness and Cognition 14: 89–114.
Messenger JB (2001) Cephalopod chromatophores: Neurobiology and natural history. Biological Reviews 76: 473–528.
Moriyama T and Gunji YP (1997) Autonomous learning in maze solution by octopus. Ethology 103(6): 499–513.
Naftolin F, Horvath T, and Balthazart J (2001) Estrogen synthetase (Aromatase) immunohistochemistry reveals concordance between avian and rodent limbic systems and hypothalamus. Experimental Biology and Medicine 226: 717–725.
Pepperberg IM (1998) Possible perceptual consciousness in grey parrots (Psittacus erithacus). American Zoologist 35 (5): 7A.
Premack D and Premack AJ (1984) The Mind of an Ape. New York: W.W. Norton & Company.
Puelles L (2001) Thoughts on the development, structure and evolution of the mammalian and avian telencephalic pallium. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 356(1414): 1583–1598.
Sanders GD (1975) The cephalopods. In: Corning WC, Dyal JA, and Willows AOD (eds.) Invertebrate Learning. New York: Plenum Press.
Savage-Rumbaugh S, McDonald K, Sevcik RA, Hopkins WD, and Rubert E (1986) Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (Pan paniscus). Journal of Experimental Psychology. General 115(3): 211–235.
Savage-Rumbaugh S, Sevcik RA, Rumbaugh DM, and Rubert E (1985) The capacity of animals to acquire language: Do species differences have anything to say to us. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 308(1135): 177–185.
Seth AK and Baars BJ (2005) Neural Darwinism and consciousness. Consciousness and Cognition 14(1): 140–168.
Seth AK, Baars BJ, and Edelman DB (2005) Criteria for consciousness in humans and other mammals. Consciousness and Cognition 14(1): 119–139.
Seyfarth RM and Cheney DL (2003) Meaning and emotion in animal vocalizations. Annals of the New York Academy of Sciences 1000: 32–55.
Skinner BF (1938) The Behavior of Organisms: An Experimental Analysis. New York: Appleton-Century Company, Inc.
Srinivasan R, Russell DP, Edelman GM, and Tononi G (1999) Increased synchronization of neuromagnetic responses during conscious perception. Journal of Neuroscience 19: 5435–5448.
Sutherland NS (1969) Shape discrimination in rat, octopus, and goldfish: A comparative study. Journal of Comparative Physiology Psychology 67: 160–176.
Tononi G and Edelman GM (1998) Consciousness and complexity. Science 282: 1846–1851.
Van Swinderen B and Greenspan RJ (2003) Salience modulates 20–30 Hz brain activity in Drosophila. Nature Neuroscience 6(6): 579–586.
Wada K, Hagiwara M, and Jarvis ED (2001) Brain evolution revealed through glutamate receptor expression profiles. Society for Neuroscience Abstract Vol. 27.
Watson JB (1913) Psychology as the behaviorist views it. Psychological Review 20: 158–177.
Weir AAS, Chappell J, and Kacelnik A (2002) Shaping of hooks in New Caledonian crows. Science 297: 981.
Wells MJ and Young JZ (1969) The effect of splitting part of the brain or removal of the median inferior frontal lobe on touch learning in octopus. The Journal of Experimental Biology 56(2): 381–402.
Wells MJ and Young JZ (1972) The median inferior frontal lobe and touch learning in the octopus. The Journal of Experimental Biology 56(2): 381–402.
White SA, Fisher SE, Geschwind DH, Scharff C, and Holy TE (2006) Singing mice, songbirds, and more: Models for FOXP2 function and dysfunction in human speech and language. The Journal of Neuroscience 26(41): 10376–10379.
Young JZ (1961) Learning and discrimination in the octopus. Biological Reviews 36: 32–96.
Young JZ (1971) The Anatomy of the Nervous System of Octopus Vulgaris. Oxford: Clarendon Press.
Young JZ (1983) The distributed tactile memory system of octopus. Proceedings of the Royal Society of London. Series B: Biological Sciences 218: 135–176.

Biographical Sketch

Dr Edelman is an associate fellow in Experimental Neurobiology at The Neurosciences Institute, in San Diego, California, where his laboratory is engaged in the study of mitochondrial dynamics in the mammalian brain. He has longstanding interest in the behavioral, neuroanatomical, and neurophysiological correlates of animal consciousness. Recently, he has begun to collaborate with Dr Graziano Fiorito (Stazione Zoologica, Naples, Italy) on a psychophysical approach to studying octopus cognition. He is an active member of the Association for the Scientific Study of Consciousness and, together with Drs Anil Seth and Bernard Baars, has published a number of papers on consciousness, most recently in the journal Consciousness and Cognition.

Artificial Intelligence and Consciousness

O Holland and D Gamez, University of Essex, Colchester, UK

Glossary

Artificial general intelligence (AGI) – A system with AGI is capable of wide-ranging human-like intelligence that is not limited to the performance of a single task.

Artificial intelligence (AI) – Systems with AI are capable of intelligent reasoning and behavior, and often have some form of learning mechanism.

Frame problem – Every action has a large number of potential consequences and it can be difficult or impossible for a robot to know whether it has worked through enough of them to accurately update its knowledge of the world.

Global workspace theory – A cognitive architecture that accounts for the difference between consciousness and unconsciousness.
A number of specialized processors receive information from the global workspace, process it, and compete to output their result back to the global workspace, with the cycle being repeated until the problem is solved or a new problem is presented.

Information integration – In a system with high information integration one element can influence many other elements in the system; in a system with low information integration the elements are mostly isolated from each other.

Intentional state – Intentional states are states of the mind that are directed toward or about objects, properties or states of affairs.

Machine consciousness – A research area that focuses on the development of machines with the behavior, cognitive characteristics and/or architecture associated with consciousness. This type of work may lead to the development of systems that are phenomenally conscious.

Synthetic phenomenology – A research area that attempts to decide whether artificial systems are capable of conscious states and tries to describe the phenomenology of these states if they occur.

Turing test – A test in which an interrogator communicates with a hidden person and a computer and attempts to identify which is which. A computer is said to have passed the Turing test if it cannot be distinguished from the person.

Introduction

At the most general level artificial intelligence (AI) is concerned with the development of systems that are capable of intelligent reasoning and behavior, and often include some form of learning mechanism. AI systems may be standalone computer programs or they may process data from the world and produce behavioral output using a robot body. After a rather shaky start AI is now a major research area and it is used to perform tasks ranging from the processing of security camera data to the control of autonomous vehicles. During its development AI has influenced our theories about consciousness and provided a good way of testing them, and there has been an extensive debate about whether consciousness can be created in a machine. Several different ways of testing for consciousness in artificial systems have been put forward and the research on machine consciousness has raised a number of ethical and legal issues.

An Overview of Artificial Intelligence

Research on AI is generally considered to have started in 1956 when a conference was held at Dartmouth with a number of the key figures. The early computer programs represented problems using logical symbols and solved them by carrying out manipulations that were defined and constrained by explicit formal rules. Some of this work was inspired by studying how humans claimed to solve similar problems, and the underlying idea that the intelligence of minds and machines could have the same basis was articulated as the physical symbol system hypothesis, which claims that a physical symbol system has the necessary and sufficient means for general intelligence. Although it was not explicitly a theory of consciousness, the physical symbol system hypothesis was held to account for a variety of cognitive abilities necessary for intelligence. Many of these cognitive abilities, such as understanding, would ordinarily be thought to imply some degree of consciousness.

While the logic-based approach to AI had some initial success, by the early 1970s it was starting to run into difficult problems. To begin with, the programs that had been developed up to that point could only handle basic examples of the problems that they were supposed to represent, and the computers of that time did not have enough memory or processing speed to accomplish anything useful or to process significant amounts of data from the world. It also became clear that although computers

could solve problems that humans found difficult, such as theorem proving, it was very difficult to program them to carry out tasks that humans found easy, such as object recognition and manipulation. A third problem that emerged was that our interactions with the world depend on a background of common sense knowledge that constrains our selection of actions in a particular situation. Manually programming all of this knowledge into a computer would have been a large and almost impossible task and there was no other obvious way in which an artificial system could gain this type of information. Logic-based AI also encountered a difficulty known as the 'frame problem,' which is that even the most basic actions have a large number of potential consequences. It often takes a long time to calculate these potential consequences and it can be impossible for a robot to know when it has worked through enough of them to accurately update its knowledge of the world.

A number of researchers responded to these problems by abandoning the logic-based approach in favor of robots that grounded their representations directly in the real world using sensors and actuators. The behavior of these systems was controlled by modules that processed the sensory data and produced a motor output, and these modules could communicate through the world and inhibit each other (to manage conflict resolution). The modules were typically organized in a series of layers connecting perception to action, with the higher layers taking the output of the lower layers into account – an arrangement known as the subsumption architecture. One problem with the robotics approach was that the difficulties and expense of working with real robots were so severe that many people used robot simulations, which often introduced misleading simplifications, such as a lack of noise.

A second response to the problems with logic-based AI was for researchers to focus on well-specified problems within a particular domain that could be addressed more easily, such as chess playing or data mining. This type of specialized AI is often contrasted with work that is attempting to replicate human intelligence completely, which is known as artificial general intelligence (AGI). More narrowly focused AI has developed a number of technologies and had some impressive success – for example, a computer beat Gary Kasparov at chess in 1997 – and machine learning is now used routinely in areas ranging from computer games to stock market prediction.

The rapid increase in computer power in recent years has made AI easier to implement and it has now developed a range of technologies that include artificial neural networks, agent systems and a variety of learning and reasoning techniques. Contemporary AI has also carried out some biologically inspired work based on simpler creatures, such as ants and salamanders. The resurgence of interest in consciousness in recent years has inspired a number of researchers to test theories of consciousness using computer models, and there has been some speculation that this could eventually lead to more intelligent machines that might actually have phenomenal states. This type of research is gradually becoming known as 'machine consciousness,' although 'artificial consciousness' and occasionally 'digital sentience' have also been used to describe it.

Artificial Intelligence and Consciousness Research

The Influence of Artificial Intelligence on Theories about Consciousness

The early work on AI had a significant impact on our thinking about the mind because it highlighted the limitations of a purely logic-based approach and pointed to a number of issues, such as the frame problem and the background of common sense, that any thinking creature would have to solve. The work on robotics suggested that there is an important link between embodiment and intelligence, and this led researchers to speculate that physical embodiment may be necessary for consciousness.

One of the most direct links between AI and theories about consciousness is Bernard Baars' global workspace architecture, which is based on the blackboard architecture that was developed in the 1970s as part of research on AI. What the two have in common is that a large number of specialized processors analyze the same problem data and compete to output their result back to the blackboard/global workspace, with the cycle being repeated until the problem is solved or a new problem is presented. However, within global workspace theory the problem data are actively distributed to all processors by a broadcast mechanism, instead of using a blackboard that all processors can inspect, and global workspace theory compares the outcome of the competitive selection process to a spotlight on the stage of a darkened theatre, which lights up the message of the winning processor for the 'audience' of all the processors. While the blackboard architecture is an AI technology that was designed to solve specific problems, global workspace theory was put forward by Baars as a cognitive model of consciousness in which the spotlit broadcast message forms the contents of consciousness and the operations of the processors are unconscious.

Applications of Artificial Intelligence to Research on Consciousness

The work that has been carried out in AI, computer science and robotics can be used to test theories of consciousness by building models. This type of test is a good way of making a theory's assumptions explicit since thought experiments can leave all kinds of things out, but a model will only work when everything necessary

has been included. While the usefulness of this approach is often limited by the fact that experimental systems are very basic, this type of modeling provides a very good way of bridging the gap between abstract theories and the real world and it can help us to understand how consciousness is implemented in the brain.

A further benefit of this type of work is that it can help us to clarify our intuitions about consciousness. If someone claims that X, Y, and Z are necessary and sufficient for consciousness, then it is generally possible to develop a machine that has X, Y, and Z. If X, Y, and Z are functions, then the processing of X, Y, and Z in the machine can be done by any form of computation, and opponents of function-based theories of consciousness have pointed out that computation can be carried out by the manipulation of beer cans or by the population of China communicating with mobile phones and satellites. While it might have been easy to believe that X, Y, and Z were necessary and sufficient for consciousness in human beings, it is much harder to believe that these eccentric implementations are really conscious. Within AI many systems have been developed that have attributes associated with consciousness, such as imagination, a global workspace architecture or information integration (see later discussion), and yet researchers are wary of attributing phenomenal consciousness to them. This suggests that the presence of biological neurons has a big influence on our intuitions about consciousness, although the absence of consciousness in deep sleep indicates that this is not sufficient.

AI has also been used in consciousness research to specify the contents of conscious experiences. It is generally acknowledged that infants and animals are likely to have very different phenomenal states from humans, and it is difficult or impossible to say in human language what it is like to be an infant or an animal. AI technology can help with this problem because it enables us to build systems whose states, interactions, and capacities can be used to describe the phenomenology of very young humans or nonhuman creatures. For example, some theories of consciousness link conscious states to our expectations about what will appear next in the visual field. The possible phenomenology of this type of system has been specified by the construction of a robot-based device, whose perceptions were based on its expectations about visual input, and not on the actual visual data that it received.

The Debate about Consciousness in Artificial Systems

Positive Theories about Machine Consciousness

Naturalistic accounts of consciousness

Many people believe that the brain is a complex mechanism that came about through natural selection and is governed by well-understood principles that apply to all other physical processes in living things. If there is nothing special about the brain compared to the liver, for example, then once we have understood how consciousness works in the brain, we should be able to produce consciousness in other systems as well. Up to this point science has not identified anything in the brain that is correlated with consciousness and would be impossible to put into an artificial system. If such magical ingredients cannot be found, then it seems reasonable to suppose that silicon chips, wires, motors, and cameras could be assembled into a conscious robot without leaving anything out that is necessary for consciousness – although it might be difficult to match the processing speed and complexity of biological systems.

Information theories of consciousness

The information integration theory of consciousness was put forward by Giulio Tononi, who claims that the conscious parts of a system have the greatest capacity to integrate information. Tononi also maintains that the amount of consciousness in a system depends on the amount of information integration. Information integration is measured using the value Φ, which is calculated by dividing the system up in many different ways and evaluating the causal influence of the parts on each other. Tononi's claim about a link between consciousness and information integration is entirely independent of the material out of which the system is made, and so any system that is capable of integrating information is conscious to some degree according to this theory. The information integration theory of consciousness has some overlap with David Chalmers' claim that conscious experiences are realizations of information states, which predicts that systems as simple as thermostats are conscious because they process information.

Cognitive theories

Many people have conjectured that consciousness may be linked to cognitive characteristics, such as emotions, imagination, and a model of the self. If consciousness depends on functions at the cognitive level, then it should be possible to realize it on any piece of hardware that is capable of carrying out the appropriate processing. One example of a cognitive theory of consciousness is the axiomatic theory of Igor Aleksander, who claims that imagination, emotion, depiction, volition, and attention are minimally necessary for consciousness, and any natural or artificial system that implements these axioms is judged to be conscious according to this theory. Global workspace theory is another cognitive architecture that can be implemented on many different types of hardware, and Thomas Metzinger's phenomenal self model and constraints on conscious experience are also largely independent of the physical system.

Other theories

Many other theories of consciousness have positive implications for the possibility of creating consciousness in artificial systems. To begin with, the pantheist claim that all matter is conscious to some degree suggests that computers and robots are conscious even when they are switched off. David Rosenthal's claim that consciousness depends on a higher-order thought about another mental state is not linked to any particular implementation, and a number of connections have been made between consciousness and virtual machines that can run on any type of hardware. Some researchers have claimed that inner speech helps to constitute our sense of self and agency and may be important to consciousness as well. If this is the case, it might be possible to use the work on language acquisition in AI to develop conscious systems.

Criticisms of Machine Consciousness

The hard problem of consciousness

People working on consciousness commonly distinguish between the easy problem of explaining how we can discriminate, integrate information, report mental states, focus attention, etc., and the hard problem of explaining phenomenal experience. Although solving the 'easy' problem is far from easy, we do at least have some idea how it could be done, but it can be argued that we have no real idea about how to solve the hard problem. If we do not understand how human consciousness is produced, then it makes little sense to try to make robots phenomenally conscious.

Objections to machine consciousness based on the hard problem may only be a caution about any current claims that we might want to make, because we may solve the hard problem in the future and discover that machines are capable of consciousness. Furthermore, it can be argued that asking questions about phenomenal states in machines and building models of consciousness are likely to improve our understanding of human consciousness and take us closer to a solution to the hard problem. It might also be possible to create conditions that allow consciousness to emerge in a system without understanding the causes of phenomenal states – for example, it has been suggested that consciousness could emerge in a detailed simulation of a human infant that develops by interacting with its environment.

The hard problem of consciousness only becomes problematic for work on machine consciousness if it can never be solved, which would make it difficult to make strong claims about the consciousness of artificial systems. This is the position of Colin McGinn, who argues that consciousness depends on a natural property, P, that cannot be grasped by human beings. If McGinn is right, we will never know if machines can have property P and we will be unable to measure P in the systems that we have built.

The Chinese room

One of the most influential attacks on the idea that AI could develop a mind was John Searle's Chinese Room thought experiment in which a person in a room receives Chinese characters, processes them according to a set of rules, and passes out the correct result without understanding what the characters mean. This processing of characters could be used to create the external behavior associated with consciousness, to simulate the cognitive characteristics associated with consciousness, or to model a conscious architecture. However, Searle argues that the person processing characters in the room would not understand the objects represented by the Chinese characters or have intentional states about them, and so the Chinese Room would never become a real mind. Although consciousness as such barely featured in the original paper, it (or its absence) has since become a key feature of the Chinese Room scenario, which is now seen by many of its proponents as an attack on the idea that a computer executing a program could become conscious, rather than merely intelligent.

One response to this argument is that the Chinese Room could be grounded in some kind of nonsymbolic lower level, such as images or sounds, which would give it the ability to understand the characters that it is manipulating and have intentional states directed toward the nonsymbolic content. This type of grounding could be achieved by building an embodied system that processed data from the real world, and neural models have been proposed as a way of grounding higher level symbolic representations in sensory inputs. A second objection to the Chinese Room argument is that both brains and computers are physical systems assembled from protons, neutrons, and flows of electrons, and Searle is happy to claim that consciousness is a causal outcome of the physical brain. Until the hard problem of consciousness has been solved, we will be unable to say whether the physical computer and the physical brain are different in a way that is relevant to consciousness, and since we have no idea about this at present, the Chinese Room argument does not offer any *a priori* reason why the arrangement of protons, neutrons, and electrons in a physical computer is less capable of consciousness than the arrangement of protons, neutrons, and electrons in a physical brain.

Consciousness is not algorithmic

Machine consciousness has been criticized by Roger Penrose, who claims that the processing of an algorithm is not enough to evoke phenomenal awareness because subtle and largely unknown physical principles are needed to perform the noncomputational actions that lie at the root of consciousness. If consciousness does something that 'mere' computation cannot, then it cannot be simulated by a computer and phenomenal states cannot be created in a machine. The most straightforward response

to Penrose is to reject his theory of consciousness, which has been heavily criticized by a number of people. However, even if Penrose's theories about consciousness are accepted, it may still be possible to develop some kind of quantum computer that incorporates the type of physical action that is linked by Penrose to consciousness.

The limitations of artificial intelligence

It has already been pointed out that AI research encountered deep problems when it attempted to implement intelligence using logical symbols manipulated by rules, and Hubert Dreyfus has argued that this attempt failed because human intelligence depends on skills, body, emotions, imagination, and other attributes that cannot be encoded into long lists of facts. While this is a good objection to using logic to try to develop systems that are as intelligent as humans in real world situations, there is no reason why machine consciousness could not be pursued in more limited ways independently of this objective. For example, some of the behaviors that require consciousness in humans could be created in a simple way, and imagination and emotion can be simulated without the expectation that they will work as effectively as human cognitive processes. It can also be argued that the work being carried out on imagination, emotions, and embodiment in machine consciousness addresses some of the areas that are lacking in traditional approaches to AI.

Research on Machine Consciousness

Current Research

Machines with the external behavior associated with consciousness

If consciousness is the result of natural selection, then the possession of consciousness must have changed organisms' behavior in a way that gave them an evolutionary advantage and it should be possible to list the behaviors associated with consciousness. One problem with identifying a definitive list is that many waking behaviors can be carried out consciously and nonconsciously, which makes it difficult to decide whether the behavior is associated with consciousness or not. A commonly cited example is that we can drive home from work with our attention on other things, and patients suffering from epileptic automatism can carry out complex actions, such as diagnosing lung patients or cleaning guns, without any conscious awareness. While these examples show that humans can carry out a limited amount of complex behavior unconsciously, the stereotypical nature of this behavior suggests that more dynamic and interactive activities, such as interpersonal dialogue, can only be carried out consciously, and many new behaviors can only be learnt when consciousness is present.

Since there is not a clear specification of the behaviors associated with consciousness, relatively few people are explicitly working on this task, although there has been some recent work on a functional specification for a conscious machine that could eventually be implemented in computer code. Most of the research in this area is being carried out by people who are attempting to mimic human intelligence completely, either as part of work on AGI or in order to pass the Turing test (see later discussion). The work on AGI has been mostly logic-based, and to overcome the difficulties with this approach one company has been attempting to build a database with enough facts that would enable it to become generally intelligent, although it has had little success so far. There has also been some work on the development of systems with memory, reasoning, and learning that are being used to control virtual characters in online multiplayer environments. The dialogue systems that have been developed to pass the Turing test are often superficially convincing for short periods, but none of them have managed to pass the test yet.

Machines with the cognitive characteristics associated with consciousness

A number of connections have been made between consciousness and cognitive characteristics, such as imagination and emotions, and it has been suggested that an internal model of the self plays an important role in consciousness. The modeling of the cognitive characteristics associated with consciousness has been a strong theme in machine consciousness and a number of different approaches have been explored.

One approach to this area has been the simulation of the cognitive characteristics associated with consciousness using neural networks with tens of thousands of neurons. In one set of experiments a series of networks was used to model basic versions of emotion, imagination, volition, attention, and depiction; in other work, a neural network was developed that controlled the eye movements of a virtual humanoid robot and used models of imagination and emotion to decide whether it looked at a red or blue cube.

There has also been some research that has used simple wheeled robots and small recurrent neural networks to simulate imagination and internal models. One set of experiments started by training a neural network to move a robot around a simple maze. The network's predictions about the next sensory state were then connected up to its sensory input and the network was able to navigate 'blind' around the maze using its 'imagination.' A second group of researchers used a simple wheeled robot and a machine learning algorithm to develop a system that explored its environment and built up an internal model that could be graphically displayed. This work on internal modeling was taken further by the development of a system that used a

virtual robot as an internal model of a real humanoid robot. The virtual environment was updated with data from the real world and the system used this virtual model to imagine the consequences of different actions. When the virtual modeling suggested that an action would have a positive outcome, the behavior was executed by the real robot.

Machines with an architecture that is claimed to be a cause or correlate of human consciousness

A number of theories have been put forward about the architectures that are associated with consciousness. Most of the machine consciousness work in this area has either modeled Baars' global workspace or created brain-based models of the neural correlates of consciousness.

The most well-known global workspace model is the Intelligent Distribution Agent (IDA) naval dispatching system developed by Stan Franklin that was created to assign sailors to new billets at the end of their tour of duty. This task involves natural language conversation, interaction with databases, adherence to Navy policy, and checks on job requirements, costs, and sailors' job satisfaction. These functions were carried out using a large number of small programs that were specialized for different tasks and the whole system was organized using a global workspace architecture. There have also been a number of neural models based on global workspace theory. One neural network was developed with 40 000 neurons and a brain-inspired global workspace architecture that enabled a virtual robot to explore the consequences of potential actions and select the one that would result in a positive stimulus. Detailed neural models of the global workspace architecture have been used to model a number of psychological phenomena related to consciousness, such as the transitions between the awake state and sleep, anesthesia or coma.

Within research on the neural correlates of consciousness, a brain-based model with 100 000 neurons was created to study neural computation, attention, and consciousness. In other work a simulated neural network was developed with areas that were based on the brain's neuroanatomy. This network had connections between its motor and sensory areas, which some people have linked to consciousness, and it was used to control a virtual child whose 'survival' depended on its communications with a human operator.

Phenomenally conscious machines

The previous three approaches to machine consciousness are all relatively uncontroversial, since they are modeling phenomena linked to consciousness without any claims about real phenomenal states. The fourth area of machine consciousness is more philosophically problematic, since it is concerned with machines that might have real phenomenal experiences – machines that are not just tools in consciousness research, but are actually conscious themselves. This approach has some overlap with the other approaches, since in some cases it may be hypothesized that the reproduction of human behavior, cognitive states, or internal architecture will lead to real phenomenal experiences. On the other hand, phenomenal states might be achievable independently of other approaches to machine consciousness. For example, it might be possible to create a system based on biological neurons that was capable of phenomenal states, but lacked the architecture of human consciousness and any of its associated cognitive states or behaviors.

The lack of consensus about consciousness has made most AI researchers cautious about attributing real phenomenal states to their systems. There have not been any strong behavioral grounds for attributing consciousness to these systems either, although it has been possible to make limited predictions about phenomenal states in artificial systems using particular theories of consciousness.

Future Developments

Progress in machine consciousness is likely to be slow and incremental, with increasingly complex systems being developed that are based on the cognitive characteristics and architectures associated with consciousness. These systems will gradually display more human-like behavior and they will be increasingly likely to possess phenomenal states. There has also been speculation that the expansion of the Internet or the apparently inexorable increase in computer power will lead to conscious systems.

The number of computers linked together in the Internet (currently some hundreds of millions) and the bandwidth of connections are increasing every year, and so in principle the Internet can be regarded as an enormous and growing distributed computing system. Some people claim that, as the Internet grows further, it will achieve a size at which it will inevitably become conscious. There is no evidence from within AI suggesting that this should be the case, although it is worth noting that the information integration theory of consciousness claims that any system of active interconnected elements is potentially conscious. However, it will be difficult to tell which parts of the Internet are predicted to be conscious according to this theory because it is very difficult to calculate information integration in large networks, and there is only likely to be any potential for consciousness in the Internet as a whole when the data transfer rate between computers exceeds the data transfer rate within each computer.

For the last 50 years, the density with which transistors can be packed onto an integrated circuit has doubled approximately every 2 years. This exponential growth rate, known as Moore's law, is expected to continue for another decade or two, and this increase in packing density leads to proportionate increases in performance and decreases in cost. It has been claimed that by around 2019

a typical desktop computer will have the computational capacity of the human brain (although the way in which this has been calculated is certainly open to question), and will cost only $1000. If the key to producing an intelligent computer is sheer processing power, then whatever power is required will become available within a relatively short time frame. Assuming that AI software continues to improve (in some unspecified manner, but presumably able to take advantage of the computational resources), then some have claimed that it will be possible to build a machine that can pass the Turing test by 2029. For those happy to accept the Turing test as a test of consciousness as well as intelligence, this amounts to a prediction that an increase in the available computing power, combined with improved AI software, is all that is needed to produce a conscious machine. However credible the proponents of this argument may be as individuals, the argument itself is fatally weakened by the lack of specification of the principles on which the AI software will be built. A variant of this argument proposes that the available computing power could be used to run an exact simulation of a human brain, assuming that the relevant neuroscientific details will be available by then. Whether this will yield a conscious machine is a matter for computational neuroscience rather than AI.

A third possibility that is often discussed in connection with this area is the extension or enhancement of consciousness with technology. In a trivial sense this happens whenever a person looks through a pair of binoculars or listens to amplified music, but advocates of this view believe that we will eventually be able to plug external devices into our brains and thus become directly conscious of the Internet, for example, without the mediation of our senses. While this would be more of a technological extension of our consciousness than a direct application of AI, it is likely that some form of AI would play a role in this, either to control what would be a very complex interface or more fancifully as an advanced AI that we would fuse with through a brain interface. However, while implanted electrodes are able to provide very basic vision and can be used to control a robot arm, more advanced extensions of our consciousness remain entirely within the realm of science fiction.

Testing for Consciousness in Artificial Systems

Behavioral Tests

Turing tests

The best-known behavioral test for consciousness is the Turing test, which was put forward by Alan Turing in 1950 as an answer to the question ‘Can machines think?’ Instead of defining what he meant by ‘machines’ and ‘think,’ he chose to limit the machines to digital computers and operationalized thinking as the ability to answer questions in a particular context well enough that the interrogator could not reliably discriminate between the answers given by a computer and a human (via teleprinter) after 5 min of questioning. In the original paper consciousness was mentioned only in the context of the objection that thought in the brain is always driven and accompanied by feeling, and so the mere generation of text by a machine in the absence of feeling, however convincing it might otherwise seem, could not be taken as a sufficient indicator of thought. After pointing out that a refusal to accept the computer’s text at face value would constitute a kind of solipsism, a doctrine he thought unacceptable, Turing expressed the opinion that most people initially supporting the argument from consciousness could be persuaded to abandon it rather than embrace solipsism, and so they would probably be willing to accept the validity of the test. In accepting that consciousness cannot strictly be inferred from behavior, Turing does not seem to be claiming that the test can establish that a machine’s thinking is accompanied by feelings, but rather that the difficult issue of determining the involvement of feelings should be set aside in view of the (hypothetical) excellence of the linguistic outputs. More recently a number of people have suggested extensions to the standard test that involve processing audio and visual data or controlling a humanoid body in a human-like way for an extended period of time. In order to pass any of the Turing tests, a machine would almost certainly have to have experience of the world, a capacity for imagination and an emotion system, since no sequence of preprogrammed responses is likely to be convincing over an extended period of time.

Other behavioral measures

The link between certain behaviors and consciousness was discussed earlier, and while this is only vaguely defined at present, it might eventually be possible to use these behaviors to identify consciousness in artificial systems. One potential problem with this approach is that we might not be prepared to attribute phenomenal states to all systems that can carry out the functions associated with consciousness, and so behavioral tests might have to be combined with other measures.

Internal Architectural Tests

Another way of testing for consciousness in artificial systems is to inspect the internal architecture and states of the system and use a theory of consciousness to make predictions about its phenomenal states. This type of work is part of the nascent discipline of synthetic phenomenology, which is attempting to answer the question whether artificial systems are capable of conscious states and trying to describe these states if they occur. This type of work

overlaps with the emerging discipline of neurophenomenology, which makes predictions about human consciousness using objective measurements of a person's brain obtained with electrodes or scanning technologies, such as functional magnetic resonance imaging (fMRI).

At the current stage of research, architecture-based predictions about the phenomenology of a system are far from certain because there is no commonly agreed explanation of consciousness, and the best that can be done is to make predictions according to a particular theory. It is anticipated that work on the neural correlates of consciousness will lead to substantial progress in neurophenomenology and this may feed into the more controversial problem of making predictions about the consciousness of artificial systems.

Ethical and Legal Issues

Many people are concerned that work on machine consciousness will eventually lead to the development of machines that take over the world. Some writers have suggested that this might be a gradual process in which we pass more and more responsibility to machines until we are unable to do without them, in the same way that we are increasingly unable to live without the Internet today. A number of people have responded to this scenario by pointing out that humans cause a great deal of death and destruction every day, and it is possible that conscious machines could run the world better and make humanity happier. Since our current machines fall far short of human intelligence and few people attribute consciousness to them, these science fiction predictions possibly tell us more about our present concerns than about a future that is likely to happen. It is also probable that our attitudes toward ourselves and machines will change substantially over the next century, as they have changed over the last one, and as machines become more human and humans become more like machines, the barriers will increasingly break down between them until the notion of a takeover by machines makes little sense.

A second ethical dimension to work on machine consciousness is the question about how we should treat conscious machines. In order to build systems that are capable of consciousness we may have to carry out experiments that would cause conscious robots a considerable amount of confusion and pain. While we want machines that exhibit the behavior associated with consciousness and want to model human cognitive states and conscious architectures, we may want to prevent our machines from becoming phenomenally conscious if we want to avoid the controversy associated with animal experiments. To regulate this potential suffering, a number of people have suggested that conscious robots should be given rights.

A final aspect of the social and ethical issues surrounding machine consciousness is the legal status of conscious machines. When traditional software fails, responsibility is usually allocated to the people who developed it, but the case is much less clear with autonomous systems that learn from their environment. A conscious machine might malfunction because it has been maltreated, and not because it was badly designed, and so its behavior could be blamed on its carers or owners, rather than on its manufacturers. It might also be appropriate to hold conscious machines responsible for their own actions and punish them appropriately.

Conclusions

AI has made significant progress over the last 50 years and developed from a purely logic-based approach into a range of successful technologies. Many of the problems encountered by AI over the course of its development have improved our understanding of consciousness and AI provides a good way of testing theories of consciousness. A substantial amount of research has been carried out on the construction of machines with the behavior, cognitive characteristics, and architecture associated with consciousness and this work is likely to make steady progress in the future, although this is unlikely to be in the ways predicted by some futurists. A number of people have raised objections to the possibility of machine consciousness, but none of these appear to be conclusive at the present time, and many theories openly embrace the possibility that consciousness could be realized in an artificial system. There has been some work on testing for consciousness in artificial systems and research on machine consciousness has raised a number of ethical and legal issues.

See also: Cognitive Theories of Consciousness; An Integrated Information Theory of Consciousness.

Suggested Readings

Aleksander I (2005) The World in My Mind, My Mind in the World: Key Mechanisms of Consciousness in People, Animals and Machines. Exeter: Imprint Academic.

Chella A and Manzotti R (eds.) (2007) Artificial Consciousness. Exeter: Imprint Academic.

Chrisley RJ, Clowes R, and Torrance S (eds.) (2007) Special Issue: Machine Consciousness: Embodiment and Imagination. Journal of Consciousness Studies 14(7).

Dennett DC (1997) Consciousness in human and robot minds. In: Ito M, Miyashita Y, and Rolls ET (eds.) Cognition, Computation and Consciousness. Oxford: Oxford University Press.

Gamez D (2008) Progress in machine consciousness. Consciousness and Cognition 17(3): 887–910.

Haikonen PO (2003) The Cognitive Approach to Conscious Machines. Exeter: Imprint Academic.

Holland O (ed.) (2003) Machine Consciousness. Exeter: Imprint Academic.

Searle JR (1980) Minds, brains and programs. Behavioral and Brain Sciences 3: 417–457.
Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5(42).
Turing AM (1950) Computing machinery and intelligence. Mind 59: 433–460.

Biographical Sketch

Professor Owen Holland graduated from Nottingham University in 1969 (BSc Hons. in psychology). After teaching experimental methods at Edinburgh University Psychology Department, he moved into commerce, and then into engineering. In 1988, he began his research on behavior-based robotics, for which he received a Department of Trade and Industry SMART award in 1990. In 1993, he moved to the University of the West of England, Bristol, (UWE) to help set up the Intelligent Autonomous Systems Engineering Laboratory. He has held visiting appointments at the University of Bielefeld (1993–94), the California Institute of Technology (1997, 2000–01), and the Ecole Polytechnique Federale de Lausanne (2007), and has also worked in private research laboratories in Cambridge and Brussels. In 2001, he moved to the University of Essex, where he is currently a professor of computer science. His main research interests include biologically inspired robotics, swarm intelligence, swarm robotics, the history of cybernetics, autonomous rotorcraft, and machine consciousness. He has published more than 100 research papers in these areas, and has been particularly closely involved in the development of the field of machine consciousness, having edited the first collection of papers on the topic, and having obtained the first major research grant in the area. With Professor Phil Husbands of the University of Sussex, he is currently working on a book about the Ratio Club, a postwar British cybernetic dining club of which Alan Turing was a member.

Dr. David Gamez completed his BA in natural sciences and philosophy at Trinity College, Cambridge, and went on to study for a PhD in continental philosophy at the University of Essex. This thesis applied methods taken from continental philosophy to contemporary problems in philosophy and science and he later developed and extended this work into a book titled *What We Can Never Know* (London & New York: Continuum, 2007). After leaving Essex he studied for an MSc in IT at Queen Mary, University of London, and then took up a research position on the EU Safeguard project, which developed an agent system that could protect electricity and telecommunications management networks against attacks, failures, and accidents. When the Safeguard project ended, he was offered a PhD position on Owen Holland's CRONOS project to build a conscious robot (see www.cronosproject.net). During this PhD he developed a theoretical framework for machine consciousness and made predictions about the representational and phenomenal states of a spiking neural network. He passed this PhD in July 2008 and is now working as a software engineer for the Trinity Mirror Group.

Attention: Change Blindness and Inattentional Blindness

R A Rensink, University of British Columbia, Vancouver, BC, Canada

Glossary

Change blindness – The failure to visually experience changes that are easily seen once noticed. This failure therefore cannot be due to physical factors such as poor visibility; perceptual factors must be responsible. Focused attention is believed to be necessary to see change, with change blindness resulting if such attention is not allocated to the object at the moment it changes.

Diffuse attention – A type of attention that is spread out over large areas of space. It is believed to be space-based rather than object-based.

Focused attention – A type of attention restricted to small spatial extents. It is believed to act on small areas of space or on relatively small objects.

Implicit perception – Perception that takes place in the absence of conscious awareness of the stimulus. It is generally believed to take place in the absence of any type of attention.

Inattentional blindness – The failure to visually experience the appearance of an object or event that is easily seen once noticed. Attention (likely, diffuse attention) is thought to be necessary for such an experience. Inattentional blindness typically occurs when attention is diverted, such as when the observer engages in an attentionally demanding task elsewhere, and does not expect the appearance of the object or event.

Introduction

As observers, we generally have a strong impression of seeing everything in front of us at any moment. But compelling as it is, this impression is false – there are severe limits to what we can consciously experience in everyday life. Much of the evidence for this claim has come from two phenomena: change blindness (CB) and inattentional blindness (IB).

CB refers to the failure of an observer to visually experience changes that are easily seen once noticed. This can happen even if the changes are large, constantly repeat, and the observer has been informed that they will occur. A related phenomenon is IB – the failure to visually experience an object or event when attention is directed elsewhere. For example, observers may fail to notice an unexpected object that enters their visual field, even if this object is large, appears for several seconds, and has important consequences for the selection of action.

Both phenomena involve a striking failure to report an object or event that is easily seen once noticed. As such, both are highly counterintuitive, not only in the subjective sense that observers have difficulty believing they could fail so badly at seeing but also in the objective sense that these findings challenge many existing ideas about how we see. But as counterintuitive as these phenomena are, progress has been made in understanding them. Indeed, doing so has allowed us to better understand the limitations of human perception in everyday life and to gain new insights into how our visual systems create the picture of the world that we experience each moment our eyes are open.

Change Blindness

Background

The ability to see change is extremely useful in coping with everyday life: we can monitor the movement of nearby automobiles (as drivers or pedestrians), notice sudden changes in the posture or location of people in front of us, and notice that the sky is quickly darkening. Given the importance

of perceiving change and the fact that most humans can survive reasonably well in the world, it follows that our ability to perceive change must be such that few events in the world escape our notice. This agrees nicely with our impression that we perceive at each moment most, if not all, objects and events in front of us.

Failures to see change have long been noticed, but they were usually taken to be temporary aberrations, with nothing useful to say about vision. This attitude began to change in the early twentieth century, when film editors discovered an interesting effect: when the audience moved their eyes across the entire screen (e.g., when the hero exited on the left side and the femme fatale entered on the right) almost any change made during this time (e.g., a blatant change of costume) would often go unnoticed. A similar blindness to change could be induced by making it during a loud, sudden noise (e.g., a gunshot), during which the audience would momentarily close their eyes.

The scientific study of this effect began in the mid-1950s, with work on position changes in dot arrays and other simple stimuli. Here, a change in one of the items was typically made contingent on a temporal gap that lasted several seconds. A separate line of studies was also begun that investigated the perception of displacements made contingent on an eye movement (or saccade). In all cases, observers were found to be surprisingly poor at detecting changes made under such conditions.

The next wave of studies, begun in the 1970s, was based on a more systematic examination of these effects. This work uncovered a general limit to the ability to detect gap-contingent changes under a wide variety of conditions; this eventually formed the basis for the proposal of a limited-capacity visual short-term memory (vSTM). Likewise, a general lack of ability was found for detection of saccade-contingent changes, which was traced to a limited transsaccadic memory. Both lines of research were eventually linked by the proposal that transsaccadic memory and vSTM were in fact the same system.

A third wave began in the mid-1990s, extending the methodology and results of earlier work in several ways. To begin with, stimuli were often more complex and realistic: images of real-world scenes or dynamic events were used in place of simple figures. Next, changes were often repeated rather than occurring just once, allowing the use of time as well as accuracy to measure performance (Figure 1). Third, blindness was induced via several new kinds of manipulation, such as making changes during a film cut or during the appearance of sudden splats elsewhere in the image. Finally, all these effects – as well as the earlier ones – were accounted for by the proposal that attention is necessary to see change. This work therefore showed that this CB is a general and robust phenomenon that can last for several seconds, and can be connected to known perceptual mechanisms. As such, it supported the idea that change perception is an important part of perception, and that studying its failure can cast light on mechanisms central to our experience of the world.

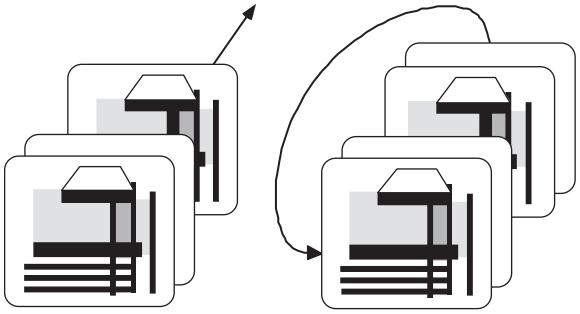


Figure 1 Example of a technique to induce CB. Here the change is made during a brief gap between the original and the modified stimulus. (a) One-shot paradigm. A single alternation of the stimuli is used, and the observer must respond to it. Performance is measured by accuracy of detection (or identification or localization). (b) Flicker paradigm. Stimuli are continually cycled until the observer responds. Performance is measured by the time taken until the change is detected (or identified or localized). Both types of measurement paradigm can also be applied to other techniques, such as changes made during saccades or splats.

Conceptual Distinctions

A clear understanding of CB – and its inverse, change perception – has been slow to emerge. This is partly due to the nature of change itself. Although this concept appears simple, attempts to formalize it over the years – from the earliest Greek thinkers to modern philosophers – have

generally encountered difficulties. However, a number of important distinctions have become clear, which have helped our understanding of CB.

Change versus motion

The human perception of temporal variation is handled by two separate systems: a motion-detection system for variation in regards to a particular location and a change-detection system for variation in regards to structure. Temporal variations in the world activate both systems to some extent. For example, a moving automobile will activate motion detectors over the relevant part of the visual field, although these will not be able to determine whether there is an enduring spatiotemporal structure that remains constant. Conversely, a whirling dust devil will be seen as a coherent moving structure, even though there is nothing beyond the flow of dusty air in space. The key challenge in perceiving motion and change is to separate out the contributions of these two systems, so that temporal variation is assigned to the correct substrate.

Change versus difference

Another important distinction is that between change and difference. In both cases, reference is to a particular structure – an object or event of some kind. However, while change refers to transformation of a single structure, difference refers to the comparison of two or more separate structures.

More precisely, change is based on the properties of the same structure at separate points in time, and can be perceived as a single dynamic visual event. For example, when a person is seen walking, their legs are seen at various positions over time. This kind of perception suggests – although does not prove – that the underlying representation has a spatiotemporal continuity of some kind. In contrast, difference is based on an atemporal comparison of structures that may or may not exist simultaneously, for example, comparing the height of two people. This kind of perception allows for the possible involvement of long-term memory elements that are only intermittently engaged.

This distinction is important. Change detection and difference detection can be distinguished, at least conceptually; it is important to consider the type of experiment used when discussing either one. Spotting the difference between two side-by-side stimuli may not engage the same perceptual mechanisms as detecting a change in successively presented stimuli. Failure to make this distinction can lead to erroneous conclusions being drawn about the mechanisms involved.

Experimental Approaches and Results

All studies to date have been based on much the same design: an initial stimulus is first presented (e.g., a picture), an altered version is then shown (e.g., some object in it is removed), and the ability of the observer to perceive the change is then measured (see Figure 1). Various aspects of perception have been explored in this manner based on particular selections among a relatively small set of design parameters.

Type of task

CB is the failure to perceive a change. Since there are several different aspects to perception, there can be – at least in principle – several different kinds (or levels) of blindness. These can be investigated by giving the observer the appropriate kind of task:

- Detection. This is the most basic and the most widely studied aspect, concerned with the simple noticing of a change. No properties of the change itself are necessarily perceived: the observer is simply asked to report whether it occurs.
- Identification. This concerns reporting the properties of the change, that is, seeing what type it is (e.g., a change in color, a change in orientation). This can in principle be separated from detection: the observer could be asked to guess the type of change even if the change itself was not detected. Identification of change appears to be more difficult than detection, indicating that somewhat different mechanisms are involved.
- Localization. This is concerned with reporting the location of the change. This can be decoupled from detection and from identification, at least in principle. Relatively little work to date has been done on this aspect of perception. Some results suggest that a separate memory for location may exist, although this has not yet been fully established.

Type of response

Another way to engage different mechanisms is to use different aspects of the response of an observer to a change in the external world. These effectively test different perceptual subsystems:

- Explicit percept. This is the approach used in most perceptual experiments, involving the conscious visual experience of the observer. A high degree of blindness can usually be induced. The proposal that attention is needed to see change is concerned with this type of experience.
- Semiexplicit percept. Some observers can have a 'gut feeling' for several seconds that change is occurring, even though they do not yet see it (i.e., do not have a picture of it). The basis of this is controversial, but it may involve nonattentional or subattentional systems.
- Implicit percept. Change that is not experienced consciously may still be perceived implicitly. If so, this must be measured by its effect on other processes. For example, some studies indicate that an unseen change can influence forced-choice guessing about its possible location. The existence of this form of perception is controversial, although evidence for it appears to be increasing.
- Visuomotor response. This involves the response of a visually guided motor system to a change that is not consciously experienced. Systems used are almost always manual pointing or eye fixation. Both kinds of visuomotor response are faster than consciously mediated ones. They also appear to be more accurate, suggesting the existence of representations with higher-capacity memory.

Attentional manipulation

A central tenet in change perception is that attention is needed to consciously experience change. In normal viewing, the local motion signals accompanying a change attract attention to its location, allowing the change to be seen immediately; this is why it helps to wave at a friend from across the room. If these local signals can be neutralized so that the automatic drawing of attention cannot help, a time-consuming attentional scan of the display will be needed. The observer will consequently be blind to the change until attention is directed to the appropriate item. A variety of techniques have explored this:

- Gap-contingent techniques. The change is made during a blank field or mask briefly displayed between the original and altered stimulus, which swamps the local motion signal (see Figure 1). Observers are very poor at detecting change if more than a few items are present; results suggest only 3–4 items can be seen to change at a time.
- Saccade-contingent techniques. The change is made during a saccade of the eyes. Observers are generally poor at detecting change if more than a few items are present; again, a limit of 3–4 items is found. CB can also be induced for position change with even one item present, provided no global frame of reference exists.
- Blink-contingent technique. The change is made during an eyeblink. Again, observers are generally poor at detecting such changes. Interestingly, blindness can be induced even if the observer is fixating the changing item.
- Splat-contingent techniques. These make the change at the same moment as the appearance of brief distractors (or splats) elsewhere in the image. The blindness induced in this manner is relatively weak, but still exists, showing that it can be induced even when the change is completely visible.
- Gradual-change techniques. Here, the transition between original and altered display is made slowly (e.g., over the course of several seconds). Observers generally have difficulty detecting such changes, even though no disruptions are used.

The results from all these approaches are consistent with the proposal that attention is needed to see change. The finding that at most 3–4 items can be seen to change at a time is consistent with the capacity of attention obtained via other techniques such as attentional tracking.

Perceptual set

An important part of perception is the perceptual set of the observer, which strongly affects the mechanisms engaged for a given task. The issue of set is important for the question of which

mechanisms are involved in everyday vision, and how these are related to performance as measured in the laboratory:

- Intentional set. Observers are instructed to expect a change of some kind. They are assumed to devote all their resources to detecting the change, which provides a way to determine perceptual capacities. A further refinement is controlling expectation for particular types of change. Changes for presence and location are not affected by expectation of type, whereas changes for color are.
- Incidental set. Observers are given some other task as their primary responsibility (e.g., count the number of sheep in an image); there is no mention of a possible change until after the task is over. The engagement of perceptual mechanisms is believed to be more representative of their use in everyday life. The degree of blindness found under these conditions is higher than under intentional conditions, indicating that relatively little is attended – or at least remembered – in many real-life tasks.

Implications for Perceptual Mechanisms

At one level, CB is an important phenomenon in its own right: among other things, it illustrates the extent to which we can potentially miss important changes in everyday life. Indeed, most people are unable to believe the extent to which they are unable to see change – in essence, they suffer from 'CB blindness.'

However, CB itself can also be harnessed as a powerful tool to investigate the mechanisms by which we see. The exact conclusions obtained from such studies are still the subject of debate, but their general outlines are becoming clear.

Visual attention and short-term memory

All experiments on CB are consistent with the proposal that attention is needed to see change. Experiments on carefully controlled stimuli suggest that 3–4 items can be seen to change at a time, consistent with the limit on attention obtained via other techniques, such as attentional tracking.

However, while a limit of some sort is involved, the nature of this limit is still somewhat unclear. Contrary to subjective impression, change perception is not an elementary process. Instead, it involves – at least in principle – a sequence of several steps: enter the information into a memory store, consolidate it into a form usable by subsequent processes, hold onto this for at least a few hundred milliseconds, compare it to the current stimulus at the appropriate location, clear the memory store, and then shift to the next item. A limit on any of these steps would limit the entire process, making it difficult to determine the relevant step in a given situation.

Most proposals for mechanisms have been couched in terms of either visual attention or vSTM. Both are similar in the results they cite and the mechanisms they propose. Part of this is caused by the extensive overlap that appears to exist between the mechanisms associated with attention and vSTM. Indeed, the difference may be largely one of terminology, caused by the vagueness in the definition of attention. As used in CB studies, attention is defined by the formation of representations coherent over space and time; such representations are not that different from those posited as the basis of vSTM.

In any event, interesting new issues are emerging. One is whether the limiting factor applies to the construction of the coherent representation, or to its maintenance once it is formed, or perhaps both. Another issue is the nature of the elements that are attended and held in vSTM. Much evidence suggests that these are proto-objects that already have a considerable local binding of features; if so, the function of attention and vSTM would be to create a representation with extended spatial and temporal coherence. Results also suggest that the 3–4 items are not independent, but, rather, may have a higher level of interaction such that their contents are pooled into a single collection point (or nexus) that supports the perception of an individual object.

Scene perception

Given that only a few items may have a coherent representation at any time, our subjective impression of seeing everything that happens in front of us cannot be correct. But care must be taken in the particular inferences drawn. Since it deals only with dynamic quantities, CB cannot say anything

about the static information that may or may not be accumulated. On the other hand, it does show the existence of severe limits on the extent to which changes are represented in those subsystems accessible to conscious perception.

To account for the impression that we see all the changes that occur in front of us – not to mention that we can actually react to many of these – it has been proposed that scene perception is handled by a virtual representation. Here, coherent representations of objects – needed for change perception – are created on a 'just in time' basis, that is, formed whenever they are needed for a task and then dissolved afterward. Coordination of this process can be achieved via a sparse schematic representation of the scene – perhaps a dozen or so items, each with some properties – formed independently of the coherent representations. Guidance could be done on the basis of both high-level factors (e.g., schemas that enable testing of expected objects) and low-level factors (e.g., motion signals that draw attention to unexpected events). All results on CB are consistent with this proposal, and results from other areas (e.g., research on eye movements) appear consistent with this as well.

The status of static scene information is still unclear. In principle, it could be accumulated to create a dense description that would match our subjective impression. However, such accumulation is unnecessary: a virtual representation could handle most if not all aspects of scene perception. Furthermore, no results to date have clearly shown storage of information beyond the relatively sparse information used for guidance and the contents of attention and vSTM. Some information about the prior state of a changed item appears to be stored in a longer-term memory not used for the perception of change; the amount of this is information not known. More generally, many scenes might be stored in long-term memories of various kinds, but the information density of each could still be quite low – perhaps a relatively limited amount of information from each of a dozen or so locations.

Whether the dynamic element is total or partial, it nevertheless plays an important role in each individual's perception of a given scene. CB has therefore been useful as a tool to investigate individual differences in perception: the faster a change is detected, the more important it is deemed to be. Studies have shown an effect of training – including culture – on the encoding of objects and the importance attached to them, along with an influence of the particular task undertaken. There is also an emerging connection here with the design of interactive visual interfaces: these owe much of their success to the engagement of these dynamic mechanisms, and are therefore subject to many of the same considerations regarding operator and task.

Inattentional Blindness

Background

It has been known for thousands of years that people engaged in deep thought can fail to see something directly in front of their eyes. Such blindness can be easily induced when an observer intensely attends to some event, for example, waiting to see if an oncoming automobile will stop in time. Under such conditions, much of the visual field can effectively disappear from consciousness, even if it contains objects that are highly visible. It was long believed that such blindness might be due to the image of an unseen object falling onto the periphery of the eye, which is relatively poor at perceiving form. But decades of research have shown that location is unimportant – the key factor is attention.

One of the earliest lines of research that encountered this IB was the study of head-up displays for aircraft pilots, which superimpose instructions and information over a view of the external world. Simulator studies in the early 1960s discovered that when pilots carried out difficult maneuvers requiring attention to outside cues, they did not see the unexpected appearance of instructions to alter course. Later studies showed the converse effect: if pilots focused on the displayed instructions, they often failed to see highly visible aircraft unexpectedly rolled out onto the runway in front of them.

The first perceptual studies to investigate this effect took place in the 1970s. These used selective looking tasks, where observers were presented with two superimposed visual events (e.g., a sports game and the face of someone talking) and asked to attend to just one of these. Results showed that observers could easily report what happened in the attended event, but did not do well in the other:

they often missed large unexpected events, such as the appearance of a woman carrying an umbrella. This happened even when the unseen stimulus was in the center of the visual field, showing that the cause of this blindness effect was not optical, but attentional.

Although this effect was surprising, the use of superimposed stimuli caused it to be considered somewhat artificial, and interest in it never really developed. But in the 1990s a new wave of studies did manage to kindle interest. These differed from the earlier studies in several ways. First, they used opaque rather than superimposed (or transparent) stimuli, greatly reducing concerns about artificiality. Indeed, it was found that even large unexpected events – such as the appearance of a human in a gorilla suit – were still not noticed by most observers under these conditions. Second, in addition to videos of events, techniques were also developed based on brief-presented static images (Figure 2), which showed that IB could apply to both static and dynamic stimuli. Third, simple stimuli were often used in both the static and dynamic case, allowing more experimental control. And finally, there was greater examination of the effects caused by the stimuli that were not seen consciously. Together, these developments showed that IB could indeed occur in everyday life, and that the mechanisms involved play a major role in everyday perception.

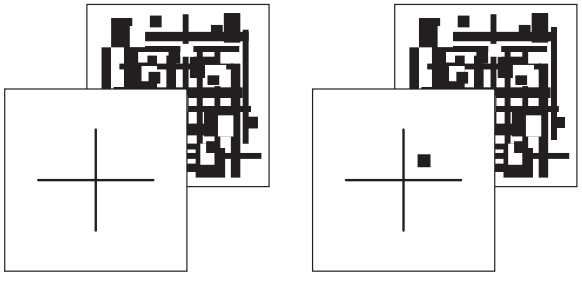


Figure 2 Example of a technique to induce IB. A static test stimulus is presented while visual attention is focused on a primary task; here, the task is to report which of the two lines is longer. (a) Noncritical trial. The primary lines alone are presented for 200 ms, followed by a mask to prevent further processing. The observer must report which is the longer line. Two or three such trials are usually given to allow the observer to get into the appropriate mental state. (b) Critical trial. Along with the primary lines, an unexpected test stimulus is presented nearby. Blindness is measured by detection (or identification or localization) of the test stimulus. An additional round of trials is often presented, in which the test stimulus is now expected; this is believed to correspond to a condition of divided attention.

Conceptual Distinctions

Work on the theoretical and experimental aspects of IB have emphasized several conceptual distinctions. Some apply not only to IB but are also relevant to more general issues of awareness.

Expression versus suppression

Given that attention is needed to see something (i.e., consciously experience it), two general possibilities exist as to how it acts. First, attention might cause selected stimuli to be expressed consciously, for example, they enter awareness by being activated so as to exceed some threshold. In this view, attention does not just facilitate the conscious perception of a stimulus – it also enables it. On the other hand, attention might act to selectively suppress some of the stimuli, so that they vanish from the conscious mind. (In both cases, the status of unattended items is open: they might be weakly expressed in some way, or available implicitly.) It is worth noting that the two possibilities are not exclusive: it could be that both are used, with attention acting to express – or at least emphasize – some items while suppressing others.

Restricted versus unrestricted effects

A related issue is the extent of any expression or suppression. It was once believed that everything registered in the visual system was experienced consciously. However, given that implicit perception of various kinds has been found, an important issue is whether the selective effects underlying IB are restricted to conscious experience, or whether they spill over to other aspects of perception.

For example, results show that attentional inhibition of a given location will increase the amount of blindness there. But does such inhibition attenuate the input strongly enough to also affect implicit processes at that location? Virtually all studies to date have assumed that selective effects are restricted to conscious experience, but this assumption has not been tested.

Experimental Approaches and Results

At the most general level, all studies of IB use the same basic approach. An observer is given a primary task that engages their visual attention, an unexpected test object (or event) then appears in the display, and the response of the observer to that appearance is then tested. Because expectation is an important factor, several trials without a test stimulus are generally presented first. Results are taken from the trial in which the test stimulus first appears (the critical trial). After this, the object is no longer entirely unexpected, and the blindness levels in subsequent trials are usually much lower, consistent with the idea that attention is now divided among the expected stimuli.

Type of task

As in the case of CB, several different aspects of perception can be distinguished, each of which can be tested by appropriate selection of task:

- Detection. This is the most basic aspect of perception, concerned with the simple noticing of something present; no properties of the stimulus itself need be involved. For isolated static items that are small, that is, have an extent of less than about 1 of visual angle, observers often fail to see anything at all in conditions that induce IB. For larger items, however, detection remains generally good. For dynamic events, informal observations suggest that the unattended event does not necessarily disappear entirely – observers often notice 'something else' going on, but cannot say what it is.
- Identification. This is concerned with determining various properties of the stimulus. For example, observers can be asked to report the color or location of an item; in principle – although not often in practice – this can be done whether or not it was detected. Observers can usually report the color of a small item that is detected; identification of shape appears to be more difficult. Detection without identification has been reported, but it has been suggested that these are simply false positives.
- Localization. Here, the task is to locate the test stimulus that appears in the critical trial. In principle, this can be tested independently of detection and identification, although this is rarely done. Observers are able to report the location of a small item that is detected, showing that there may be some interaction between these two aspects of perception.

Type of response

Performance can also be studied in terms of the kind of percept experienced. Operationally, this can be done by testing different aspects of the observer's response to the appearance of a stimulus. These different aspects effectively involve different perceptual subsystems:

- Explicit percept. This is the kind of percept used in most IB experiments. The observer is asked to respond to their conscious experience, that is, the picture they have of something. The usual definition of IB is in terms of the failure to have such an experience; the proposal that attention is needed to see is likewise concerned with this aspect.
- Implicit percept. Items that are not experienced consciously may still be perceived implicitly. Given that the observer is blind to the stimulus, responses cannot be measured via direct subjective experience; they must instead be based on indirect effects. For example, relatively little blindness is found for emotionally laden words or pictures. This suggests that such words have been perceived implicitly, with attention drawn to them on the basis of their meaning. Similarly, a length illusion can occur in a set of lines perceived consciously, even if the background lines that induce the illusion are themselves not reported.
- Motor response to primary task. The response to the appearance of a test stimulus can be measured by its effect on the speed or accuracy of the primary task. This is essentially the approach developed to study attention capture, which measures the effect of a new item on motor response times for the main task, regardless of whether the new item is seen. (An interesting variation would be to determine if such effects are conditional on the experiential state of the test stimulus.)
- Visuomotor response. Here, the appearance of an item causes the eye to move toward it. This is another approach developed in the area of attention capture; again, this response is measured

regardless of whether the item is seen. This might be useful for studying IB if measurement were made conditional on whether the test stimulus is consciously experienced.

Attentional manipulation

An important factor in inducing IB is to present the test stimulus while the observer has their visual attention engaged on an unrelated task. This has been done in several ways:

- Superimposed stimuli. Two independent dynamic events (e.g., a basketball game and two pairs of hands playing a game) are presented simultaneously. In earlier studies, this was often done via half-silvered mirrors; more recent studies do this electronically. The result is perceived as a set of transparent or ghostly images. Observers can easily attend to one of these, but can often miss events in the other.
- Interspersed stimuli. Here, two different sets of stimuli are presented; all are opaque and appear on the same display (see Figure 2). Tests on static stimuli generally use this method. Observers are asked to carry out a primary task on one of the sets (e.g., make a length judgment on a pair of lines); a test stimulus is presented near these lines a few trials later. Dynamic events have also been tested this way. A high degree of blindness can be found, even when the items in the two sets of stimuli are intermingled.
- Dichoptic presentation. Here, two independent events are presented, each to a different eye; the observer is asked to pay attention to one of them. Events in the unattended eye often fail to be reported. In contrast with superimposed stimuli, where observers often have an impression of 'something else' going on, observers here fail to experience anything of the unattended set – it is simply not there. This may be related to the blindness experienced in binocular rivalry suppression.

Perceptual set

An important aspect of IB is the perceptual set of the observer. There are several ways this could influence performance:

- Control of selectivity. This is the extent to which selection – either expression or suppression – is invoked in the primary task. For example, observers can be required to attend only to the white-shirted players in a game, while ignoring the black-shirted players. In such tasks, blindness appears to be due at least in part to observers suppressing the features of the ignored stimuli: the greater the similarity of the test item to the ignored items, the greater the blindness. In non-selective tasks, there is no need to screen out any stimuli, at least up to the critical trial. Blindness is still induced here, but it remains unknown whether this is due to a failure of expression or an invocation of suppression.
- Control of capture. It has been suggested that high-level control may be exerted over the kind of stimuli that can capture attention, and thus be seen. Attention is usually drawn by the appearance of a new item or by the presence of a unique property. However, work on attention capture has shown that this can be overridden by a high-level attentional set. Results on IB are consistent with the proposal that an attentional set governs what is consciously seen, with distinctive stimuli experienced consciously only if they fit into the observer's expectations.

Implications for Perceptual Mechanisms

The finding that observers can often fail to see highly noticeable objects and events touches on several issues concerning the way we cope with our world. For example, it suggests that we might not be aware of the extent to which we fail to see various aspects of our immediate environment, even if these are important and highly visible. (This might be termed IB blindness.) This has obvious implications for tasks such as driving, where the ability to accurately perceive objects and events is literally a matter of life and death, and where knowing about our limitations could well affect how careful we are.

Meanwhile, work on IB can also tell us about the mechanisms involved in visual perception. In particular, it can provide a unique perspective on the mechanisms that underlie our conscious experience of the world.

Visual attention

All results to date are consistent with the proposal that attention is needed to see an object or event. For example, the degree of blindness has been found to increase with distance from the center of the attended location, in accord with space-based models of attention. In addition, the greater the attentional load of the primary task, the greater the blindness to the test stimulus, which is consistent with the proposal of a limited attentional capacity. Indeed, the inability of an observer to follow more than one coherent event supports the proposal that only one complex object or event can be attended at a time.

Results also provide tentative support for the proposal that the high-level control of attention is achieved via an attentional set which determines the kinds of information that can capture attention, and perhaps also the kinds of information that can enter conscious awareness. This development potentially connects work on IB to work on attention capture.

Another possible connection is with preattentive vision, usually studied by the rapid detection (or pop-out) of unique items in a display. Work here has pointed toward a considerable amount of processing achieved in the absence of attention. But this assumes that little or no attention is given to most items in a display at any moment. Results on IB suggest that this supposedly inattentional condition may be better viewed as a case of diffuse attention, with the inattentional condition characterized as one where no conscious perception exists.

However, there are several findings that make these connections less than certain. For example, an observer in an IB experiment can usually detect the individual items in a group, although the grouped pattern itself cannot be identified. In addition, the pop-out of a unique item in a group occurs only if the group appears in the critical trial; if this group is shown earlier in noncritical trials (with the unique item the same as the others), pop-out no longer occurs. So what is the effect of the unique item here? If it is to draw attention to the group, why are the individual elements already seen in the earlier, noncritical trials? And why should pop-out occur in one condition, but not the other?

Scene perception

Relatively little work has studied the aspects of natural scenes perceived under conditions of inattention. Studies using briefly presented scenes as test stimuli found that observers could usually report the scene gist (i.e., its overall meaning, such as being an office or a forest), along with several objects of indeterminate description. Some confabulation is also found. This is broadly similar to results on the ability of observers to rapidly perceive the gist of scenes from brief exposures of 100 ms or less, which is also likely done with little or no focused attention.

Individual differences in the coding of scenes and events can be measured in terms of how blindness varies with the primary task. For example, experts in basketball could better detect the appearance of an unexpected object while they were attending to a basketball game. This suggests that they had encoded the scenes and events in a way that allowed them to divert some of their attention to occasionally monitor other items, without seriously affecting performance on the primary task.

The model of scene perception often used to account for results on IB is the perceptual cycle. Here, sustained attention is believed to activate the conscious percept of a stimulus; once this has been done, the stimulus has entered the cycle, and helps select the appropriate schema to determine what information to admit next. In contrast, stimuli that do not become part of the predictive cycle may never be seen at all. Such stimuli – especially low-level signals – can guide the process, but do not enter awareness on their own. Similarly, information assembled at a preattentive level can also guide this process, although it too does not enter awareness automatically. This model has some similarities to the dynamic scene representation often used to account for CB, as well as the reentrant models used to account for conscious experience.

General Issues

Change Blindness versus Inattentional Blindness

CB and IB both involve a failure to perceive things that are easily seen once noticed, and both are

believed to be due to a lack of attention. It is therefore likely they are related. But exactly how?

First of all, the difference between CB and IB does not depend on the kind of input. Both can be found using dynamic images, and both can be found using static images. And the particular contents of the input do not matter for either. Instead, the critical difference between the two is the status of the information under consideration. IB is entirely concerned with first-order information – the simple presence of quantities. In contrast, CB involves second-order information – the transitions between these quantities. These can be separated: an alternating sequence of two images, say, could be experienced as a 50% presence of each image over time (same first-order distributions), but with different amounts of change (different second-order distributions) if the alternation rates are not the same. And just as CB can say little about first-order (static) information, IB can say little about second-order (changing) information. The two therefore refer to largely complementary aspects of the visual world.

This distinction has consequences for the perceptual mechanisms involved. For example, the kind of attention required for each aspect of perception may be different – or at least, have different effects. The kind of attention involved in IB is space based: the degree of blindness increases with increasing distance from the center of attention. It is also easily diverted – hence the common use of a test stimulus that is completely unexpected. In contrast, the kind of attention needed to perceive change is object-based and is much less easily diverted (or at least slower), since telling the observer that a change will occur – and even giving them practice at perceiving it – does not affect performance to any great extent. Loosely speaking, IB might be identified with the absence of diffuse attention, and CB with the absence of focused attention. But a final determination of this must await a better understanding of attention itself.

Blindness versus Amnesia

An important issue in regards to the status of these induced failures is whether they are failures of perception or of memory. It might be, for instance, that an observer did experience something under conditions that induced CB or IB, but then forgot it before they could make their report. If so, these effects would not be forms of blindness, but forms of amnesia.

In the case of CB, the resolution of this issue is reasonably straightforward. Perception of change is often measured by asking the observer to respond to the change as soon as they see it; all that is needed to trigger this is a minimal conscious experience. When observers are asked to respond to a single change, only a few hundred milliseconds exist between its presentation and the initiation of the report (e.g., the pressing of a button). Thus, if the experience of change is forgotten in the absence of attention, it would be exceedingly brief and incapable of causing a response to be initiated. To all intents and purposes it would be as if it never existed, at least at the conscious level. (Note that if taken seriously, the possibly of such a fleeting perception would not be restricted to CB – it would apply to any failure of perception.)

The situation for IB is more complex. In one sense, this issue was resolved by the finding that unattended – and thus unseen – items are indeed perceived, in that they can indirectly affect aspects of conscious experience. Entry into conscious experience itself, however, is not addressed by this. The observer is usually asked about a possible item only after the primary task has been completed, which allows several seconds to possibly forget it. And, the observer cannot be asked to prepare to respond to the test stimulus, since this sets up an expectation of the item, which severely diminishes the attentional diversion, and thus the degree of blindness.

This issue has been grappled with in several ways. One is to present a highly surprising or meaningful item, and hope that the observer will spontaneously report it. However, this has not generally been successful: even if a human walks around in a gorilla suit or if an airplane is wheeled out onto the runway where the pilot is about to land, most observers still do not respond. It has been proposed that the observers do see the visual elements, but do not assign meaning to them. In essence, this phenomenon becomes one of inattentional agnosia. (This may explain the reports in some selective looking experiments, where observers see 'something else' going on, but do not know what it is.)

Another perspective derives from studies of the neural systems involved. Patients with lesions to particular parts of the cortex can suffer conditions such as neglect and extinction, in which attention cannot be easily allocated to objects. Such patients, however, do not appear to experience forgetting – they simply do not report perceiving such stimuli, even when asked with the object in full view. In addition, functional imaging of the brains of normal observers shows that words are not consciously identified in the absence of attention, even if the observer is looking directly at them. These results make it highly likely that the failure is one of perception, and not memory.

Visual Attention versus Visual Experience

CB and IB can be regarded as two forms of the perceptual failure created by the diversion of attentional resources. They can be distinguished at the functional level by the type of information involved (second- or first-order information, respectively). They also appear to be distinguished by the type of attention involved (focused or diffuse) and the kinds of operations (e.g., comparison) associated with these. This division may correspond to the two modes sometimes proposed for conscious visual experience: an object mode associated with focused attention and a background mode operating as default. Beyond this, however, only partial and tentative conclusions can be drawn regarding the issue of how visual attention relates to conscious visual experience.

In the case where attention of any kind is absent, there does not appear to be any conscious experience of stimuli (second-order quantities for focused attention; first-order quantities for diffuse). However, results still point to a considerable amount of processing being carried out. For example, work on IB indicates that unattended – and therefore unseen – items can influence the perception of attended items. Similarly, some models of CB posit low-level representations with a degree of detail and feature binding (proto-objects) that are formed in the absence of this kind of attention.

It is worth pointing out that observers in IB experiments often report that they can detect something about the nonselected stimuli, even though they cannot always identify it. Importantly, this kind of experience is found only in those experiments involving superimposed or interspersed stimuli; for dichoptically presented stimuli, there is a complete absence of perception of the nonselected event. This suggests that in the superimposed and interspersed conditions diffuse attention is given to nonselected events, with the main event given focused attention. If so, this would suggest that identification and localization may require more focused attention (or related resource), and that both diffuse and focused attention may be allocated simultaneously to different stimuli.

This proposal would be consistent with work on CB. Focused attention is needed only for the perception of complex quantities such as change; background items not given focused attention might still be seen, but only in regards to detection and perhaps a limited form of identification based on relatively fragmented pieces of static items.

Conclusions

Work on CB suggests that focused attention is needed for the conscious experience of change: without it, observers will be blind to even large changes, at least at the conscious level. Some ability to implicitly perceive change may exist; if so, this does not appear to require focused attention.

Similarly, work on IB suggests that diffuse attention is needed to detect an unexpected object or event, that is, to see it as 'something.' Some results indicate that further (attentional) processing may be needed to more completely identify or locate it. There also appears to be some ability to pick up and process information about static items in the absence of any form of attention, even though such items are not experienced consciously.

Beyond this, our current understanding is poor. More work is needed to expand our empirical knowledge of the basic phenomena. More work is also needed on the basic conceptual issues involved, in particular, on our understanding of terms such as attention and awareness. However, much exciting progress is being made on these fronts. And some of the most powerful sources of these new developments are the phenomena of CB and IB.

See also: Attention: Selective Attention and Consciousness; Neglect and Balint's Syndrome; Perception: Unconscious Influences on Perceptual Interpretation.

Suggested Readings

Brockmole JR and Henderson JM (2005) Object appearance, disappearance, and attention prioritization in real-world scenes. Psychonomic Bulletin & Review 12: 1061–1067.

Irwin DE (1991) Information integration across saccadic eye movements. Cognitive Psychology 23: 420–456.

Levin DT (2002) Change blindness blindness as visual metacognition. Journal of Consciousness Studies 9: 111–130.

Luck SJ and Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390: 279–280.

Mack A and Rock I (1998) Inattentional Blindness. Cambridge, MA: MIT Press.

Moore CM and Egeth H (1997) Perception without attention: Evidence of grouping under conditions of inattention. Journal of Experimental Psychology: Human Perception and Performance 23: 339–352.

Most SB, Scholl BJ, Clifford E, and Simons DJ (2005) What you see is what you set: Sustained inattentional blindness and the capture of awareness. Psychological Review 112: 217–242.

Neisser U and Becklen R (1975) Selective looking: Attending to visually significant events. Cognitive Psychology 7: 480–494.

Rensink RA (2002) Change detection. Annual Review of Psychology 53: 245–277.

Rensink RA, O'Regan JK, and Clark JJ (1997) To see or not to see: The need for attention to perceive changes in scenes. Psychological Science 8: 368–373.

Simons DJ (ed.) (2000) Change Blindness and Visual Memory. Hove, East Sussex, UK: Psychology Press.

Simons DJ and Chabris CF (1999) Gorillas in our midst: Sustained inattentional blindness for dynamic events. Perception 28: 1059–1074.

Simons DJ and Rensink RA (2005) Change blindness: Past, present, and future. Trends in Cognitive Sciences 9: 16–20.

Wilken P (ed.) (1999) Mack & Rock: Inattentional blindness. Psyche 5–7. http://psyche.cs.monash.edu.au/symposia/mack_rock/.

Wolfe JM (1999) Inattentional amnesia. In: Coltheart V (ed.) Fleeting Memories, pp. 71–94. Cambridge, MA: MIT Press.

Biographical Sketch

Ronald A Rensink is an associate professor in the Departments of Computer Science and Psychology at the University of British Columbia (UBC). His interests include human vision (particularly visual attention), computer vision, and human–computer interaction. He has presented work at major conferences and in major journals on visual attention, scene perception, computer graphics, and consciousness. He received his PhD in computer science (in computer vision) from UBC in 1992, and was then a postdoctoral fellow for 2 years in the psychology department at Harvard University. This was followed by a position as a research scientist at Cambridge Basic Research, a laboratory sponsored by Nissan Motor Co., Ltd. He returned to UBC in 2000, and is currently a part of the UBC Cognitive Systems Program, an interdisciplinary program that combines computer science, linguistics, philosophy, and psychology.

Attention: Selective Attention and Consciousness

P D L Howe, K K Evans, R Pedersini, T S Horowitz, J M Wolfe and M A Cohen, Brigham and Women's Hospital, Harvard Medical School, Cambridge, MA, USA

Glossary

Balint's syndrome – A neurological disorder caused by bilateral damage to the parietal cortex that results in the inability to perceive more than one object at a time.
Binding problem – The difficulty of mentally conjoining features that belong to the same object.
Blindsight – A neurological disorder, typically caused by damage to the primary visual cortex, that results in the patient being unaware of any stimuli located in a particular part of the visual field while still being able to detect them.
Feature – An attribute of an object (e.g., its color).
Illusory conjunction – An illusory combination of features from different objects.
Receptive field – The region of the retinal image to which a particular neuron responds.
Pop-out – In visual search, the situation where the speed or accuracy of target detection is independent of the number of distractors.

Introduction

We are aware of the world around us, but not in a uniform fashion. We selectively attend to some stimuli and are consequently less aware of others. You are probably sitting, at present. If you selectively attend to the pressure of your posterior on the seat, you will become more aware of that sensation than you were a moment before. This article deals with the relationship between conscious awareness and selective attention. There are three possibilities. It could be that selective attention has no effect on our conscious awareness. As the opening example makes clear, this is not a promising hypothesis and we will ignore it. At the alternative extreme, it has been proposed that conscious awareness is fully determined by selective attention that we are conscious only of the current contents of attention. A more moderate position is that attention modulates awareness, but we have some awareness of unattended stimuli. This middle position reflects our view.

At the outset, justifying this position is made difficult by the many uses of terms like consciousness and attention in common speech and technical writing. In this entry, we will restrict ourselves to the conscious awareness of visual stimuli, though the same questions arise in the other senses, between sensory domains and perhaps even when monitoring one s own thoughts.

We will often use the term object another term with a problematic definition. For example, consider an image of a face. One could consider the entire face to constitute a single object. Alternatively, one could consider the eyes, the nose, the mouth, and so on to be objects. Indeed, each of these objects could in turn be decomposed as you attend to, say, a pupil or a nostril. As there is no general agreement on what constitutes an object, we avoid the issue. Instead, we ask for the reader s indulgence and use the term as a layman would imprecisely.

Colloquial speech tends to incorrectly treat awareness (like attention) as a single entity. In fact, we can profitably distinguish between the type of awareness that accompanies attention and the type of awareness that seems to occur in the absence of attention. This is an old idea. In 1780, Etienne Bonnot de Condillac asked his readers to imagine arriving at a chateau late at night. The next morning, you wake in a completely darkened room. Then the curtains are thrown open for just a moment on the scene out the window with its

farms, hills, forest, and so on. Condillac argued that you would initially see something, perhaps just patches of color, throughout the scene, but you would be unable to identify what you were seeing until you had directed attention to different parts of the scene. Condillac s patches of color are what we are calling awareness in the absence of attention. We contrast this level of awareness with that obtained from attending to one of those colored patches and consequently realizing that it represents, say, a meadow in the summertime.

We will argue that in the absence of attention we can, at most, be aware of object attributes but not how they are related. For example, if an object is composed of a red vertical bar and a blue horizontal bar, then, in the absence of attention, we might be aware that there was a vertical bar and a horizontal bar and that there was red and blue. However, we would not know which bar was which color. To be able to relate (or bind) a bar s color to a bar s orientation requires that the bars be attended. To understand why this might be the case we need to consider the binding problem.

Feature Integration Theory, Object Recognition, and Awareness

When we attend to an object, we usually feel that we are aware of multiple features of that object. For example, we might be aware of a round, red, revolving disk. That type of awareness requires that we bind the roundness, redness, and motion to the same object. The neurons that analyze different attributes of an object are often located in different regions of the brain. Consequently, binding features together to form a coherent representation poses a problem. In a world filled with many objects, often in close proximity, how do we know that the red computed in this part of the brain goes with the motion analyzed in this other part? This issue is known as the binding problem. One proposed solution is that it is the act of attending to an object that allows different features of the same object to be conjoined and features from other objects to be excluded. Indeed, this may be the main function of selective attention.

In principle, our brains could have been constructed in such a way that we would not suffer from the binding problem. For example, the optic tectum of the common toad (*Bufo bufo*) contains a class of fly-detector neurons that signal the location of small, moving black dots. The method effectively avoids the binding problem because the toad can detect the fly directly without having to first measure the fly s individual attributes such as its motion, color, and size. Unfortunately, this method allows for the detection of only a small number of different types of objects, as it needs a dedicated group of neurons for each type of object that it is to detect.

Human visual systems (indeed mammalian visual systems, in general) have a flexible ability to represent arbitrary combinations of attributes like color, size, orientation, and so forth. In order to understand the relationship of attributes, these visual systems have had to solve the binding problem.

Our understanding of the structure and function of the visual system (Figure 1) is obtained from multiple sources including neuroimaging techniques like functional magnetic resonance imaging (fMRI) in human observers and more invasive neuroanatomical and neurophysiological methods performed mainly in animal models such as the cat and the monkey.

Visual information flows from the retina to the lateral geniculate nucleus (LGN) of the thalamus. The LGN in turn relays that information to the primary visual cortex (V1) located on the rear surface of the brain, mostly inside the calcarine fissure. From here the pathway divides, with the dorsal and ventral streams being particularly important

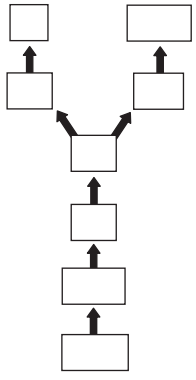


Figure 1 A schematic representation of the ventral (left) and dorsal (right) visual pathways. Also known as the ‘what’ and ‘where’ pathways. Abbreviations are explained in the text.

subdivisions. The dorsal stream includes visual areas V1 and V2, the middle temporal (MT) area, and the medial superior temporal (MST) area. The ventral stream also includes areas V1 and V2 and then proceeds to area V4 and to areas in the inferior temporal (IT) cortex. As one progresses along either stream, the neural activity become less purely stimulus driven, more readily modulated by changes in the attentional state, and increasingly likely to mirror the reported conscious percept.

The dorsal stream is often referred to as the "where" pathway as it is particularly sensitive to spatial information. For example, the MT area is especially sensitive to the motion of an object. To a large extent, the impression of the object's motion is closely related to the activity in this area. Electrical microstimulation of neurons in the MT in macaque monkeys influences their judgment of motion. Damage to the MT can cause akinetopsia, an inability to perceive motion. Sufferers of this condition can report that an object was in one position and is now in another. However, they have no conscious perception of the movement of the object. Conversely, for those that do not suffer from akinetopsia, it is possible to have a conscious percept of motion even when the visual stimulus does not move. For example, if one stares fixedly at a coherent moving pattern, such as a waterfall, and then fixates on a stationary object, the stationary object will appear to move (a motion aftereffect) and MT will be activated. If transcranial magnetic stimulation (TMS) is used to prevent the MT from activating, then the motion aftereffect is not seen.

The ventral stream is often referred to as a "what" pathway, as it is particularly sensitive to the identity of the object. For instance, activity in the IT closely reflects the subject's impression of the object's shape. This was elegantly demonstrated by David Sheinberg and Nikos Logothetis in a series of neurophysiological recordings in the macaque IT. First, they would isolate a neuron and find an image to which it responded strongly and one that did not excite it. Then, they would present one of these images to one of the monkey's eyes while simultaneously presenting the other image to the other eye. The monkey had been trained to pull a lever to indicate which image it saw. As with humans, the percept reported by the monkey alternated between the two images, even though the images themselves were constant, a phenomenon known as binocular rivalry. They found that almost every IT cell responded only when the monkey reported seeing the image that had previously been shown to excite that cell. Crucially, these cells did not respond when the monkey reported seeing the cell's nonpreferred image even though the preferred stimulus was still present on the retina of the other eye. The activity of these cells reflected conscious perception, as opposed to the unchanging retinal image.

It should be stressed that the activity in the MT and IT does not cause the perceptual awareness of motion and shape, respectively. Indeed, when monkeys are rendered unconscious by an anesthetic, MT and IT continue to be active. Instead, it seems that when a monkey is conscious of an object, much of its awareness of motion and shape is reflected by the activity in these areas. For present purposes, the important observation is that when you see a moving object, its shape and motion are typically bound into a single coherent percept. In this physiological framework, the binding problem is the problem of understanding how motion information from the MT, shape information from the IT, and various other bits of information from other visual areas come to be unified in a bound percept.

Feature Integration Theory

Anne Treisman's feature integration theory (FIT), first proposed in 1980, holds that attention is critical to the formation of bound representations of objects and, by extension, it proposes that attention is critical to our conscious experience of those bound representations. In FIT, following the understanding of the visual neurophysiology given above, the visual system first decomposes the visual scene into its composite features, arrayed in a set of "feature maps." The preattentive description of a scene or object comprises a list of such features. The term "preattentive" has been controversial, but it can be operationally defined here as the representation of a stimulus before selective attention is directed to that stimulus.

In FIT, the approximate position of each feature is recorded on its preattentive feature map. For example, if the visual scene contains two red

objects, the feature map corresponding to redness would be activated at two points roughly corresponding to the locations of the red objects. If each feature were associated with a precise region in space, this might solve the binding problem. Features that correspond to the same region in space could be automatically conjoined, thus guaranteeing veridical perception. Unfortunately, the location of many features is measured in an imprecise fashion. For example, the smallest receptive fields in IT, the region whose activity correlates well with shape perception, have a spatial extent of a few degrees of visual angle. Within this region, the cell will respond to an object in an approximately translation invariant manner. Thus, a neuron in the IT cannot signal the location of a particular shape with a precision of better than a few degrees, while the perception of coherent objects requires a much finer resolution.

Because of the poor resolution of these feature maps, if two objects are close together, then there is the potential that the features from one object may become conjoined with the features of the other object thus creating a percept of an object that did not in fact exist. For example, if the visual scene contains a red vertical bar and a blue horizontal bar then one might see a blue vertical bar and a red horizontal bar. Such inappropriate combinations of features are known as illusory conjunctions. FIT suggests that attention hinders the formation of illusory conjunctions.

Supporting this assertion is a series of classic experiments by Treisman and her colleagues showing that, if attention is occupied elsewhere, illusory conjunctions are, in fact, reported. In one version of the experiment, observers viewed a display of five characters aligned horizontally. The outer two characters were always digits and the inner three characters were always letters. While the digits were always black, the letters were colored. The observer s primary task was to name the digits. After doing that, the observer reported the letters and their associated colors. When the display was presented sufficiently rapidly, observers would often report seeing an incorrect conjunction of a color and a letter. For example, if the display contained a red X and a green T, they might report seeing a red T. Crucially, these illusory conjunctions occurred at a much higher rate than could be attributed to the observer simply misperceiving a given feature. Generally, the observer correctly perceived the features present in the display. It was the conjoining of features that proved to be problematic.

When asked to report how confident they were that they had actually seen an object, observers were just as confident when they reported seeing an illusory conjunction as they were when they correctly reported the features of an object. Indeed, although all observers were told that the digits would always be black (and in fact always were) about half the observers spontaneously reported that the digits sometimes appeared to be colored, sometimes even going as far as to argue with the experimenter about the issue! This raises an interesting problem in the study of attention and awareness. In tasks of this sort, one can only ask about what was seen, after the fact. If one asks about the current status of a visible object, the observer will attend to it in order to answer the question and will be unable to give an accurate report of the unattended state. Nevertheless, the phenomenology of illusory conjunctions does show that, within a fraction of a second of the disappearance of a display, observers can be quite convinced that they have seen something that was not, in fact, present. Subsequent studies have shown that illusory conjunctions can be perceived even when the subject attends to the objects, especially if the objects are perceptually grouped. Clearly, attention does not always succeed in solving the binding problem.

There is neuropsychological evidence, from studies of patients with Balint s syndrome, which supports the idea that attention can inhibit the formation of illusory conjunctions. This syndrome occurs when both the left and right parietal lobes are damaged. As these areas help govern the deployment of attention, such patients have great difficulty in directing their attention to a given object, resulting in the inability to perceive more than one object at a time. As would be expected, they are also prone to suffer from illusory conjunctions, experiencing them even when the image is displayed for several seconds.

Neurophysiological support also comes from work by Robert Desimone and colleagues. They performed a series of extracellular studies in area V4 of the macaque monkey that have shown that attention can help solve the binding problem.

First, they would find a stimulus that, when presented on its own, would elicit a strong response from the neuron in question (the preferred stimulus), and another that would elicit only a weak response (the nonpreferred stimulus). They would then present both stimuli simultaneously so that both were within the neuron s receptive field. In the absence of attention, the cell would simultaneously respond to both stimuli, with its response (spike rate) lying between that generated by each stimulus when presented on its own. In other words, the response reflected contributions from both stimuli, meaning that the cell could not distinguish between the two. However, when the monkey attended to one of the stimuli, the situation changed and the cell responded primarily to the attended stimulus. Specifically, when the monkey attended to the preferred stimulus, the cell would respond strongly, but when the nonpreferred stimulus was attended, only a weak response was elicited. In this case, attention is able to solve the binding problem, at least at the neuronal level, by shrinking the receptive field of the cell to include just the selected item, thereby removing the influence of the unattended item.

This constriction of the receptive field does not explain how signals about one feature analyzed in one cortical area can be bound to signals about another feature from another area. Other mechanisms have been suggested to account for this aspect of binding. Several of these are based on the idea that neurons in different cortical areas that respond to the same object synchronize their activity, so that they create action potentials at the same time. Consequently, a third brain area could determine whether two neurons in two different parts of the brain are responding to different features of the same object by being sensitive to this synchrony. As attention is known to increase neural synchrony, theories based on synchrony are consistent with the notion that attention is needed to solve the binding problem.

Features

While it is easy to say that the visual system decomposes a visual object into its constituent features, it is harder to be precise about what this statement might mean. In particular, there is imperfect agreement about the list of features that might be available to be bound. Various tests have been proposed, most of them based on the premise that the attributes on this list can be analyzed in the absence of attention. For example, if an item is the only item in a display that has a particular feature, that item will tend to pop-out of the display, summoning attention (as long as the other items are not too similar to that target item nor too different from each other). The set of items that pop-out in this manner is one definition of features. Figure 2(a) shows a case where the target has a unique feature. It is the only red object in the scene. Consequently, it pops-out and can be located very quickly, independent of the number of other items. Conversely, in Figure 2(b), the target and distractors share the same features. The target is a rectangle that is green on the left but red on the right, whereas the distractors are red on the left and green on the right. In this case, finding the target is a slow process.

Texture segmentation is another test. Consider two regions of a display, one with a putative feature and the other without. If a border between those regions can be effortlessly detected, one could declare that there is a feature difference that permits the segmentation. Unfortunately, these and other methods for identifying features agree imperfectly. It is quite clear that some attributes, like color, motion, and orientation pass all the tests. Other attributes (e.g., various aspects of form) are more problematic.

Feedforward Models of Object Recognition

Even if attention is needed for binding, it is not necessary and probably incorrect to hold that

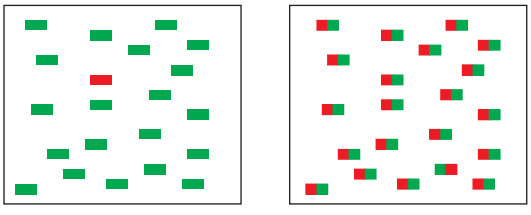


Figure 2 Two visual search experiments. (a) The target (the red rectangle) is easy to find. (b) The target (the green-on-the-left-red-on-the-right rectangle) is hard to find.

attention is always needed for object recognition. A class of feedforward models of object recognition shows how some quite sophisticated object recognition could occur without explicitly invoking selective attention. As we will see later, this fits well with evidence that humans have some ability to detect objects in one part of the field even when their attention is occupied elsewhere. For example, Maximillian Riesenhuber and Tomaso Poggio developed a theory of object recognition based on an idealization of the hierarchy of the monkey visual system. As one progresses through the monkey visual system, cells become selective for increasingly complex visual stimuli. For example, in the LGN some cells have an on-center off-surround receptive field organization. The stimulus that most excites these cells is a spot of light on a black surround. On entering the primary visual cortex, we find cells whose optimum stimulus is more complex, perhaps a bar of a particular orientation and length. In V2, there are cells whose optimum stimulus is two bars in a particular configuration. In V4, some cells are most excited by a collection of bars joined together in a particular manner. Riesenhuber and Poggio were able to build a feedforward model that was able to explain how these selectivities were generated. Crucially, this was achieved without invoking any feedback mechanisms. Since attention must be mediated by feedback, they argued that this showed that at least some recognition can occur in the absence of attention. The original model was applied to shapes that resembled bent paper clips. In subsequent work, they and others have developed models that can recognize objects like cars and faces. Does this mean that attention is unnecessary? Probably not, since the models tend to fail when there are many objects in the display. These models do show that it is possible, in theory, to have some degree of recognition without attention. This, in turn, makes it plausible that one might have some awareness of an object, even if that object is not the target of selective attention.

Reverse Hierarchy Theory

Reverse hierarchy theory (RHT), proposed by Shaul Hochstein and Merav Ahissar, is an example of a model combining feedforward and feedback components. The feedforward component is similar to the Riesenhuber and Poggio model. It explains how the hierarchy of the visual system allows for visual scenes to be processed to some degree in a feedforward manner in the absence of attention. It is the feedback component that allows for more detailed perception to occur (hence reverse hierarchy). For example, it suggests that to appreciate fine differences in orientation, the brain would deploy feedback from a high-level representation of the stimulus, in order to pay attention to the detailed orientation information held in cells in the early visual cortex. One interesting aspect of this proposal is that it suggests that the high-level information (e.g., animal or face) might reach awareness before information about low-level features.

To summarize the argument to this point; precise binding of features to objects is a problem that the visual system seems to solve by the use of selective attention. It follows that attention is required for awareness of those bindings and for awareness of object identities that rely on those bindings. However, the example of Condillac s chateau indicates that there will be awareness of something in regions not yet visited by selective attention. Moreover, feedforward models show that, at least in theory, the unattended awareness of something need not be limited to raw local features. Some quite sophisticated analysis and awareness might be possible away from the current focus of attention. It is to that awareness without attention that we turn to in the next section.

The Relationship between Attention and Awareness

In this section, we wish to distinguish between the hypothesis that some awareness occurs outside of the current focus of attention and the hypothesis that we are aware only of the current contents of attention. Recall the phenomenon, described above, of illusory conjunctions in which observers correctly report the colors and letters in a display but fail to correctly report which letter goes with which color. It could be that the experience of unbound colors and letters represents awareness without attention. However, there is a contrary

point of view. Perhaps the imperfect awareness of the letters arises from imperfect attention to the letters. There is no guarantee that naming the digits in an illusory conjunction experiment withdraws all attention from the letters. Perhaps if all attention had been really withdrawn from the letters then the observers would not have been able to report any features of the letters at all. That would be the prediction if awareness cannot occur in the complete absence of attention.

When Awareness Requires Attention

Although we will argue against this extreme viewpoint, we will discuss a set of phenomena that have been used to argue for this strong link between attention and awareness: inattentional blindness, change blindness, and the attentional blink. These topics are more extensively discussed in other articles of this encyclopedia.

Inattentional blindness was first described by Arien Mack and Irv Rock. They had observers performing an attentionally demanding perceptual task (e.g., which of the two lines is longer?). On one critical trial, the briefly presented display contained an unexpected item. Observers were frequently unable to report that it had been presented. Any awareness of that item left no trace that could be reported after it was gone. Perhaps, attention to the primary task, prevented irrelevant items from ever rising to conscious awareness. As an experimental tool, one problem with this task is that it produces only one trial per observer. Once you ask about the unexpected item on one trial, other unexpected items on other trials tend to be successfully reported.

Change blindness is a more resilient phenomenon. While the phenomenon was initially discovered in the late 1950s and early 1960s, a major renaissance on the topic emerged in the mid-1990s. Dan Simons and Dan Levin, as well as Ronald Rensink and Kevin O Regan and their colleagues presented observers with complex natural scenes (e.g., a photo of an airplane on the tarmac) and measured the ability to detect fairly large changes to these scenes (e.g., the plane s engine disappearing and reappearing). Critically, the visual transient generated by the change was masked by an eye movement, a brief blank interval, or some other visual transient. Observers thus had to actually detect the change in the image, rather than just the transient caused by the change. Change detection under these circumstances turns out to be very difficult. Observers could fail to notice changes even though they spent many seconds examining both versions of the display. If an object was attended during the transition between two frames, the change could be noted. Otherwise observers were unable to report it. One interpretation of these data would be that the observer was only truly aware of the currently attended object, while the apparent awareness of the rest of the display was, in some sense, an illusion.

The attentional blink, originally described by Donald Broadbent, and later characterized by Jane Raymond, Kim Shapiro, and Karen Arnell, is a quite different phenomenon that might point to a similar conclusion. In a typical attentional blink experiment, observers monitor a stream of images, letters, for example, appearing at fixation, at a rate of one every 100 ms. The observers are looking for particular targets, say E and X. At this rate of presentation, an observer can easily report a single target letter appearing anywhere in the stream. However, if there are two targets, the second one is much more likely to be missed if it appears 200 500 ms after the first. This is not simply a matter of perceptual masking from the first target to the second. Given the same stream of letters (e.g., J W E B P X L), the target X will be easily detected if the observer does not have to report the E, but will likely be missed if she does. Something about the attention to the first target causes an attentional blink that makes it harder or impossible to report the second. Interestingly, blinked items can be shown to have effects on the observer. A blinked word can produce semantic priming effects indicating that it has been read. Perhaps the type of attention that is tied up by the first target in an attentional blink display is the type of attention that permits awareness of an object.

Phenomena such as inattentional blindness, change blindness, and the attentional blink provide some evidence that in the absence (or, at least, near-absence) of attention, the observer may be unable to recognize or even see an object. In its strongest form, this argument proposes a tight

linkage between selective attention and visual awareness. The fact that in some situations an observer is unable to report an unattended object does not prove that she is never able to do so. In the following, we describe phenomena that indicate that the proposed strong linkage of attention and awareness is too strong. The linkage of attention, binding, and awareness is challenged by studies that seem to show that there can be some degree of recognition and awareness of objects in the near-absence of selective attention.

When Awareness Does Not Require Attention

In a series of experiments done by Fei-Fei Li and her collaborators, observers performed an attentionally demanding visual search task in one location while concurrently monitoring another portion of the visual field for some class of targets (e.g., animal or vehicle). The goal was to tie up selective attention with the search task and to determine what, if anything, could be detected elsewhere at the same time. Some tasks (e.g., determining if a red square was to the right of a conjoining green square or vice versa) are profoundly disrupted when attention is thus engaged. Interestingly, however, detection of the presence of animals or vehicles in a briefly presented scene is no worse when selective attention is occupied than when it is not.

Note that detecting that a scene contained an animal is not the same as determining exactly what that animal was. Observers in Fei-Fei Li s experiments were not necessarily sure what type of animal they had detected or where it was in the display. Moreover, when Karla Evans and Anne Treisman asked the observers to find animals in a rapid sequence of scenes, they found that the observers were significantly impaired if the stream also contained humans. In the near-absence of attention, some image statistics seem to permit awareness of the presence of an animal (human or other) or a vehicle, but it would be going too far to argue that these data make the proposed feature-binding role of attention unnecessary. Awareness of this specific animal or vehicle, which would presumably require feature-binding, appears to require attention.

A number of other phenomena also challenge the attention-binding-awareness linkage. For instance, Mary Potter and others have shown that high-order representations (i.e., gist) can be accessed very rapidly from natural scenes presented at rates of up to 10 per second, far too little time for selective attention to be directed to more than a small handful of items in that scene. Gist in this case refers to a broad categorical label for the scene: beach, kitchen, and so on. Does this challenge the relationship of attention to binding to object recognition? It certainly would if recognizing the gist of a scene involved promiscuous binding of multiple objects without attention. Alternatively, these results might be demonstrating that performance of this scene-categorizing task does not require binding. If information in the unbound feature statistics could support performance of the task, then it would not be necessary to assume binding. Aude Oliva and Antonio Torralba have shown that this can be done, in principle, for scenes. They have created filters whose output can be used to put a label on a scene (indoor, outdoor urban, beach, etc.) based on the raw image statistics of the scene; without the need to parse the image into objects, regions, and so forth. This unbound analysis is not adequate to identify a specific scene, for example, Crane s Beach in Ipswich, MA, but it could provide the gist of a scene in the near absence of attention.

As with the detection of animals with selective attention occupied elsewhere, findings of this sort suggest that awareness of a visual stimulus is not a unitary, all-or-none experience. Awareness of an unattended scene may be different than awareness of a well-attended scene, but the scene is seen in both cases. This distinction may be reflected in neurophysiological findings showing that categorization and perception are mediated by different cortical areas. Specifically, those neurons that can categorize a target are not necessarily the same neurons that can, say, locate it in the image. Neurons in the visual cortex are able to signal a target s location, but are generally insensitive to whether an object is a target or not (i.e., they cannot categorize it). However, David Freedman and colleagues have elegantly demonstrated that neurons in the prefrontal cortex can be highly sensitive to the categorical status. They used a photographic morphing

technique to create a picture of an animal that represented a combination of a cat and a dog. By varying the morphing parameters, they could vary the similarity between this computer-generated animal and prototypical cat and dog images. They presented a series of these computer-generated pictures to monkeys that had previously been trained to indicate whether each picture more closely resembled a cat or a dog. They found cells in the lateral prefrontal cortex that encoded the monkey s categorization. Such cells responded similarly to images that belonged to the same category, even when the images appeared very different. Conversely, the cells responded very differently to images that appeared very similar, but which belonged to different categories. These cells therefore responded to the categorization of the images, as opposed to the visual image itself.

We have seen that a strict linkage of attention, binding, object recognition, and awareness leads us to a theory that does not have room for the full range of phenomena. Still, it seems likely that attention is required for the recognition of specific objects and that, as Condillac argued, this act of attention changes the state of our visual awareness. Not being privy to more recent developments, Condillac does not tell us what he thinks we would see if we were performing an attentionally demanding task at fixation when the curtains were thrown wide, revealing the scene outside the chateau for the first time. However, it seems likely that his answer would have been much the same. You would have some impression of the outside world, but you would not understand what you were seeing until you had attended to the scene.

As he became familiar with the current literature, Condillac might agree that some information about the gist of the scene might be available in that first moment, but his description of the initial state of awareness, modified by subsequent attention, would remain essentially unchanged.

When Attention Does Not Imply Awareness

It is hard to gain any introspective access to this with real scenes because we are too good at analyzing them. However, the very unreal scene of Figure 3 may serve the purpose.

Note that when you first look at this scene, you are aware of the patches of color and orientation across the entire stimulus. More than that, you are aware of some structure in the scene. There are plusses everywhere with a scattering of items with more than four line terminations. There are blue and yellow objects in the upper left and red-green elsewhere. However, if you are asked to detect red vertical components, you will need to direct your attention to specific items over a period of time and, having deployed your attention, your awareness of the stimulus will change. Now you will find that the red and green plusses are not all the same. There is a region of red verticals in the upper right and an isolated example at the bottom center.

You have awareness with and without selective attention. What about attention without awareness? Can you select and bind an object without being aware of it? Returning to the figure, imagine you are asked to locate the five-pointed item. It is entirely possible that you had already attended to

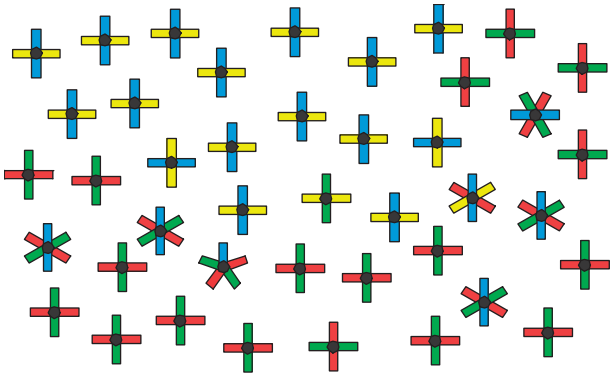


Figure 3 A display that gives some idea of one's level of visual awareness. Please see the text for details.

that object during a search for red vertical without becoming aware of it. Thus, it seems possible that you can attend to an object without ever becoming aware of it. The point is tricky since one could argue that you were aware of the five-pointed item when you putatively attended to it, but you forgot about it prior to being queried about it. It is often difficult to distinguish between having been unaware and being amnesic.

Studies of blindsight patients provide converging evidence for the hypothesis that attention is not always sufficient for visual awareness. In blindsight, damage to the primary visual cortex results in a condition where patients report being unaware of part of the visual field (the unaware area). Interestingly, when asked to guess what stimuli are located in the unaware area, patients can perform at above chance levels. Furthermore, when spatially cued to the unaware area, they exhibit speeded discrimination of targets that subsequently appear in that area, demonstrating that they can attend to stimuli that they cannot see. Evidently, attention does not necessarily result in awareness.

Postattentive Awareness

Consider the red-vertical plus at the bottom center in Figure 3. We can posit that you had some representation of it before it was selected. That representation can be called preattentive. You also had a representation of the plus while it was attended. Call that an attended representation. What is the representation of that plus when you then move your attention to the blue-vertical star, up and to the right? This can be called the postattentive representation. A series of experiments show that changes to the plus (e.g., changing it to red-horizontal, green-vertical) once attention has been shifted elsewhere, will go unnoticed, until attention is directed back to the item (providing that the transients produced by such changes are masked by, say, a blink, a saccade, or a visual transient). Observers show no more awareness of the current binding of features in a postattentive object than in a preattentive object. At the same time, you the observer are aware that there was a red-vertical plus at that location so, in that sense, your postattentive awareness of that particular plus is different than your preattentive awareness.

Of course, you would also be aware of the plus, in that sense, if the lights went out and you relied on memory. The already difficult topic of visual awareness becomes more difficult once we admit a role for memory. Your awareness of a familiar face is different from your awareness of a new face. That difference is tied to memory and it is hard to know whether one should consider this to be part of the definition of visual awareness. Whatever one concludes from this question, the postattentive vision research suggests that selective attention affects postattentive awareness of an object through memory and not through some sort of persistent binding that continues once selective attention is disengaged from an object.

Awareness of Awareness

The account outlined above suggests that we think we are aware of more than we actually are. We greatly overestimate our own awareness. Our naïve impression of our visual awareness is that we are aware of a large visual scene at a high resolution. Yet this is not so. At any moment, the only part of the visual scene we can see in high resolution is the small area around the current point of fixation. A particularly striking demonstration of this (beloved by people who sell eye trackers!) is to use an eye tracker to monitor the observer s point of fixation. The observer s task is to read some text presented on a computer monitor, similar to that shown in Figure 4(a).

At some point, the eye tracker salesperson pushes a button and, during the observer s next saccade, the letters in the display become jumbled except for those in the words near the point of fixation (Figure 4(b)), which, in this figure, are assumed to be in the top-left corner. Every time the observer saccades to a different point, the letters in the words near that point become unjumbled, while the letters in all other words either become or remain jumbled. Provided all changes to the display occur during a saccade, the observer is unaware of the scrambling. She reports that she is simply reading a normal text. Similarly, if the image away from fixation is appropriately blurred, an observer would be unaware of this degradation and will have the impression of looking at the usual, apparently well-focused scene.

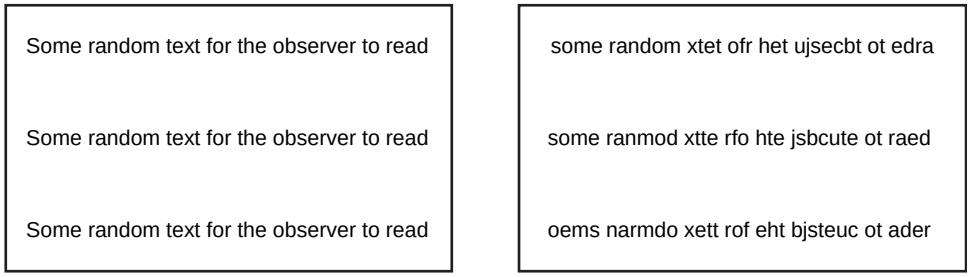


Figure 4 The observer starts to read the text shown in (a). During an eye movement, all the text that is not near the observer's point of fixation (assumed to be at the top-left corner) becomes jumbled (b). The observer does not notice the change and so cannot differentiate between (a) and (b), thereby demonstrating how limited visual awareness really is.

The observer in these illustrations is aware of something, but it turns out not to be a true assessment of the contents of the current visual representation. Beyond a simple contribution of memory to awareness, this indicates a contribution of theory to awareness. In the reading example, the observer is not aware of the scrambling occurring away from the fixation. Wherever she fixates on the page, the letters form themselves into English words. It is a reasonable theory that the page consists of readable English words and our observer s awareness incorporates that theory. Her visual awareness is affected by what she thinks she knows. Returning to the role of attention in awareness, we see that selective attention alters not only the awareness of the attended object, but potentially, the awareness of other unattended objects, potentially divorcing that awareness from the actual perceptual facts.

Awareness of Attention

If one were inclined to propose a tight linkage of attention and awareness, one might propose not only that observers are only aware of the objects of attention, but that observers are aware of all of the deployments and, thus, all the objects of attention. However, in visual search experiments, it is estimated that observers can attend to 20 50 items per second. This rapid selection seems to occur without a clear awareness of which items in a display have or have not been selected. At least observers do not use any such awareness to guide their search. For example, Todd Horowitz and Jeremy Wolfe conducted a visual search experiment in which all the items in the display were randomly relocated every 111 ms. This made it impossible for observers to keep track of which items they had already attended to. Surprisingly, the search was no less efficient in this case than in the control condition where the items were static. This showed that, at least in some cases, visual search had no memory. Observers acted as if they were unaware of what they had attended to.

Conclusions

In this article, we have considered the relationship of conscious awareness and attention from the perspective of vision. Following Condillac, we found it helpful to differentiate between the awareness that results from attending to an object or group of features and that which occurs in the absence of attention. Condillac is more famous for his statue than for his chateau. He asked his readers to imagine the mental life of a statue with no senses. In his honor, we can imagine a statue with senses, but without attention or, perhaps better, with attention disabled. In the absence of attention, the evidence indicates that our statue would retain some visual awareness, but would be unable to form any percepts that would require the binding of two of more features. This level of awareness might allow our statue to classify scenes (beach, mountains, etc.) or declare that they did or did not contain an animal, but would not allow it to determine specifics such as which animal occurred in a given scene. If we now endow this statue with selective attention, it can then solve the binding

problem, and have a more complete awareness. Specific objects can be selected, perceived, and identified. If we now allow the statue to have a memory, the statue will know that a given object was at a particular location and, in the absence of contradicting information, the statue is likely to assume that the object continues in that location. If we add a theory-building capability, the statue can generalize from the fact that all selectively attended samples are seen in sharp focus to the assumption that all objects really are in sharp focus and then use this assumption to modify visual awareness accordingly. The statue now has an approximation of human visual awareness.

See also: Attention: Change Blindness and Inattentional Blindness; Mind Wandering and Other Lapses; Neglect and Balint's Syndrome; Neuroscience of Volition and Action.

Suggested Readings

Evans K and Treisman A (2005) Perception of objects in natural scenes: Is it really attention free. Journal of Experimental Psychology: Human Perception and Performance 31: 1476–1492.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291: 312–316.

Freedman DJ, Riesenhuber M, Poggio T, and Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. Journal of Neuroscience 23: 5235–5246.

Hochstein S and Ahissar M (2002) View from the top: Hierachies and reverse hierarchies in the visual system. Neuron 36: 791–804.

Horowitz TS and Wolfe JM (1998) Visual search has no memory. Nature 394: 575–577.

Kentridge RW, Heywood CA, and Weiskrantz L (2004) Spatial attention speeds discrimination without awareness in blindsight. Neuropsychologia 42: 831–835.

Li FF, VanRullen R, Koch C, and Perona P (2002) Rapid natural scene categorization in the near absence of attention. Proceedings of the National Academy of Sciences USA 99: 9596–9601.

Moran J and Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229: 782–784.

Potter MC and Faulconer BA (1975) Time to understand pictures and words. Nature 253: 437–438.

Riesenhuber M and Poggio T (1999) Are cortical models really bound by the 'binding problem. Neuron 24: 87–93.

Sheinberg DL and Logothetis NK (1997) The role of temporal cortical areas in perceptual organization. Proceedings of the National Academy of Sciences 94: 3408–3413.

Treisman AM and Gelade G (1980) A feature-integration theory of attention. Cognitive Psychology 12: 97–136.

Treisman A and Schmidt H (1982) Illusory conjunctions in the perception of objects. Cognitive Psychology 14: 107–141.

Biographical Sketch

Piers Howe graduated as an exhibitioner from Oxford University in 1998, with a masters in physics. Winning a Presidential University Graduate Fellowship, he obtained his PhD from Boston University in 2003, under the guidance of Stephen Grossberg. His PhD thesis was titled Cortical mechanisms of depth and lightness perception: Neural models and psychophysical experiments. He then worked as a Helen Hay Whitney postdoctoral fellow with Margaret Livingstone at Harvard Medical School before moving on to Brigham and Women s Hospital to work as a research fellow with Todd Horowitz and Jeremy Wolfe. His research has involved a variety of techniques including computational modeling, macaque neurophysiology, human fMRI, and human behavioral experiments. He has published articles on lightness perception, motion perception, depth perception, and

visibility. More recently his focus has shifted to visual attention and to devising computational techniques for determining brain connectivity from fMRI data. He is a member of the Harvard fMRI Center for Neurodegeneration and Repair. He has taught a psychology course at the University of Massachusetts, Boston, MA and a cognitive modeling course at Boston University. He lives in Boston, MA.

Karla Evans graduated summa cum laude from Trieste University, Italy, in 1998, with a degree in psychology and went on to obtain her PhD in 2007, from Princeton University, studying with Anne Treisman. Her PhD thesis was entitled Crossmodal interactions of corresponding auditory and visual features. Evans then became a postdoctoral associate at the McGovern Institute of Brain Research at MIT. During that period, she published papers on natural scene perception, crossmodal correspondences, and affective person knowledge and face perception. She is currently a research fellow working with Jeremy Wolfe and Todd Horowitz. Karla Evans was a recipient of a science/engineering fellowship, Princeton University and the Summer Institute Fellowship in Cognitive Neuroscience at Dartmouth College. She lives in Somerville, MA.

Riccardo Pedersini graduated cum laude from the University of Padova in 2001, with a degree in experimental psychology. He continued in Padova and Amsterdam obtaining his European masters in mathematical psychology in 2003, and his PhD in perception and psychophysics in 2006, from the University of Padova. His PhD was obtained under the guidance of Prof. Elisabetta Xausa and Prof. Marc Le Menestrel. He visited the University Pompeu Fabra in Barcelona and Sussex University in Brighton, and he taught several graduate courses at the faculty of medicine in Padova. He was a two-time regional winner of the Mathematics Olympiad. In addition to two Marie Curie fellowships, he won scholarships from both the University of Padova and the European Community before moving to the United States, to work as a research fellow with Todd Horowitz and Jeremy Wolfe. His research interests include decision theory, game theory, nonlinear dynamical systems, and perceptual decisions.

Todd Horowitz obtained his BS in psychology from Michigan State University in 1990. He then went on to study under Anne Treisman at the University of California, Berkeley, receiving his PhD in cognitive psychology in 1995, for his thesis titled Spatial attention: Inhibition of distractor locations. In 1995, he came to Brigham and Women s Hospital and Harvard Medical School for postdoctoral work on selective attention and circadian rhythms with Charles Czeisler and Jeremy Wolfe. In 2000, he joined the faculty of Brigham and Women s and Harvard Medical School, where he is currently an assistant professor of ophthalmology. His work is currently funded by the National Institutes of Health. He has published over 30 papers on visual attention, visual search, multiple object tracking, and circadian rhythms in journals such as Nature, Psychological Science, Neuropsychologia, American Journal of Physiology, Psychonomic Bulletin and Review, and Journal of Experimental Psychology. He is currently a consulting editor for Perception & Psychophysics, and a member of the Association for Psychological Science, the Psychonomic Society, and the Vision Sciences Society. He lives in Medford, MA.

Jeremy Wolfe graduated summa cum laude from Princeton in 1977 with a degree in psychology and went on to obtain his PhD in 1981 from MIT, studying with Richard Held. His PhD thesis was entitled On binocular single vision. Wolfe remained at MIT until 1991. During that period, he published papers on binocular rivalry, visual aftereffects, and accommodation. In the late 1980s, the focus of the lab shifted to visual attention. Since that time, he has published numerous articles on visual search and visual attention. In 1991, Wolfe moved to Brigham and Women s Hospital and Harvard Medical School where he is a professor of ophthalmology. The lab is funded by the US National Institutes of Health, Air Force, and Department of Homeland Security. Wolfe teaches psychology courses at MIT and Harvard. Jeremy Wolfe is a past-president of the Eastern Psychological Association. He won the Baker Memorial Prize for teaching at MIT in 1989. He is a fellow of the AAAS, the American Psychological Association (Div. 3 and 6), the American Psychological Society, and a member of the Society for Experimental Psychologists. He lives in Newton, MA.

Michael Cohen graduated from Tufts University in 2007, with a BA in philosophy and his minor in cognitive science. While at Tufts, he worked extensively with Daniel C Dennett on the problem of consciousness and its relation to visual attention. He previously worked in the Tufts NeuroCognition Lab under Philip Holcomb, where he used ERPs to investigate how the brain processes conceptual anaphors. He is presently a research technician in the visual attention lab at the Brigham and Women s Hospital. His current research focuses on multiple object tracking, attentional capture, and visual short-term memory.

Autobiographical Memory and Consciousness

M A Conway, University of Leeds, Leeds, England

Introduction

The term 'autobiographical memory' refers to long-term memory for personal experiences and personal knowledge of an individual's life. A specific autobiographical memory usually consists of at least one detailed memory of a personal experience (an episodic memory) and various associated items of knowledge. Consider a typical example of this mix of episodic memory and autobiographical knowledge, taken from a web survey of self-defining memories:

> I was about to go away for a year to Chile to work as an English teacher when I discovered that my dad had to have a major heart operation. It was a dream of mine to go, but that changed when my dad went into hospital. I went to visit him to tell him that I wouldn't go. The hospital was really hot, and smelt of years of disinfectant. Dad was in a room by himself, looking very gaunt and worried. We talked about anything but my trip until I was due to go, but before I could say anything he got up and hugged me, saying that he would be more upset if I didn't go. We both started to cry, I couldn't speak. Dad said it would help him to know that I was fulfilling a dream. I couldn't believe how much he loved me. He walked me to the lift and had to wrench himself away. I said ''Thank you Dad, I love you.'' and the doors closed.

This memory description has all the major features of a specific autobiographical memory: sensory–perceptual episodic details, affect, information about goals, personal interactions, and more conceptual autobiographical knowledge that contextualizethe episodic details in the rememberer's life. Here we will explore how these different types of knowledge give rise to different types of conscious feelings of memory, how these may function in everyday life, and what occurs when they malfunction.

Cognitive Feelings: Remembering, Familiarity, and Knowing

When a specific autobiographical memory comes to mind then a rememberer has 'recollective experience.' That is, they experience remembering consciously and have what has been termed 'autonoetic consciousness.' Typically images enter conscious awareness, often visual in nature, attention turns inward, other highly specific knowledge may also feature too, and there is a strong sense of the self in the past. Additionally there is a distinct 'feeling of remembering'; a feeling that what is in consciousness is a memory. Such feelings are part of a class of mental experience that have been termed 'cognitive feelings.' Cognitive feelings let us experience our mental states and without them we would have to, perhaps consciously and laboriously, infer what state we were in at any given time. Thus, the feeling of remembering, triggered by mental content such as visual images of past experiences, lets us know automatically and experientially that we are remembering – no further inferential reasoning is required to determine the state. The conscious feeling of remembering may be important too in convincing a person that they are indeed remembering and then to act on that. To give a trivial but not uncommon example, if a person remembers that they locked the door when they left the house, may be a visual image comes to mind and perhaps other details, and they have a feeling of remembering then they almost certainly would not go back to check. If, however, the feeling of remembering was not triggered by the knowledge brought into consciousness then determining whether what is in consciousness is in fact a memory and not, for instance, a generic visual image or set of images is more difficult and in some cases perhaps not possible at all. In which case, the probability of repeating one's actions is increased. The feeling of remembering and autonoetic consciousness may then have quite powerful, direct, and important effects on behavior.

The conscious experience of remembering can be contrasted with other states of memory awareness. For example, feelings of familiarity and knowing occur when autobiographical knowledge is brought to conscious awareness without associated

episodic memories. Remembering a school attended, the name of a friend, a work project, a holiday, a repeated event, etc., without remembering any single specific event is associated with familiarity and knowing but not with recollective experience. This type of remembering features what has been termed 'noetic consciousness.' Noetic consciousness does not feature specific representations of the self in the past such as those that are represented in episodic memories. Noetic consciousness cannot then trigger the feeling of remembering, although as will we shortly see this may occur in malfunctions of memory.

Cognitive feelings characteristic of noetic states appear to be related partly to orienting processes and partly to metamemory functions. A person's face for example might trigger a feeling of familiarity and lead to a metamemory inference that that person has been recently encountered. This may prime the memory system for detail recollection if current tasks come to require that, on the other hand, it may not lead to full recollection so reduce the attentional costs that would be incurred by constructing a full and detailed memory in consciousness. This 'feeling of familiarity of recent occurrence' may then serve to optimize cognitive performance and keep processing online and task-orientated while still supporting some recognition of past occurrence.

In a similar way a more pervasive feeling of familiarity triggered by features of an individual's habitual environment may facilitate orientation to that environment and fluent processing of it. In other words this 'feeling of familiarity of the habitual' may reduce attentional and processing demands on cognition and consciousness by signaling what does not need close attention and more detailed processing. It is particularly interesting that many brain-injured patients, who are disoriented in time and space, often claim that everything is unfamiliar, as though the feeling of familiarity triggered by the habitual had in some way been dysfacilitated. There may well be further shades of cognitive feelings here related to familiarity in the temporal dimension. For example, a conscious feeling triggered by the beginnings and endings of events, the feeling that events are proceeding fluently, feelings that one's autobiographical memory is continuous, anticipatory feelings of imagined future events (which share activation of many of the brain areas that are activated when remembering), are all memory-related cognitive feelings that await further investigation.

The feeling of knowing may be more associated with metamemory functions than with orientation and again there seem to be two functions of this feeling: one is to let us feel what it is we know and other is to let us feel what we might know. For example, during the process of learning at universities, students initially rely on memories of knowledge presented in lecture, encountered in books, in conversation with teachers, or even in conversation with each other. Later, however, as they acquire knowledge they no longer depend on memories of times when knowledge was encountered and instead come to 'just know.' That is they just know, if they are psychology students, that the behaviorists rejected introspectionism, Freud discovered the unconscious, and in a with-subjects design the same participants take part in all the experimental conditions, and so on. Verification of these 'facts' is accompanied by a 'feeling of knowing' – a feeling that lets the individual know what they know experientially and without having to engage in further extended processing. People just know that Paris is the capital of France, that last year's holiday was on a Greek island, that they once lived in London, and so on. Conversely there is knowledge in long-term memory that is available but not currently accessible to consciousness, usually because an appropriate cue cannot be located that would activate the knowledge to a level where it could enter consciousness and be experienced as known. Nonetheless extensive evidence shows that people can discriminate on the basis of 'a feeling of knowing' between items that they would be able to recognize and even remember with an effective cue from those they would not be able to recognize or remember even with a cue. Presumably this is because the knowledge in long-term memory is sufficiently activated to trigger a feeling of knowing, but not support incorporation of that item in a conscious representation. Were such an available item to become consciously accessible, then there is nothing to preclude it triggering the 'just know' feeling state. The relation between the feeling of knowing for temporarily inaccessible items and the feelings these same items trigger when they eventually do enter consciousness is not known.

Memory Feelings in Everyday Life

One of the key features of conscious feeling states is that we act on them. If we remember the items we intended to buy at the supermarket we buy them, if we remember where we parked the car we go to that location, if we remember the emails we had to send, calls to make, letters to write, we do not repeat these. Similarly if we know we are meeting a colleague for lunch, if we know we completed a work project, we act on these conscious states of remembering. But what if, instead, we had not completed the project but only imagined it? Or perhaps a person might imagine going round the supermarket locating items on the shelves, in the fridges and freezers, in anticipation of actually making the shopping trip. How do imaging, remembering, and actually doing differ from one and other?

One of the problems of consciousness is a reflexive one and that is to know what state (of consciousness) one is currently in, in order to act on it appropriately and adaptively. In the case of memory the feelings of remembering, familiarity, and knowing allow us to experience our conscious states 'as' particular states that can be acted on in ways that are adaptive for the state. However, these memory feelings can also be triggered by conscious representation of the future too. Indeed, imagining the future may often feature access of knowledge and memories in long-term memory and when this happens corresponding memory feelings will be triggered. Our model of the future, particularly the near future, is based on the past, especially the near past. This probably has adaptive value because much of the time (but certainly not always) the near future is closely related to the near past. Indeed, it has recently been found that the brain regions that become active during remembering and during imaging plausible false events are highly similar. Thus, it may be more correct to talk of a 'remembering-imaging system' than of a memory system alone. Within the remembering-imaging system, conscious states relating to the near past, present, and near future are colored by memory feelings that fluently arise in response to the contents of consciousness and allow us to feel continuity and familiarity in time – in short, what James so memorably called the 'stream of consciousness.'

However, our memories for the immediate past, what happened earlier today, yesterday, and may be 1 or 2 days further in the past are highly detailed and often very specific and frequently trigger the feeling of remembering. It seems likely that our prospective memory for the future, our imagings of future events, and our goals and plans may also have a similar specificity for the near future, perhaps just a few days, and gradually become more generic and abstract as the future becomes more distant. Memory for the past and the future and conscious feelings that are triggered may operate in a remembered-imagined window of time. As this window moves, linearly, through time, the past and future components are constantly changing but we experience this as fluent and continuous because the patterns of memory feelings bind together moments of consciousness. When these patterns of conscious memory feelings are interrupted, conscious experience becomes discontinuous and fluency is lost. Under such circumstances making sense of the world is compromised and effortful.

Malfunctions of Memory Feelings

One of the most severe cases of amnesia described in the literature is that of the musician Clive Wearing. He suffered brain damage due to a viral infection and this resulted in two types of amnesia. Retrograde amnesia refers to the loss of memories that were formed before a brain injury. Often this is temporally graded in the sense that the amnesia reaches back in time to a point where memories can be recovered again. For instance, in many closed head injuries (essentially a bang on the head) the amnesic period may only be for a few hours, days, or possibly weeks. In more severe cases it may cover years and even decades often stretching back to the early adulthood and teenage years. Wearing's retrograde amnesia was extremely dense and cover all his prebrain damage life preserving only a few pieces of autobiographical knowledge, for example, he knew he was a musician, had a wife, had children, and a few more factual aspects of his life. This patient, very unfortunately, also suffered a second type of amnesia, anterograde amnesia (the more common type) and this was an inability, following his brain damage, to

encode new information into long-term memory. His short-term memory was, however, intact and so he had a window of consciousness that persisted for about a 2 min period, but which could not be updated. One of his disturbing behaviors was that he would write in a note book, many times a day, 'I have just woken from being dead,' filling many notebooks over the first period of his illness. Without the dynamic change remembering-imaging window of conscious driving memory feelings, he was wholly unable to operate on the world. Despite the personal tragedy of this case it is interesting to note that he was aware that consciousness was changing over time and must have had a feeling of change every few minutes.

Clive Wearing is an unusual and extreme case. Other less severe types of amnesia may have other consequences. Patients have been described who have dense anterograde amnesia and cannot encode new experience at all or only to a limited extent but who have less dense temporally graded retrograde amnesia. One patient, for example, a man in his 60s, could recall reasonably well his early 20s, adolescence and childhood. In fact, in his late teens he had been on naval duty during World War II. It is clear from his descriptions of his memories that he had recollective experience and the feeling of remembering when he brought these to mind. The unfortunate consequence was that he acted on his memory feelings and came to behave like a 19-year old and have beliefs about the world that explained it as though he was still a naval rating, that is, that television constantly showed science fiction programs. One of the functions of conscious memory feelings such as recollection is to provide a basis for comprehension of the world and for adaptive actions. What one remembers of the past and the feelings which that triggers, drive our engagement with the world and in this case, as well as several other similar ones, demonstrate how powerful memory feelings can be in this respect.

Understandings of the world that are erroneous but generated in good faith on the basis of mental representations are sometimes referred to as 'confabulations' or as one researcher insightfully put it 'honest lies.' What is most interesting about plausible confabulations of the past is that they have often been found to be accompanied by recollective experience. Patients have in consciousness images of past experience and knowledge of their lives but configured in ways that are incorrect. One patient for example recounted his memory of the death of his brother in a car crash when he was much younger – in fact the brother was visiting him in hospital the same day. Such confabulations are not infrequent in certain types of brain-injured patients in whom false memories are experienced as true memories.

It should be noted here too that such memories are not uncommon in the everyday life of healthy adults and, moreover, it has been found to be relatively simple to induce false recollections in laboratory settings. Inducing false memories of events dating to childhood; false memories that come to be recollectively experienced is well documented as is inducing false memory details in adults, details that again are recollectively experienced. One explanation of these experimentally induced false memories is that they arise from errors in 'source monitoring.' The individual, for whatever reason, fails to evaluate the source of the memory information in consciousness and attributes it to memory rather than some other source, for example, imagination. How this then gives rise to the conscious experience of remembering is not known. One possibility is that information in consciousness that is evaluated, by nonconscious processes, to be of memory content, perhaps for example it is found to be highly associated with other memory content, then triggers the experience of remembering. Whatever the case, it is clear that normal, intact, memory is also prone to at least the occasional honest lie.

Consider the reverse situation and the conscious state that might occur if a person had recollective experience but no specific autobiographical memory entered conscious awareness. This powerful experience of remembering but with no memory in mind is the state that has been termed déjà vu. But déjà vu means 'having seen before' and as such approximates more to familiarity and knowing than to recollective experience. Researchers have suggested that the term déjà vu vecu, which is the experience of 'having lived the present moment before,'

better captures the experience of recollection without a memory in conscious awareness. Some brain-damaged patients suffer from the relatively frequent experience of *déjà vu vecu* and this may be because the brain circuits that mediate recollective experience are no longer constrained to activate only when a specific autobiographical memory is in conscious awareness. One proposal is that there is a brain circuit which when activated mediates the conscious feeling of remembering. The circuit is activated when specific memory content enters consciousness. However, specific memory content is most probably being automatically activated all the time by cues encountered in the environment and generated in the brain. One view is that control process gate this information entering consciousness and suppress the memory feeling circuit. When control processes are damaged such that they can no longer suppress the memory feeling circuit, but can still gate what knowledge enters consciousness then it is possible to have frequent experiences of recollection when no memory content is in consciousness.

The consequences of this for patients who suffer this malfunction of consciousness are severe and far-reaching. For these patients the present moment is consciously felt as a memory. Because of this they act on their *déjà vu vecu* feelings because for them they are simply remembering. Thus, they stop reading newspapers (which only report old news), they cease watching television because it all repeats, they stop answering letters, sending letters and cards, fail to keep appointments and by their disturbing behavior induce anxiety, anger, and incomprehension in their carers (usually a life partner). A conscious feeling of remembering can then powerfully drive a wide range of behaviors. More generally, the experience of recollection without a specific memory in mind may occur in anyone if an activated memory triggers recollection but the memory itself does not enter conscious awareness. The fact that this is a relatively rare, but by no means unknown, malfunction of memory and consciousness suggests that the function of recollective experience, to allow us to know in an experiential way that we are remembering, is a particularly important function of this type of memory consciousness.

Conclusions

The conscious experience of memory lets us feel what states we are in, allows us to identify distinct states, and supports the flow of experience through time in a window of remembering and imagining, and when it malfunctions disorientation, confusion, and confabulation arise. Furthermore, and most importantly, conscious feelings of memory drive behavior and when they are triggered inappropriately they can give rise to dysfunctional patterns of behavior. A further possible function of the conscious experience of remembering might also be to stimulate the adoption of plans and goals. Recollectively experiencing events which featured goal processing, knowing the outcomes of previous actions, maintaining the sense of familiarity, might drive the creation of new plans extend into the remembered-imagined future. It is particularly interesting that one deficit in amnesia appears to be the loss of the ability to imagine the future and the loss of goals extending into that future. Without a dynamic constantly changing and renewing window of remembered-imagined consciousness it seems that a future may not be possible.

See also: Memory: Errors, Constructive Processes, and Conscious Retrieval; Self: Personal Identity; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Addis DR, Wong AT, and Schacter DL (2007) Remembering the past and imagining the future: Common and distinct neural substrates during event construction and elaboration. Neuropsychologia 45: 1363–1377.

Conway MA (2005) Memory and the self. Journal of Memory and Language 53(4): 594–628.

Gardiner JM (2008) Remembering and knowing. In: Roediger HL, III (ed.) Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, Byrne J (ed.), 4 vols, pp. 169–198. Oxford: Elsevier.

James W (1950) Principles of Psychology. New York: Dover.

Johnson ML, Hashtroudi S, and Lindsay DS (1993) Source monitoring. Psychological Bulletin 114: 3–28.

Moulin JAC, Conway MA, Thompson R, James N, and Jones RW (2004) Disordered memory awareness: Recollective confabulation in two cases of persistent déjà vecu. Neuropsychologia 43: 1362–1378.

Roediger HL and McDermott KB (1995) Creating false memories: Remembering words not presented in lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 21: 803–814.

Sacks O (1985) The Man Who Mistook His Wife for a Hat. London: Duckworth.

Schacter DL and Addis DR (2007) The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. Philosophical Transactions of the Royal Society of London Series B 362: 773–786.

Tulving E (1983) Elements of Episodic Memory. New York: Oxford University Press.

Tulving E (1985) Memory and consciousness. Canadian Psychology 26: 1–12.

Tulving E (2002a) Episodic memory: From mind to brain. Annual Review of Psychology 53: 1–25.

Tulving E (2002b) Chronesthesia: Awareness of subjective time. In: Stuss DT and Knight RC (eds.) Principles of Frontal Lobe Functions, pp. 311–325. New York: Oxford University Press.

Tulving E and Pearlstone Z (1966) Availability versus accessibility of information in memory for words. Journal of Verbal Learning and Verbal Behaviour 5: 381–391.

Wheeler MA, Stuss DT, and Tulving E (1997) Towards a theory of episodic memory: The frontal lobes and autonoetic consciousness. Psychological Bulletin 121: 351–354.

Automaticity and Consciousness

W Schneider, University of Pittsburgh, Pittsburgh, PA, USA

Glossary

Automatic attention response – A specific type of automatic process that does priority assessment of stimuli based on the priority for future processing and places information in working memory with a priority signal can interrupt ongoing controlled processing and facilitates the processing of the high-priority stimulus.

Automatic process – Cognitive process that develops for consistently executed component behaviors that are performed rapidly, with minimal effort or attention involving a long-term associative connections and require an appreciable amount of consistent training to develop fully. The processing is domain specific, occurring in multiple perceptual/response area of the brain.

Cognitive control network – A network of brain areas that provide endogenous control of attentional operations performing controlled processing operations including domain general cortical areas. The brain areas include anterior cingulate cortex/presupplementary motor area (ACC/pSMA), dorsolateral prefrontal cortex (DLPFC), inferior frontal junction (IFJ), anterior insular cortex (AIC), dorsal premotor cortex (dPMC), and posterior parietal cortex (PPC).

Controlled process – Cognitive process that involves attention and top-down control that performs novel or varied tasks that is characterized as being slow, serial, effortful, easy to set up and alter, laying down a strong memory trace, and nonrobust (e.g., sensitive to stress, alcohol, sleep deprivation). It occurs in the domain general cognitive control network of the brain.

Spatial neglect – A neurological condition after damage to the parietal cortex that produces a deficit in attention to and awareness of one side of space.

Visual search – A task in which an observer is given a memory set of items to search for, for example, letters, words, and pictures and search of display of potentially multiple items, and responding if there is a match (e.g., push a button if you see the letters ‘c’ or ‘b’ in the string of ‘detbn’).

Introduction

The concepts of automaticity and consciousness have been discussed together extensively for centuries and this continues in the current literature. The concept of automaticity was a major focus in Williams James’s *Principles of Psychology* (1890), which contrasted habit and ideational/will processing. In modern times, automatic processing has been an important issue in attention, skill acquisition, social perception, and cognitive neuroscience.

Most human behaviors involve automatic processing of consistent components of the task. For example, the word encoding of this text is an automatic process. When you look at a familiar word, the meaning of the word is activated. This takes little effort or awareness of the component processes of perception. In contrast, if we alter the words such that the automatic processes fail, reading becomes slow and effortful and you become aware of the processing. Interpreting the text string where the letter order is reversed (e.g., what is ‘‘ssensuoicsnoc’’) is hard and error prone, and disrupts the normal reading process severely. Note, however, that you were likely more aware of the rearranged version of ‘consciousness’ read in reverse order than you were the other times that word occurs on this page. The type of processing done on the reversed text is typically referred to as controlled processing,

which is characteristic of novel or early trials of a task. The distinction between automatic and controlled processing relates to dual process theories of human cognition. These two processing modes differ in the types of processing they perform and the brain mechanisms involved.

Multiple theoretical positions align controlled and automatic processing with the concepts of conscious and nonconscious processing. Although there is a high overlap of concepts where most control processes are conscious and automatic processes are not, there are well-known exceptions that indicate consciousness has a more nuanced relationship. Consciousness is related to a subset of control process events and some automatic processes can place information into consciousness.

Automaticity and Dual Processing Theory

The concept of automaticity relates to dual process theories that assume that most human behavior results from the interplay of two types of processing referred to as automatic and controlled processing. The theoretical and empirical work of Walter Schneider and Richard Shiffrin provided modern day definitions and quantitative models of these cognitive processing modes. 'Automaticity' or 'automatic processing,' refers to the cognitive process that develops for consistently executed component behaviors that are performed rapidly with minimal effort or attention, involve long-term associative connections, and require an appreciable amount of consistent training to develop fully. Automatic processing is typically characterized as being fast, parallel, requiring low effort, minimal memory demand, and is a robust type of processing. Practiced consistent behaviors involve automatic processing in tasks such as driving, brushing teeth, word encoding, and social perception.

In contrast to automatic processes, 'controlled processing' occurs typically in novel or varied situations (e.g., searching for a random subset of letters in a display). Controlled processing is characterized as being slow, serial, effortful, easy to set up and alter, laying down a strong memory trace, and nonrobust (e.g., sensitive to stress, alcohol, sleep deprivation). A novel task is initially executed by controlled processing. Controlled processing involves attentive operations that can shift attention, compare memory states, execute sequential rules, induce learning, and initiate responses.

Empirical results show automatic and controlled processing differ in six characteristics: (1) extended consistent training is required for automatic processing while controlled processing requires only a few trials to develop; (2) automatic processing is fast and parallel, while control processing is slow and serial; (3) automatic processing requires low effort whereas controlled processing is high effort; (4) automatic processing is robust to stressors while controlled processing is not, (5) automatic processing is hard to alter whereas controlled processing is easy; and (6) learning is dependent on the amount and nature of controlled processing where as little is learned in pure automatic processing.

These differences between controlled and automatic processing are quite marked and often interpreted as indicating qualitatively different forms of processing. For example, with training humans can perform surprisingly fast parallel automatic tasks (e.g., looking for words, icons, numbers colors at the same time and processing 96 stimuli per second at high accuracy). In an example of semantic category search (e.g., find words that are members of the categories, press a button if the word is from the categories 'vehicles, land formations, units of time, and fruits' then be presented words such as 'hammer, foot, bank, peach'). If this task is novel or the stimuli have a varied mapping (i.e., which stimuli are targets or distracters varies over trials) controlled processing is active. Processing requires about 200 ms for each comparison, so with 16 comparisons the reaction time would be 3.2 s. The mental effort would be slow and serial as each category would be checked against each word.

In contrast, if you searched for the same four categories consistently for many trials, automatic processing would develop. Automatic processing would be fast (1 ms/comparison, 100) and parallel in memory for small displays (two words). An important trait of automatic search is the search appears to be a parallel search for example, searching for a match of many possible previously detected stimuli can be as fast as searching for any specific stimulus.

The terminology of automatic/controlled processing emphasizes the difference in control of the processing. Controlled processes are easy to set up and change (e.g., respond to red stimuli on one trial, and on the next, respond to green stimuli). However, if you perform a given operation many times, it becomes easier to perform and may become automatic where it is difficult to inhibit the processing of the task. For example, consider the Stroop color naming task, where participants name the ink color of words which are incompatible with the color meaning of the text (e.g., the text 'blue' printed in red ink and the response is to answer 'red'). This slows reaction time dramatically (i.e., 70% slower response time), relative to naming the colors of squares without the incompatible words. This effect is a result of developing automatic word processing, which evokes the color response even when it is not intended. Altering this automatic process requires effort and takes a week of training to overcome.

Action slips also show and lack of control in automatic behaviors. Action slips are common errors such as getting out of an elevator on the wrong floor because the doors opened up a floor early or continuing on your standard driving route to work when you intended to go somewhere else illustrate that an automatic process can determine behavior with minimal consciousness. Addictions such as cigarette smoking and pathological gambling illustrate the difficulty of altering well practiced consistent behaviors even if they are very detrimental.

These two forms of processing provide complementary benefits. Tasks can be learned quickly via controlled processing but are slow, serial, and effortful, and nonrobust. After much consistent practice in the task, it can be performed much faster with lower effort by automatic processing. In martial arts contests, the master's speed is far beyond that of the novice and in a fight the master would have great advantage relative to the novice. However, the novice's ability to hear instructions from the master allows the student to execute rudimentary forms of the moves in few trials. The student is consciously aware of the specific steps of the moves and often rehearses them before execution. With instruction, students learn good moves, much more rapidly than without instruction, and conscious control of their actions. The presence of control processing provides early skill and produces the memory representations that with practice allow fast execution of effective moves.

The two forms of processing dynamically interact resulting in cognitive processing that is a result of exogenous and endogenous processing. Stimuli with certain features (e.g., loud noises, moving/looming visual objects, pain) and learned responses (your name, cry of your child), produce an 'automatic attention response' that activates the related representation and focuses attention on the stimulus. Note this can occur without any intention (e.g., missing a stair). The classic 'cocktail party effect' is where you hear your name in a presumably unattended conversation, which causes you to shift your attention to that conversation, and you become consciously aware people are talking about you.

Control processing can also alter automatic processing. In a fraction of a second you can alter how you walk. This is remarkable in that you can alter a skill you have perhaps practiced for decades and dramatically alter your motor movements (e.g., try to walk but do not let the outsides of your feet touch the floor). Your walking will likely be much slower, demand much more attention, and you will be far more consciously aware of the act of walking.

Consciousness in Dual Processing Theory

The concepts of controlled and automatic processing are closely related to concepts to consciousness and nonconsciousness processing. This is a simplified view that captures the high overlap of the situations in which both occur but this view does not detail the different nuanced conditions in which this simplification breaks down. In order to identify the differences one must be clear about the definitions of consciousness and automaticity that one is considering.

There are also multiple definitions of consciousness. The distinction between access consciousness and phenomenal consciousness by Ned Block provides a useful distinction to organize the discussion of two types of consciousness. Briefly, 'access consciousness' relates information made available to a

global workspace, what Block refers to as ''the brain's 'consumer' systems: systems of memory, perceptual categorization, reasoning, planning, evaluation of alternatives, decision-making, voluntary direction of attention, and more generally, rational control of action.'' Phenomenal consciousness refers to the experience of the event and the ability to introspect, reflect and report on the nature of the event. This metacognitive report may be a different process that provides an experience based on processing events occurring in the global workspace.

Table 1 provides a matrix of dual processing concepts and the relationship to consciousness. The upper left cell, automatic nonconscious behaviors, represents the bulk of human cognition. You are unaware (either phenomenological or via access) of the enormous knowledge stored in the connection weights between brain regions. You can become aware of the result of using those connections but not of the connections themselves. For example: What is the name of the first person to walk on the moon? How do you tie your shoes? You likely had neither access nor phenomenological awareness to those two facts before you read the questions. You can drive and make many corrections for changes in the road or other drivers that influence your actions. Yet you may have very limited awareness of much of what is perceptually processed.

Dan Simons' study of 'change blindness' illustrates that you are aware generally what is attended to and the gist of the environment. Simons examines what is detected in movies between camera cuts. In a conversation of two people much can be changed (e.g., object on a table clothing of the actors, the actors themselves) and occur without awareness. We are aware of the gist of the situation and what we are attentively focused on (e.g., changing an actor from a male to female in a conversation of two people is detected as it changes the expected nature of interaction).

The lack of consciousness in nearly full automatic processing is perhaps best illustrated in very high workload, high stress, zombie behaviors, and epileptic seizures and alcohol intoxication. High workload (e.g., driving while engaged in a heated argument) can result in driving with little conscious awareness of the drive (e.g., what stop lights did you stop at). In high stress situations such as combat, there can be a nearly complete disassociation between extensive automatic behaviors (e.g., hand combat motor actions) and the conscious mind activity (e.g., will my mother forgive me if I die in the next minute?). Zombie behaviors are highly trained, automatic, fluid visuomotor behaviors in which people are on automatic pilot (mountain biking, playing sports, driving). Sleepwalking can involve complex behaviors (e.g., rearranging furniture) with no sign of awareness or memory. Automatic capture errors (e.g., exiting an elevator on the wrong floor) illustrate autopilot performance. Robert Penfield describes an epileptic walker: ''Thus, the automaton can walk through traffic as though he were aware of all that he hears and sees, and so continue on his way home. But he is

Table 1 Relations of consciousness and dual processing modes

	Nonconscious	Access conscious	Phenomenal consciousness
Automatic processing	Connection strength	Pattern activations in working memory	Automatic attention responses
	Most pattern activation		Gist
	Zombie behaviors		
Controlled processing	Sub 100 ms component processes	Pattern activation in working memory	Slow processes involving changes in controlled processing events
	Some extended execution of controlled processing pre results	Selection of information for comparison or remembering	Completed component processes
			Rehearsed working memory items

aware of nothing and so makes no memory record. If a policeman were to accost him he might consider the poor fellow to be walking in his sleep.''

These all illustrate situations where past consistent practice has created associations that can execute complex behaviors with little controlled processing or awareness. The behaviors are complex and can be executed for extended durations without the need for controlled processing.

Automatic behaviors can place information into working memory. Most of it is not phenomenologically available but can be if attention is shifted to it (e.g., the pressure of the chair you are sitting on as it presses against the leg).

One class of automatic processing, an automatic attention response, does place information into both access and phenomenal consciousness. This is best illustrated in visual search tasks. For example, if a subject is consistently detecting a specific set of letters, for example, ('X' or 'S'), after hours of practice, one can detect letters at very high rates (e.g., process displays of 4 letters occurring every 80 ms, or 50 events per second with high accuracy >90%). At very high rates nontrained observers are unable to see character shapes at all. Yet, to the trained observer the letters pop out of the blur perceptually seeming to stop the display and become consciously available. Hearing your name in an 'unattended conversation,' missing a step and attending to your feet, or trying to retrieve a reference in memory and it 'pops' to mind, are examples of an automatic process activating a stimulus, directing attention to the stimulus, and generating awareness.

The phenomenon of spatial neglect illustrates how losing the automatic attention responses can produce a loss of access to information. Spatial neglect is a condition after damage to parietal cortex that produces a deficit in attention to and awareness of one side of space. Patients with neglect do not see stimuli on one half of the visual field or perform actions on half of their body (e.g., not dressing the right side of their body). This is not a deficit in vision because they can see the stimuli (e.g., can read a single word crossing both visual fields but when asked to name random letters will neglect to read the ones on the right side). This lack of stimuli interrupting controlled processing is very debilitating for operating in the normal world making it difficult to dress oneself, perform skills (driving) and interact socially. Controlled processing is tightly associated with conscious processing but there are some control processes that are not conscious. Controlled processing is slow, serial, and involves attention. Given that consciousness is assumed to involve longer processing of hundreds of milliseconds, it is reasonable to expect that the results of controlled processing are both consciously accessible and phenomenologically available. In general, individuals are aware of what they have attended to, targets they have detected, and working memory items that they have rehearsed. In controlled search tasks (ones with novel or varied targets), whenever a target is detected, the subject can report the target. False alarms are rare indicating the detection resulted in phenomenological awareness.

There are control processes that seem unavailable to consciousness. A dramatic example of this occurs in a search task when the subject finishes the search, makes a response indicating no target was detected, feels they have quit the task, but then sometimes after a substantial delay (several hundred milliseconds) they have an 'oops' response where they realize they missed the target. This suggests that the control process was continuing to check information but there was no conscious phenomenal awareness.

Subjects have difficulty reporting details of fast control processes. For example, in search task comparing small numbers (4 or less comparisons, 200 ms comparison time) subjects typically execute an exhaustive search continuing to compare items even after a match is found. When this is discussed in classes, students often disbelieve that they could be using such a strategy. As another example, subjects may perform very predictable search patterns (e.g., in driving around a curve, subjects look at the tangent point of the curve), but few are aware of the strategy they deploy. These are examples of control processes that result in minimal awareness.

Function of Consciousness in Dual Processing Theory

Automatic, controlled, and conscious processing operate together with different functions

performed by each. The functions of automatic processing are to perform routine, consistent cognitive actions and to alert controlled processing/consciousness of important stimuli and the gist of the current situation. The functions of control processing are to control attention, perform control process operations such as selective attention, comparison, response release, performing interpretive execution of control programs, set up patterns to allow learning, and to alter/modify the result of automatic processing. Each processing mode has strengths and weaknesses but together they provide greater functionality and survivability to the individual than each separately.

There are many disparate views of the functions of consciousness. These views vary in terms of the types of consciousness involved and the breadth to which they feel the term of consciousness applies.

Lee Pierson and Monroe Trout view the role of consciousness to explicitly counteract serious problems with purely automatic processing. They state ''Consciousness makes volitional attention possible'' and that ''Although the ultimate function of volitional attention is to make volitional movement possible, its proximate function is to override automatic attention that is not well-suited to the situation at hand. The capacity for volitional attention gives the conscious organism the flexibility to nondeterministically yet nonrandomly sustain attention on a particular conscious content longer than the default set by neural processes.'' They continue commenting that consciousness can correct ''misapplied automatisms, infinite loops, and tendencies that were adaptive when they evolved but have outlived their usefulness.''

In a changing world there is a need to alter some automatic behaviors. A common example might be when a software 'upgrade' changes a function of key strokes. The motor actions that were usefully automatic become painfully difficult to inhibit and correct. In search tasks, it can take more time to alter automatic behaviors than it did to learn them. In the case of addictive habits (e.g., smoking) this can make people have enormous difficulty kicking the habit even if they know it is killing them. In a visual search task if you learned to search for numbers and ignore letters, detection of numbers and the search for them becomes automatic and quick, requiring low effort. However, if you switch to detect letters and ignore numbers, the numbers pop into consciousness and it is difficult to process the letters, making you worse that a novice with no practice at all. Subjects report that after reversal, the task that was easy and perhaps boring is now aversive. They feel their attention is dragged to the old targets when they are trying to attend to the new ones. They are very phenomenally aware that automatic processing is not working to their advantage. The conscious effort to abort processing the old targets and maintain attention on the new letter targets occurs slowly, and after thousands of trials new automatic processing is established. Consciousness awareness of the processing and controlled effort to inhibit the old automatic processing allows for alteration of automatic processes and the eventual replacement of automatic processes with new ones.

Many consciousness researchers take a wider view on the role of consciousness. Bernard Baars lists a variety of functions including executive control, decision making, recruiting and control actions, adaptation and learning; prioritizing the cognitive system's concerns, facilitating problem solving, optimizing the trade-off between organization and flexibility, detecting errors and editing action plans, creating access to the self, facilitating learning and adaptation, and in general increasing access between otherwise separate sources of information.

In this broader view of consciousness there is a great deal of overlap between the concepts of control processing and the nature of access consciousness. It is unclear which of these differences are just terminology referring to the identical underlying functions. The access consciousness working memory and controlled processing working memory are perhaps very similar constructs.

The functions of phenomenological consciousness provide tasks that suggest more reflective processing systems than is typically associated with controlled processing. This is perhaps a processor that has access to a subset of control operations and working memory (see above controlled processing without phenomenological awareness) but can reflect on those operations and their consequences.

Consciousness could have multiple functions operating as a metaprocess to controlled processing. It might not directly perform control

operations but rather monitor the success of those operations and initiate tweaks in control processing. For example, in controlled searches, if observers are missing too many targets, they often slow down the comparison rate. If there is a demand for more speed they might increase the rate. If there are multiple concurrent tasks they may shift more resources to the task with higher payoff. In these situations observers are phenomenally aware of their performance and take volitional actions to alter processing.

The metaprocess would be beneficial for accomplishing the broad list of Baars' expanded functions of consciousness. These functions generally relate to monitoring control process operations, assessing whether current operations are accomplishing the desired goals at the intended levels of performance, determining if the existing control operations are successful and using control state variables to influence controlled and automatic behaviors.

To illustrate the metaprocess, consider control processing related to reading instructions to assemble a bookshelf from the instructions. This is not automatic, each step of the instructions must be read, converted to a control process routine, and executed. If the instructions are clear, easy to interpret, accurate and read, there is little need for controlled metaprocessing. The control instructions typically involve search for this part A, move position X on part B, and perform motor action Y on part B until some state Z is reached. Then go on to the next step. However, if a control subroutine fails (e.g., you can not find a part A) or the motor action fails (e.g., the screw does not go in), a metaprocess kicks in to find alternatives to execute the intended step. This could take many forms such as tweaks (look longer for A), rereading/re-encoding the step, deciding if the instructions are wrong and interpreting what they 'meant to write,' recalling the last time you performed a similar task, deciding whether to delegate the task to someone else, or deciding to give up. These metacontrol operations have great utility in a world where control processes must be learned or altered across situations. They also have the characteristic of not being part of the control routine that would have been executed if the step had succeeded.

Information Access and Transfer between Automatic, Controlled, and Conscious Processing

Given the large amount of parallel automatic processing and the small amount of serial controlled and conscious processing it is important to recognize the limited information flow between the three systems. The relationships in Table 1 indicate consciousness has very limited access to the full processing of the brain represented by automatic processing. Schneider and Chein propose that automatic processing occurs in a network of ten thousand hypercolumn populations and at any point controlled processing can monitor a only small fraction of these (<1%). Consciousness might have access to some of the working memory information that controlled processing has access to, as well as, information about the executing control processes.

Information flow between the systems of automatic processing to controlled processing/consciousness and back again is likely limited to avoid overloading serial systems. Schneider and Mark Pimm-Smith suggest that controlled processing access is limited to an 'inner loop' between the sensory-motor areas processing highly processed representations (e.g., not features but meaningful objects or motor action routine). This limited monitoring of the actions of a modest number of regions (less than ten rather than hundreds). The prioritization of the automatic attention responses allows the control system to order modules for processing. The serial all or none phenomenological nature of control and conscious processing, attenuate information input to reduce the risk of overload and message interference.

Conscious monitoring has the potential for very specific conscious control of action. Automatic processing can process the consistent world, controlled processing can sequence attention to accomplish goals. However, if these operations are based on only activating modules, coding actions would be very coarse (e.g., get food). If consciousness contains specific messages it could set the control programs to find specific codes (e.g., 'chocolate covered almonds').

Top-down transmission of codes from consciousness to controlled processing enable consciously

guided control operations. For example transmitting an object message and action (e.g., 'find chocolate covered almonds') could launch very specific control process search behavior. If after executing search behaviors the target is found, consciousness could initiate a follow-on complex behavior ('give found object to youngest daughter'). The ability of the metacontrol process to inject specific codes into control process operations provides great specificity. It is hard to draw definitive boundaries between a metacognitive conscious processing system and the controlled systems that executes the actions. These functions may or may not be in different brain tissue.

Top-down conscious processing can also alter automatic processing for both good and bad. An expert tennis player may use consciousness to assess the opponent's strategy and skills and then activate high level state strategy state variables (e.g., fatigue opponent, move opponent back, conserve my strength). These alter high state variables that provide context information for the automatic processing that modulate the execution of hundreds of muscle movements.

There are downsides to conscious control of automatic behaviors. Attending can alter the weighting of information, disrupting the automatic parallel flow of information. Asking an expert to attend to a component is a known trick of experts (e.g., ask your opponent in tennis 'when you serve do you throw the ball before you move your racket?'). Sian Beilock showed that expert golfers improved their putting if they performed a dual task that blocked conscious/control top-down processing. In contrast, novices, without automatic routines, showed poorer performance when control processing was disrupted.

Brain Systems for Automatic and Controlled Processing Relating to Consciousness

Michael Cole characterized a brain network dedicated to controlled processing using functional magnetic resonance imaging (fMRI) methods. He has shown that controlled processing involves a network of domain-general brain areas (anterior cingulate cortex/presupplementary motor area (ACC/pSMA), dorsolateral prefrontal cortex (DLPFC), inferior frontal junction (IFJ), anterior insular cortex (AIC), dorsal premotor cortex (dPMC), and posterior parietal cortex (PPC)). These areas are domain general in that they are active whether processing visual or auditory stimuli, simple features or complex tasks (oriented lines vs. radar control signals). Figure 1 (left) shows the brain areas active during a face search task with varied targets.

In contrast, automatic processing occurs in the sensory motor representation areas of the brain. As automaticity develops, the control system areas decrease in activity, often dropping out as the task become fully automatic. Figure 1 (right) shows the network cortical areas active in an auditory search condition after extended consistent practice. There are two important things to note.

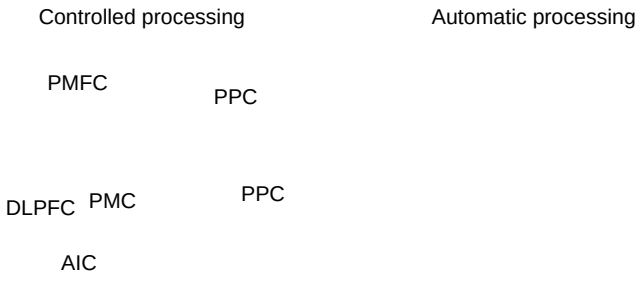


Figure 1 (Left) Controlled processing areas activated during activation of cognitive control net regions during visual face and auditory search. Activated areas: ACC/pSMA, DLPFC, IFJ, AIC, dPMC, and PPC. (Right) Automatic processing after extended search for auditory targets. Here the control network has dropped out and only the sensory areas remain active processing the stimulus via automatic processing.

First, the control areas involved in novel search are at this point inactive. Second, the sensory areas involved in processing the stimuli are active even in the absence of control processing activity. This is consistent with the proposal that after automaticity develops processing can occur without the assist of control processing.

There is a small but growing imaging literature regarding conscious processing. The problems of definition between control and conscious processing and the nature of what is the resting control (e.g., if one thinks to oneself during resting during a resting control, you can not use that condition as a subtraction to find areas of consciousness). In a metaanalysis of five imaging studies of consciousness using techniques such as bistable figures (e.g., Necker cube), Geraint Rees found areas of the prefrontal cortex and posterior frontal cortex were active. These are the same areas active in controlled processing search tasks. As we get better tasks to discriminate controlled processing from consciousness, studies will determine if these two processes have differential cortical areas.

Conclusion

The concepts of automaticity, controlled processing, and consciousness are tightly coupled concepts, relating to likely different processes that has important value to the cognitive function of humans. These processes interact making it difficult to identify the boundaries between them, the physiological structures, and the dynamics of interactions. The developing behavioral, computational, imaging, and neurology literature support the view that these are coupled processes. Automatic processing occurs in many representation-specific brain areas and supports parallel, low effort processing that is difficult to control. Controlled processing operates from a set of domain general areas that exhibits serial effortful processing in novel or changing situations. Automatic processes in the form of automatic attention responses can activate conscious awareness to stimuli that many have been unattended. Most controlled processing events such as target detection are consciously aware but there are some fast component control processes that are not consciously aware. Consciousness may be a metacontrol process monitoring key control process events and relating those events to goals to allow flexibility of behavior. At times, consciousness may tune controlled and automatic processes and at times override them. The interplay of the three types of processing produce a cognitive system that can perform fast automatic behaviors, rapidly learn and process new behaviors, and alter both the control and automatic processes to deal with a changing world.

Suggested Readings

Baars BJ (1997) In the theater of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. Journal of Consciousness Studies 4(4): 292–309.

Beilock SL, Carr TH, MacMahon C, and Starkes JL (2002) When paying attention becomes counterproductive: Impact of divided versus skill-focused attention on novice and experienced performance of sensorimotor skills. Journal of Experimental Psychology: Applied 8(1): 6–16.

Cole MW and Schneider W (2007) The cognitive control network: Integrated cortical regions with dissociable functions. NeuroImage 37: 343–360.

Goodale MA and Milner DA (2004) Sight Unseen: An Exploration of Conscious and Unconscious Vision. Oxford: Oxford University Press.

James W (1890) The Principles of Psychology, vol. 1. New York: Holt.

Koch C and Crick F (2001) The zombie within. Nature 411: 8.

Koch C and Naotsugu T (2006) Attention and consciousness: Two distinct brain processes. Trends in Cognitive Sciences 11(1): 16–22.

Mesulam MM (2000) In: Attentional Networks, Confusional States, and Neglect Syndromes. Principles of Behavioral and Cognitive Neurology, 2nd edn., pp. 174–256. London: Oxford University Press.

Pashler HE (1999) The Psychology of Attention. Boston, MA: MIT Press.

Pierson L and Trout M (2005) What is consciousness For? Cogprints. http://cogprints.org/4482/1/whatisconsciousnessfor.pdf.

Schneider W and Chein JM (2003) Controlled & automatic processing: From mechanisms to biology. Cognitive Science 27: 525–559.

Schneider W and Pimm-Smith M (1997) Consciousness as a message aware control mechanism to modulate cognitive processing. In: Cohen J and Schooler J (eds.) Scientific Approaches to Consciousness: 25th Carnegie Symposium on Cognition, pp. 65–80. Mahwah, NJ: Lawrence Erlbaum.

Schneider W and Shiffrin RM (1977) Controlled and automatic human information processing. I. Detection, search, and attention. Psychological Review 84(1): 1–66.
Shiffrin RM (1988) Attention. In: Atkinson RC, Hernstein RJ, Lindzey G, and Luce RD (eds.) Steven's Handbook of Experimental Psychology: Learning and Cognition, vol. 2, pp. 739–811. New York, NY: Wiley.
Simons DJ (2000) Current approaches to change blindness. Visual Cognition 7: 1–15.

Biographical Sketch

Walter Schneider is a cognitive neuroscience researcher professor of psychology, a senior scientist at Learning Research and Development Center, a faculty of Center for the Neural Basis of Cognition, at the University of Pittsburgh, USA. He received his undergraduate degree at the University of Illinois in 1971 and his PhD in mathematical psychology in Indiana in 1975. His dissertation came out as a pair of articles in *Psych Review* that quickly became a science Citation Classic with over 20 000 citations. He did a postdoctoral in neurophysiology at Berkeley. He established himself in the fields of attention and skill acquisition at the University of Illinois till 1985 and to the University of Pittsburgh Learning Research and Development Center since then. In Pittsburgh with Jay McClelland he set up the Neural Processes of Cognition Program that evolved into the Center for the Neural Basis of Cognition training students in modeling behavioral science and neuroscience. In 1991, he was one of the first psychologists to use fMRI and become a leader in fMRI brain imaging work in attention and the study of controlled processing and skill acquisition. He is a fellow of American Association for the Advancement of Science, 1995; a fellow of the American Psychological Society, 1997. He holds several patents and awards for computer software programs including the and coauthor of E-Prime. His 2003 *Cognitive Science* paper became a top download paper describing the CAP2 model relating brain systems and cognitive function.

Bistable Perception and Consciousness

P Sterzer, Charité Campus Mitte, Berlin, Germany
G Rees, University College London, London, UK

Glossary

Ambiguous figures – Images that can be interpreted as representing two objects or scenes, or two versions of an object.
Binocular rivalry – A phenomenon of visual perception that is characterized by perceptual alternations between different images presented to each eye.
Bistability – The ability of a system to exist in either of two steady states.
Extrastriate cortex – A belt of visually responsive areas of cortex surrounding the primary visual cortex.
Functional magnetic resonance imaging (fMRI) – A brain imaging technique that measures the hemodynamic responses related to neural activity in the brain.
Gamma distribution – A probability density function that plays an important role in statistics; the exponential distribution and chi-square distribution are special cases of the gamma distribution.
Neural correlate – A neurobiological state whose presence regularly correlates with a specific content of experience.
Parietal cortex – Superior posterior part of the human brain, the most important function of which is the integration of sensory information from different modalities.
Prefrontal cortex – The most anterior part of the frontal lobes of the brain. It has been implicated in planning complex cognitive behaviors, personality expression, and regulation of social behavior.
Primary visual cortex – The first cortical area to receive inputs from the eye via the geniculostriate pathway; also referred to as V1, area 17, and striate cortex.

Introduction

Bistable perception is a perceptual phenomenon characterized by changes in subjective perception while sensory input remains constant. Usually, perception alternates spontaneously between two mutually exclusive interpretations of the same sensory input. Rarely, more than two interpretations are possible, in which case the term multistable perception is used.

While bistability is also known for auditory stimuli, it is best known and has been most extensively studied in the visual domain. Bistable perception occurs when incoming sensory information is ambiguous or conflicting and when no additional cues are available that allow the visual system to converge on one unique interpretation. Typically, one interpretation remains stable for a few seconds until perception switches spontaneously and unpredictably to the alternative interpretation and then switches back again after another few seconds, and so forth. Famous examples include the Necker cube, Rubin's face-vase illusion, bistable apparent motion, and binocular rivalry (Figures 1(a)–1(c)).

Human fascination with such phenomena can be traced back to antiquity, but interest in bistable perception has recently been boosted by the advent of modern brain imaging techniques such as functional magnetic resonance imaging (fMRI). Such techniques can be used to measure brain activity in humans noninvasively, and therefore offer the opportunity to study the neural correlates of conscious visual perception. Importantly, to identify neural activity related to conscious awareness of a visual stimulus, one needs to show that this activity is not simply the consequence of physical stimulation, as such activity could occur irrespective of whether the stimulus is consciously perceived or not. The use of stimuli evoking bistable perception, where conscious perception changes while sensory

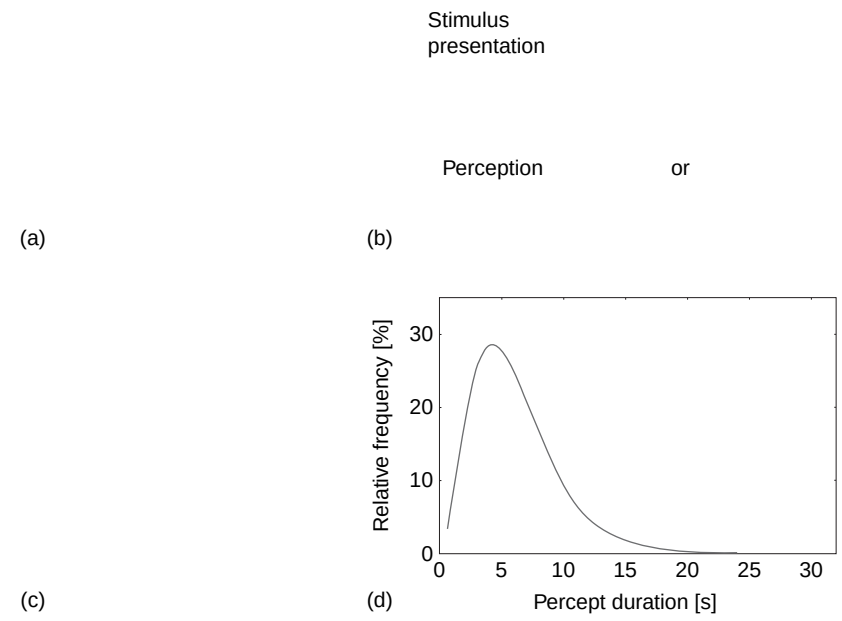


Figure 1 (a) Rubin's face/vase illusion, a figure-ground reversing stimulus that can be seen to alternate between a central white shape on a black background and two black profiles of a face against a white background. (b) The apparent motion quartet: Two dots are flashed in diagonally opposite corners of an implicit square in rapid alternation with two dots appearing in the other two corners, typically yielding bistable perception of two dots moving either horizontally or vertically. (c) Binocular rivalry: different images are presented to the two eyes, resulting in spontaneous perceptual alternations between each monocular view. (d) The distribution of perceptual phase durations during bistable perception typically follows a gamma distribution.

stimulation remains constant, is an elegant experimental approach to distinguish neural activity related to conscious perception from that related to physical stimulus properties. Over the past decade, a number of studies have chosen this approach to investigate various aspects of conscious visual perception. This article summarizes the evidence that has been derived from the neuroscientific study of the neural correlates of consciousness using the intriguing phenomenon of bistable perception.

Theoretical Considerations

Types of Consciousness and the Question of Causality

When we talk about neural correlates of consciousness, we first have to define what we mean by consciousness. One fundamental and widely accepted distinction is between level of consciousness and the contents of consciousness. Level of consciousness refers to the state of being conscious (as opposed to being asleep, anesthetized, or otherwise unconscious) and can be thought of as an enabling factor that is required for awareness but does not reflect specific conscious experiences. In contrast, the content of an individual's consciousness refers to that individual being conscious of something (e.g., an object in the environment) versus not being conscious of it. Much of the most interesting works on the neural correlates of consciousness has used bistable phenomena to characterize the neural states that are associated with the specific contents of consciousness.

A second important theoretical consideration is the need to clarify the relationship between subjective phenomenal experience of consciousness and the neural states associated with that experience that can be measured. Importantly, it is not clear how any physical process, such as neural activity, can give rise to a subjective phenomenon such as conscious awareness of something, and even the possibility of such a causal relationship is controversial. It must be kept in mind that

empirical research on the neural correlates of consciousness should be neutral to the question of causality. Instead, this research can identify and characterize patterns of neural activity that specifically correlate with conscious as opposed to unconscious perception.

Anatomical-Location versus State-Change Theories of Consciousness

What kind of information can we expect from identifying neural correlates of conscious visual perception? Most of the studies that used bistable perception to investigate the neural correlates of conscious visual perception sought to characterize neural information processing that leads to awareness in contradistinction to processing that proceeds unconsciously. One popular approach is based on the general idea of anatomical location. Such anatomical location theories, also referred to as localizationism, propose that there is one neural structure (or a set of neural structures) in the brain that can generate conscious awareness by virtue of its activity. Information may be processed unconsciously through several stages of the brain but will only reach consciousness through processing in a particular structure of the brain. The most extreme form of an anatomical location theory postulates the existence of one single structure in the brain that is necessary for any information to become conscious. Alternatively, a certain class of neurons or brain regions that share some critical functional properties may equally be able to generate consciousness, but which exactly is critical for consciousness depends on the type of information that is processed. The most advanced anatomical location theories are known as global workspace theories and propose a common functional arrangement, but place the crucial anatomical locus (or loci) in various sites. Most studies investigating bistable perception with fMRI in humans, or electrophysiological recordings in monkeys, are based on such anatomical location theories as they have tried to localize brain activity associated with conscious states or with changes in conscious perception.

In contrast to anatomical location theories, so-called state-change theories postulate a special state of neural activity resulting from a particular way of processing that gives rise to consciousness. The most influential theory of this type proposes that oscillations in neuronal firing and their exact synchrony are the crucial features for conscious awareness. Some studies using electrophysiological measurements in humans or nonhuman primates have used bistable perception to elucidate the role of synchronous oscillations in consciousness, while the fMRI signal is too slow to measure these temporal aspects of neural processing at a sufficiently high resolution. It is important to note that anatomical-location and state-change theories, fundamentally different though they appear, are not mutually exclusive. On the contrary, it is plausible that both aspects of neural activity, location and oscillatory synchrony, may be important aspects of the neural correlates of consciousness.

Bistable Perception and the Neural Correlates of Consciousness

What can we learn from bistable perception about the neural correlates of consciousness? It could be argued that the visual ambiguity that gives rise to bistable perception may be a special case, a situation that can be generated under laboratory conditions but is rarely encountered in natural scenes. What we can learn from the investigation of a special case such as bistable perception may thus be very limited and tell us little, if anything, about the neural mechanisms associated with conscious experience under normal conditions. However, the observation that we usually do not experience ambiguities in natural scenes is misleading. In fact, human vision is routinely faced with conflicting or ambiguous information that necessarily requires active interpretation guided by contextual information, prior experience, and intentions. Accordingly, bistable perception can be seen as a patent manifestation of vision as an active, interpretive process that under normal circumstances is so efficient and effortless that one seldom becomes aware of its inherent difficulty. The study of bistable perception can therefore help understand the interpretive processes that most of the time go unnoticed but give rise to visual consciousness in the form of unified, coherent, and unambiguous percepts. These interpretive processes include the integration of prior expectations that are based on

experience and of contextual information in the visual scene. Moreover, visual processing may also play an active role in continuously reevaluating the current interpretation of the visual input, in stabilizing perception and in weighing and selecting the information available according to its behavioral relevance.

Visual Stimuli Evoking Bistable Perception

Bistable perception can arise from a wide variety of different visual stimulus types. The common denominator of all bistable phenomena is that, while the visual input pattern remains the same, perception changes spontaneously and unpredictably between two states that typically remain stable for a few seconds. All bistable phenomena have in common that periods of dominance and suppression are characterized by sequential stochastic independence, i.e., the timing of the next perceptual switch cannot be predicted from the history of previous switches. Another feature that is shared by all types of bistable perception and, in fact, is so typical that it is sometimes used as a criterion for bistability, is the characteristic gamma distribution of phase durations (see Figure 1(d)). This distribution is typically skewed toward shorter durations. Depending on the stimulus type, changes can occur for various types of visual features, including object category and identity, depth, direction of motion, and visibility. Stimulus types that are known to evoke bistable perception can be classified into reversible figures, ambiguous motion stimuli, and binocular rivalry.

Reversible Figures

Reversible figures include the most famous examples of bistable perception, such as the Necker cube and Rubin's face/vase illusion. The Necker cube is a perspective-reversing wire-frame figure of a cube that can take on two distinct 3D configurations. Further examples for perspective-reversing figures are the Schroeder staircase and the folded card. Rubin's face/vase illusion is a figure-ground reversing stimulus, which can be seen to alternate between a central white shape on a black background and two black profiles of a face against a white background (Figure 1(a)). An interesting feature of figure-ground reversing figures is that high-level stimulus properties such as recognizability can strongly influence the time-course of alternation. For example, with such stimuli it is easier to hold a meaningful shape in the foreground voluntarily when it is upright than when it is inverted. Yet another group of ambiguous stimuli are the classic reversing pictures, the most famous example being Boring's classic illusion "My wife and my mother-in-law," a drawing that can be seen as depicting a young or an old woman. Reversing pictures of this type are usually complex drawings that accommodate two different images by containing ambiguous image-defining features.

Ambiguous Motion

Bistable motion phenomena were first described by German Gestalt psychologists in the first half of the twentieth century. A typical example is the so-called Stroboskopische Alternativbewegung (stroboscopic alternative motion), which is nowadays usually referred to as the apparent motion quartet (Figure 1(b)). Two dots are flashed simultaneously in diagonally opposite corners of an implicit square in rapid alternation with two dots appearing in the other two corners. This sequence yields bistable perception of two dots moving back and forth either horizontally or vertically (and in some observers multistable perception with circular motion as a third possibility), the perception of these directions being mutually exclusive. A related phenomenon is the spinning wheel illusion, where two spoke wheels that are offset to each other by half the spoke distance are presented in rapid alternation. This yields perception of a wheel rotating with the same speed either clockwise or counterclockwise. Another impressive example for ambiguous motion are random-dot kinematograms where, for example, a field of dots each moving in simple harmonic motion around a common axis gives rise to the bistable 3D percept of a rotating cylinder, the direction of rotation being ambiguous.

The ambiguous motion stimuli mentioned so far are symmetrical in the sense that there are in

principle no categorical differences between the two percepts. An example for bistable motion perception that involves changes between two different types of motion are plaid stimuli. They are composed of two orthogonal diagonal gratings that are superimposed in the same display. Observers can perceptually segregate these surfaces and perceive the gratings sliding on top of one another in opposite directions (component motion); or the two components can be integrated perceptually into a single surface that moves in a direction intermediate to the motion directions of the component gratings (pattern motion).

Binocular Rivalry

When dissimilar images are presented to the two eyes, instead of the two images being seen as superimposed or blended, perception alternates spontaneously between each monocular view (Figure 1(c)). Binocular rivalry has many similarities with other bistable percepts. For example, the temporal characteristics of perceptual alternations in binocular rivalry are similar to those in other forms of bistability. Alternations are largely stochastic and percept durations typically follow a gamma distribution (Figure 1(d)), and are similar to other bistable phenomena. However, there are also important differences. With other types of bistable stimuli, different percepts are mutually exclusive. However, this is not always the case in binocular rivalry. For example, when rivalrous stimuli are presented very briefly (<500 ms), the two images can be perceived as superimposed on each other. Piecemeal rivalry can also occur, especially with large stimuli, when a patchy mixture of the two monocular images is perceived. Furthermore, unlike reversible figures, it is very difficult to willfully influence the perceptual alternations in binocular rivalry. For example, perception of reversible figures can be strongly biased by focusing attention on one of the two percepts or on a particular image feature, while such attentional affects are considerably weaker in binocular rivalry. One possible explanation for such differences is that the nature of conflict is quite different between binocular rivalry and other forms of bistable perception.

In the case of reversible figures or ambiguous motion stimuli, the same information is presented to both eyes and the ambiguity of the pattern gives rise to perceptual conflict between two possible interpretations. While perceptual conflict between two images or patterns is also likely to be relevant in binocular rivalry, conflict between monocular signals that are processed separately at the lowest levels of the visual system is also thought to play a significant role in the resolution of binocular rivalry. It is important to keep the differences between binocular rivalry in mind when comparing the results of studies investigating binocular rivalry and those of involved in other forms of bistable perception.

Theoretical Models of Bistable Perception and Behavioral Evidence

What are the neural processes that govern the timecourse of perceptual alternations in bistable perception? The emerging view is that the neural correlates of bistable perception are dispersed throughout many areas of the visual cortex and beyond. However, different researchers have put different emphasis on the involvement of low-level and high-level processes, respectively.

Low-Level Theories of Bistable Perception

Proponents of low-level theories argue that bistability results from neural activity fluctuations at the sensory processing level (Figure 2(a)). Neuronal populations that code for the two possible perceptual interpretations of the visual input are thought to be in dynamic competition. Spontaneous fluctuations and adaptation of percept-related neural activity are two mechanisms that could contribute to perceptual dominance of either percept during ongoing rivalry. In bistable motion stimuli, for instance, different subsets of direction-selective neurons responding to visual motion (e.g., in motion-sensitive area V5/MT in the lateral occipitotemporal cortex, see below) code for the two alternative percepts. The neuronal population that codes for the currently dominant percept will exhibit adaptation over time, with the firing rate of these cells waning. Such adaptation will change the competitive equilibrium between the

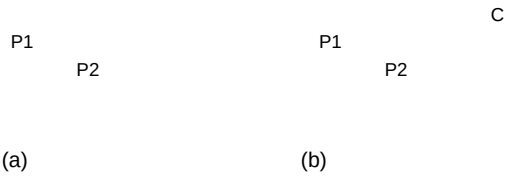


Figure 2 (a) Low-level theory of bistable perception: spontaneous alternations of perceptual states are explained by activity changes of neuronal populations ('P1' and 'P2') in visual cortex that code for the two possible percepts. Possible sources of neural activity changes are adaptation to the currently dominant percept, spontaneous fluctuations, and direct competitive interactions between the two neuronal populations P1 and P2. (b) High-level theory of bistable perception: Perceptual alternations are governed by a central process ('C') involving, for example, frontal brain regions. This central process is thought to evaluate the sensory input and to modify activity in sensory areas via feedback connections. Whenever processes in visual cortex (e.g., adaptation or competitive interactions) act to destabilize activity that underpins the currently dominating percept, higher-order evaluative processes can take effect and initiate a perceptual reorganization.

sensory representations of the two alternative percepts. This could in turn result in the currently dominant neuronal population, and hence the subjective perceptual state, to reverse.

Spontaneous fluctuations of neural activity may additionally affect the dynamics of perceptual alternations. Especially after adaptation of activity coding for the currently dominant percept, spontaneous increases in neuronal activity that codes for the nondominant percept could contribute to a perceptual reversal. In addition to adaptation and spontaneous fluctuations, competitive interactions between neurons coding for rivaling percepts may be an important factor in perceptual bistability. Direct inhibition of neural activity representing the suppressed percept by neurons that code for the currently dominant percept has been implemented in computational models for binocular rivalry, but might also play a role in other forms of bistability.

There is ample evidence to support the idea that adaptation plays an important role in bistable perception. For instance, after prolonged presentation of an unambiguous version of one of the possible interpretations of an ambiguous stimulus, the previously adapted percept is less likely to occur upon subsequent viewing of the ambiguous stimulus. Similarly, prior adaptation of a monocular stimulus presented to one eye increases the likelihood of perceiving the stimulus presented to the other eye during subsequent binocular viewing. The phase duration of a percept can also be influenced by changing the position of ambiguous and binocular rivalry stimuli in the visual field. Perceptual alternation rates are slowed when binocular rivalry or ambiguous stimuli move within the visual field, thereby avoiding local adaptation. Alternation rates can also be dramatically reduced by showing bistable stimuli intermittently, which can be interpreted along similar lines, implying that intermittent removal of the stimulus prevents adaptation. It should be noted, however, that some features of this stabilizing effect of intermittent stimulus removal are difficult to explain with adaptation alone. For example, even if the stimulus is removed for 30 s or more, observers tend to have the same percept when the stimulus reappears as they had before it disappeared. It is not clear why the neuronal populations coding for the two possible percepts should stay in the same firing state over such prolonged periods of stimulus absence.

Another empirical observation that is difficult to accommodate within models relying on adaptation and competitive interactions at the sensory level alone is that periods of dominance and suppression in bistable vision are characterized by sequential stochastic independence. This means that the longer one percept is suppressed, the more strongly it should compete for dominance in the next cycle, if adaptation was the major determinant of bistability. However, such a relationship between sequential dominance phases is never observed in bistable perception. This has led to the suggestion that other processes in addition to those described at the level of sensory processing should be involved in bistable perception.

High-Level Theories of Bistable Perception

Such an alternative view proposes that perceptual alternations are influenced by a central process

that is involved in the evaluation of sensory input and planning behavioral actions, and that can modify activity in sensory areas via feedback connections (Figure 2(b)). A strong argument in favor of this view is based on the observation that the temporal dynamics of bistability are similar to other exploratory behaviors, such as spontaneous exploratory eye movements, and might therefore reflect a general process serving the continuous automatic exploration and reevaluation of the sensory environment. Accordingly, perceptual alternations may not be the consequence of a specialized mechanism limited to situations of extreme sensory ambiguity. Rather, they could reflect a general mechanism that only becomes evident in cases of extreme ambiguity in the sensory input, but is present continually in normal vision and serves an iterative reevaluation of hypotheses or inferences about the perceptual significance of the current sensory input pattern. The general notion that bistable perception is influenced by higher-level, or at least nonsensory, processes is supported by a number of observations. It is a well established finding that voluntary control can strongly influence the temporal dynamics of bistable perception. Observers can not only willfully increase the probability of one percept over the other but it is also possible to change the rate of alternations voluntarily. The latter effect cannot be attributed to retinal shifts due to eye movements (which can also cause perceptual changes), as it is still present when visual stimuli are presented as afterimages or are stabilized on the retina by other means. Other examples for nonsensory influences on the dynamics of bistable perception include expectations, intelligence, mood disorders, laughter, and meditation.

The Special Case of Binocular Rivalry

As mentioned above, binocular rivalry needs special consideration because of the fundamental difference from other types of bistability that there is not only conflict between two interpretations of one sensory input pattern, but between two different patterns presented to the two eyes. Thus, in addition to competition between the neural pattern representations, competition between eye-specific representations at early stages of central visual information processing may contribute to rivalry. Current models of binocular rivalry propose a multilevel process involving competitive neural interactions at both monocular stages (eye rivalry) and binocular stages (pattern rivalry) of the visual processing hierarchy. The degree to which eye rivalry and pattern rivalry are involved in the resolution of binocular rivalry depends on a variety of factors, including the exact timing parameters of stimulus presentation and the type of information contained in the rivaling input patterns. Complex object stimuli are more likely to involve rivalry between high-level pattern representations than simple grating stimuli, where competition at monocular processing levels may play a more decisive role.

Neural Correlates of Bistable Perception

A variety of methodologies have been used to study the neural correlates of bistable perception in humans as well as nonhuman primates, including invasive electrophysiological recordings, electroencephalography (EEG), magnetoencephalography (MEG), and fMRI. For a better understanding of the empirical observations made with these techniques, they are briefly described in the following section.

Measuring the Neural Correlates of Bistable Perception

Using invasive electrophysiological techniques to record the spiking activity of neurons is often considered to be the gold standard that is necessary for a quantitative explanation of perception and behavior in terms of its underlying constitutive elements. However, it requires electrodes to be inserted through the skull into the brain, which puts subjects at risk of injury and infection. With the exception of the rare case that electrodes are implanted in humans for diagnostic reasons, e.g., for presurgical mapping in epilepsy, such measurements are limited to nonhuman primates or other animals for ethical reasons. In principle, such electrodes can be used to measure single-unit activity, multiunit activity, and local field potentials (LFP). Single- and multiunit activity reflects

the action potentials (the basic electrical signals required for information processing in the nervous system) recorded from one or more neurons in the vicinity of an electrode. LFPs represent the aggregate activity of a population of neurons located close to the electrode and are thought to reflect mainly synaptic activity associated with both input to and output from a cortical area.

In human subjects, fMRI has become a popular tool to study the neural correlates of bistable perception. fMRI is based on the measurement of a blood-oxygen level dependent (BOLD) signal, which is a component of the hemodynamic response that is associated with local neural activity. The BOLD signal is essentially linearly correlated with multiunit activity and even more strongly with LFPs. It might therefore reflect pre- and postsynaptic activity and hence the processing in an area more closely than multiunit activity. One drawback of fMRI is that it does not measure neural activity directly and that it has a limited temporal resolution (in the range of seconds) due to the sluggishness of the hemodynamic response. Despite great technical improvements over the last years, fMRI has also limited spatial resolution (in the range of millimeters) and can therefore only measure signals that are generated by relatively large populations of neurons. However, the spatial resolution is still considerably better than that of other noninvasive methods as EEG and MEG. The greatest strength of fMRI is its capacity to record activity from the entire brain essentially simultaneously. In contrast to single- or multiunit recordings, fMRI, therefore, in principle does not require prior assumptions as to where activity is expected to occur.

EEG and MEG measure the electric activity of the brain by recording from electrodes placed on the scalp and the magnetic fields produced by electrical activity in the brain, respectively. EEG and MEG responses to external stimuli, so-called event-related potentials (ERPs) are also thought to reflect the summed electrical effects of synaptic neurotransmission in large neuronal populations. Similar to the association between the BOLD signal and LFPs, there is also a nearly linear relationship between ERP amplitude measured with EEG and BOLD signal measured with fMRI in sensory cortices. While EEG and MEG have a much higher temporal resolution than fMRI, their spatial resolution is inferior to fMRI and especially EEG, is limited to recording only electrical activity that is generated in superficial cortical regions within the reach of scalp electrodes.

Taken together, there seems to be an overall good agreement between these various measures of neural activity, but it is important to bear in mind that the noninvasive methods available (fMRI, EEG, and MEG) measure neural activity associated with populations of neurons only indirectly and therefore need to be interpreted with caution. On the other hand, single cell measures have – in addition to the fact that they are usually only available from nonhuman primates and can therefore not easily be extrapolated to humans – the disadvantage that the number of neurons that can be recorded simultaneously is limited.

Choosing the Bistable Stimulus Paradigm

The majority of studies have investigated binocular rivalry rather than other forms of bistability. This tendency is most pronounced in electrophysiological studies in animals, which may be due to the fact that bistable perception is easier to achieve and to validate in nonhuman subjects using binocular rivalry stimuli than with reversible figures, the ambiguity of which may rely more on higher cognitive functions. But also in human studies binocular rivalry has been a very popular paradigm. One reason for this may be that binocular rivalry is a particularly intriguing phenomenon, but more mundane factors such as practical feasibility and functional anatomy of the visual system can also bias the choice of experimental paradigm. For example, binocular rivalry is very flexible with regard to the stimulus material used. In principle, any stimulus class can be used in binocular rivalry – from grating stimuli optimal for the investigation of low-level processing, to complex object stimuli that are processed in higher-level visual areas. The use of object stimuli in binocular rivalry has the additional advantage that the processing of some object categories is to some degree spatially segregated in human visual cortex (see below). Responses to two rivaling object categories (e.g., faces and houses) can therefore easily be dissociated using fMRI. To a limited degree, this

may also be possible with selected reversible figures involving complex objects, but is much more difficult with geometrical stimuli like the Necker cube or bistable motion stimuli, where responses to the two alternative percepts are not so clearly segregated. As noted above, however, it should be kept in mind that the interpretation of binocular rivalry studies with regard to the neural correlates of consciousness is complicated by the involvement of mechanisms specifically operating at the monocular levels of stimulus processing.

Subcortical and Early Cortical Processing

The first stage of central visual processing is the lateral geniculate nucleus (LGN) of the thalamus. The retinal projections from each eye terminate in different laminae of the LGN, so that they remain segregated and processing is hence strictly monocular. Some monocular processing is still preserved at the lowest level of cortical visual processing, in the ocular dominance columns of primary visual cortex (V1). Beyond V1 visual processing is essentially binocular. That is, input from corresponding retinotopic locations from the two eyes is processed together and information about the eye of origin appears not to be preserved at these higher levels.

The role of the earliest stages of visual processing in the brain, the LGN and V1, in bistable perception is controversial. This is because studies using different methodologies (single unit recordings and fMRI) in different species (monkeys and humans) have yielded diverging results. Single unit recordings in the LGN of awake monkeys during binocular rivalry provide no evidence for a correlate of rivalry in the LGN, whereas fMRI studies in humans found that BOLD responses in the LGN reflect eye-specific dominance and suppression. A similarly incongruent picture has emerged from studies investigating the role of V1 in binocular rivalry. In electrophysiological recordings in awake monkeys, only a small percentage of cells showed activity that reflected dominance and suppression in binocular rivalry, while human fMRI studies have demonstrated strong effects of binocular rivalry in V1. Differences between methodologies and species may contribute to such divergent findings, which await further clarification. With regard to consciousness, it has been proposed that the evidence from human fMRI studies indicates an important role for early visual processing, at least in V1, for conscious awareness. However, activity in LGN and V1 may merely reflect the dominance of one monocular processing channel over the other, thus gating what information reaches higher-level visual areas rather than being directly related to conscious perception. A similar role for early visual cortex is suggested by fMRI findings in bistable apparent motion perception. Whenever apparent motion is inconsistent with additional image cues (e.g., color), early visual cortex activity is suppressed, which may reflect regulatory mechanisms that flexibly gate early visual feature processing in accord with an overriding perceptual decision.

Intermediate Levels of Visual Processing

Beyond V1, the primate visual system is organized in a distributed fashion, with different aspects of the visual scene being analyzed in different cortical areas. Visual areas beyond V1 (striate cortex) are at large called extrastriate visual cortex. Extrastriate processing is divided in two major pathways, the dorsal and ventral streams. The dorsal stream is associated with motion processing and representations of object locations. It extends from V1 through areas V2, V3A, and V5/MT to the inferior parietal lobe. Area V5/MT is in humans located in the lateral occiptotemporal junction and plays a prominent role in the processing of motion signals (Figure 3). The ventral stream passes from V1 through V2 and V4 to inferotemporal cortex. V4 is located in the ventral occipital cortex anterior to V2 and is thought to preferentially process color information. The inferotemporal cortex has an essential role in higher visual functions, such as object recognition. In humans, inferotemporal cortex accommodates various functionally characterized areas that are associated with certain object categories, most notably the fusiform face area and the parahippocampal place area.

Electrophysiological recordings in early extrastriate areas (V4 and V5/MT) of awake macaque monkeys reporting rivalry showed activity modulations that were stronger than in V1 but still modest compared with the perceptual changes experienced during rivalry. In contrast, responses were markedly different in inferotemporal cortex.

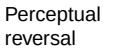


Figure 3 Spontaneous reversals of perceived motion direction (from vertical to horizontal motion perception, as shown here, or vice versa) during bistable apparent motion (see also Figure 1(b)) are associated with fMRI activations in motion-sensitive visual areas, most notably V5/MT (black circle), and in frontal and parietal brain regions.

Recordings showed that most of the inferotemporal neurons were active only when their preferred stimulus was perceived and showed essentially no activity during the perceptual suppression of the stimulus, indicating that inferotemporal cortex represents a stage of processing beyond the resolution of perceptual conflict. In humans, invasive electrophysiological recordings were made from the temporal lobes of patients in whom electrodes had been implanted for diagnostic reasons and showed results compatible with those obtained from monkeys. During binocular rivalry, medial temporal lobe neurons tuned to a specific stimulus category fired selectively when their preferred stimulus was perceived, but not when it was perceptually suppressed and invisible. Findings from fMRI studies in humans experiencing binocular rivalry support the notion that binocular conflict is fully resolved at higher levels of the visual processing hierarchy. During rivalry, fMRI responses that are recorded in the fusiform face area to face stimuli – and in the parahippocampal place area to images of places – are large and equal in amplitude to responses evoked by nonrivalrous stimuli. Interestingly, temporary removal of binocular rivalry stimuli, when the stimulus perceived on reappearance tends to be the one in awareness as they disappeared, is accompanied by persistent activity in functionally specialized regions of human visual cortex representing the last percept before stimulus removal. Such activity during stimulus absence is not measurable during removal of nonrivalrous stimuli, suggesting a role for specialized extrastriate cortex in stabilizing perception in situations of perceptual conflict.

Studies with reversible figures and ambiguous motion stimuli support the notion that intermediate levels of visual processing are involved in the resolution of visual ambiguities and conflict. In parallel with the observation in binocular rivalry using house and face stimuli, fMRI activity in the fusiform face area is greater during face perception during viewing of Rubin's face/vase illusion (Figure 1(a)). Percept specific activity in extrastriate visual areas can also be found during bistable motion perception. Electrophysiological recordings in monkeys showed that during viewing of a bistable random-dot kinematogram, neuronal firing in area V5/MT correlates with the reported direction of rotation. Distinguishing activity associated with different directions of motion is more difficult with fMRI because there is no spatial segregation between cells selective of different motion directions that could easily be resolved on the basis of fMRI signals. However, motion stimuli that are ambiguous with respect to the type of motion have shown differential activity in human motion-sensitive extrastriate cortex. During viewing of ambiguous plaid stimuli where perception alternates between global motion and component motion, stronger activity in V5/MT is found during component-motion perception. Similarly, the comparison of object-motion with illusory self-motion during large-field stimulation with a rotating stimulus is accompanied by differential activation in V5/MT and also in area V3A. Finally, when a simple apparent motion stimulus composed of two dots shown in alternation in different visual field locations is titrated such that perception of apparent motion is equally likely to occur as just a flickering of the two dots, V5/MT is more strongly activated whenever human observers report apparent motion perception.

In contrast to the substantial evidence for an involvement of V5/MT in bistable motion perception, evidence for involvement of early extrastriate areas in the ventral stream of visual processing, e.g., V4, are scarce. This is because bistability between basic object features such as color is difficult to achieve experimentally. However, when color or luminance cues are used to bias bistable apparent motion perception, activity in V4 is specifically increased whenever perception is consistent

with color cues, suggesting that extrastriate color processing influences the resolution of perceptual ambiguities.

While the work mentioned so far has capitalized on identifying differences in activity associated with different perceptual states, another line of research has been primarily concerned with the neural events associated with and time-locked to perceptual reversals. fMRI studies investigating reversal-related activity have used a variety of bistable paradigms including binocular rivalry, reversible figures, and ambiguous motion stimuli. In addition to distributed activations in frontal and parietal brain regions (see below), these studies have consistently found transient activity increases in extrastriate visual areas in association with perceptual reversals. These activations are tuned to the visual feature or attribute that is perceived to change. While changes involving face or object percepts are accompanied by activations in object processing areas of the ventral stream, perceived changes in motion direction or motion type are associated with activations in motion-sensitive areas, most notably V5/MT (Figure 3).

Taken together, electrophysiological and fMRI studies have established that neural activity in functionally specialized extrastriate cortex correlates with subjective perception during bistability. While it is difficult to make strong quantitative statements on the basis of fMRI measurements, single-cell recordings in monkeys suggest that the degree of correlation increases at successive stages of the visual processing hierarchy. While the exact role of LGN and V1 is still controversial, it seems now firmly established that high-level areas in inferotemporal cortex represent a stage of processing beyond the resolution of perceptual conflict. These findings have often been taken as evidence for a crucial role of higher-level visual cortex in consciousness. It has been argued that if neurons in these areas show such strong modulations in accord with perception, neural activity in these areas should be a prerequisite for visual information to gain access to consciousness.

Parietal and Prefrontal Cortex

The possible role of brain regions outside the visual system has been investigated almost entirely in human subjects. Neuroimaging investigations into the neural correlates of bistable perception have indicated that activity in the parietal and prefrontal cortices might be associated with conscious perception in normal subjects. As mentioned in the section titled 'Intermediate levels of visual processing,' these studies measured brain activity time-locked to spontaneous perceptual reversals both during binocular rivalry and for other bistable stimuli. In addition to reversal-related activations in extrastriate visual areas, cortical regions with activity that reflects perceptual transitions include inferior parietal and inferior frontal cortex, regions previously implicated in regulating access of sensory information to consciousness. While extrastriate areas are equally engaged by nonrivalrous perceptual changes, parietal and prefrontal regions show significantly greater activation associated with perceptual alternations during viewing of rivalrous or ambiguous stimuli.

The attempt to define in more detail the functional significance of reversal-related activations in higher-order brain structures inevitably leads into a primacy debate resembling the chicken-and-egg problem. In one view, changes in perception are caused by neural activity fluctuations in visual cortex (the low-level theory propounded above, see Figure 2(a)). Whenever a perceptual change occurs, the underlying neural event in visual cortex is communicated to higher-order areas and entrains their activation in a feed-forward fashion, similar to external stimulus changes. An alternative view also proposes that visual cortical areas house the competing perceptual representations and that perceptual dominance is ultimately underpinned by their relative degree of activity. However, in this view, the reorganization of activity in visual cortex during perceptual reversals is initiated and instructed by frontal and parietal brain structures (the high-level theory described above, see Figure 2(b)). These two scenarios differ in the causal chain assumed to underlie changes in visual awareness, but it remains difficult to infer causality from correlative neurophysiological measures. Still, temporal precedence is generally considered good evidence in favor of a putative causal role. Invasive neurophysiological recordings would therefore appear ideally suited to resolve this question but suffer from the uncertainty of where exactly to place recording electrodes for instance in the frontal lobe.

Recent analytical approaches have made it possible to resolve latency differences between fMRI responses in the range of a couple of hundreds of milliseconds. Indeed, chronometric analyses of fMRI activations in association with spontaneous changes on apparent motion perception showed that activations in right inferior frontal cortex occur earlier during bistable perception than during externally induced perceptual changes, while such temporal precedence is absent in extrastriate visual cortex. This suggests that prefrontal structures may participate in initiating spontaneous reversals during bistable perception. In line with this notion, studies in patients with brain lesions have shown that spontaneous perceptual alternations are slowed down in cases of focal damage to the right prefrontal cortex. The low-level theory that adaptation of percept-related neural activity plays an important role in perceptual bistability is not irreconcilable with a causal role of higher-order processes in initiating perceptual changes. The actual alternations of perception could be determined by the joint effect of local processes embedded into a more global process. That is, whenever local processes (e.g., adaptation) act to destabilize activity that underpins the currently dominating percept, higher-order evaluative processes can take effect and initiate a perceptual reorganization (Figure 2(b)).

In addition to participating in the initiation of perceptual changes, higher-order brain structures may also contribute to the stabilization of perception. As mentioned above, perception can be stabilized by intermittently removing a bistable stimulus, and this effect is still present when the stimulus disappears for intervals in the range of tens of seconds. An individual observer's tendency to stabilize a percept across such periods of stimulus removal strongly correlates with brain activity in frontal and parietal regions previously implicated in working memory. This suggests that higher-order brain structures in prefrontal and parietal cortices may not only play a role in the initiation of perceptual reversals but also in the stabilization of perception.

The Role of Neural Synchronization

This account of the neural underpinnings of bistable perception has so far almost exclusively been concerned with their anatomical location. Relatively few studies have addressed the question of whether bistable perception may be related to neural synchronization, an approach based on state-change theories of consciousness. The role of neural synchronization in binocular rivalry was studied in awake cats viewing dichoptically presented drifting gratings with orthogonal orientations. Electrophysiological recordings from visual cortex show that neurons representing the dominant stimulus increase their synchrony, whereas cells that process the suppressed pattern decrease their temporal correlation. In contrast to neural synchrony, however, the firing rates of cells responding to the dominant and the suppressed stimulus did not differ. This finding may provide a link between the divergent results from electrophysiological recordings and fMRI measurements of early visual processing, especially as the fMRI signal also correlates with the degree of neural synchronization.

In humans, MEG was used to identify synchronization in neural activity between spatially distributed cortical areas with dominance phases in binocular rivalry. Rival gratings flickering at different frequencies were used to tag the MEG signals associated with the two gratings. Over a wide array of sensor locations encompassing the occipital, parietal, temporal, and frontal lobes, the amplitudes of the MEG responses correlated with observers' reports of dominance and suppression. Dominance phases of rivalry were also associated with marked increases in synchronization of MEG signals recorded from widely distributed sensors, with the most prominent examples of synchronization arising in frontal areas of the brain.

Taken together, there is thus evidence demonstrating that neural synchronization, both locally within visual cortex and globally among spatially remote brain regions, may contribute to the resolution of visual conflict in bistable perception.

Conclusions

What have we learned from bistable perception with regard to the neural correlates of consciousness? Quite apart from the findings from electrophysiological and neuroimaging studies, bistable perception is an impressive illustration of the constructive nature of perception. The fact that visual

perception can change while the physical input remains constant clearly demonstrates that visual perception results from active interpretive processes in the brain. While this general argument is historically important, empirical studies of bistable perception over the last decade have yielded the following insights:

Bistable perception is a multilevel process. Neural correlates of bistable perception are measurable throughout the visual system and beyond.

Neural activity fluctuations in neuronal populations representing the competing percepts are involved in determining the perceptual outcome and hence influence conscious awareness.

High-level visual areas play a key role in consciousness. Activity in these areas reflects the actual perceptual outcome of the processes involved in the resolution of visual ambiguity and conflict.

Central processes involving prefrontal and parietal brain regions can exert an influence on bistable perception by stabilizing the current percept or by initiating perceptual reversals. Conscious perception is therefore a result of local sensory processes embedded into more global evaluative and interpretive processes.

Neural synchronizations are likely to be involved in generating a coherent conscious percept.

We conclude that bistable perception has proven a valuable tool for neuroscientific research and yielded important insights into the neural correlates of conscious vision. Some aspects, such as the prominent role of high-level extrastriate visual areas in conscious vision, are supported by a significant number of studies using very different paradigms and methodologies and therefore appear to be firmly established. Research into the neural mechanisms of bistable perception has also shed light on the role of early visual processing, which may be more important for conscious vision than previously thought. However, more research is needed to resolve controversies that have arisen from seemingly contradictory results, and to provide a more detailed account of the role that low-level processing, especially in V1, plays in consciousness. Similarly, fascinating new insights were gained about interactions between higher-order frontoparietal regions and visual cortex and about the role of neural synchronizations in conscious vision. Future research using bistable perception and related phenomena should try to provide a more detailed account of how these processes contribute to human consciousness.

See also: The Neural Basis of Perceptual Awareness; Perception: Unconscious Influences on Perceptual Interpretation; Visual Imagery and Consciousness.

Suggested Readings

Alais D and Blake R (2005) Binocular Rivalry. Cambridge, MA: MIT Press.

Blake R and Logothetis NK (2002) Visual competition. Nature Review, Neuroscience 3: 13–21.

Haynes JD, Deichmann R, and Rees G (2005) Eye-specific effects of binocular rivalry in the human lateral geniculate nucleus. Nature 438: 496–499.

Kleinschmidt A, Büchel C, Zeki S, and Frackowiak RSJ (1998) Human brain activity during spontaneously reversing perception of ambiguous figures. Proceedings of the Royal Society of London, Series B 265: 2427–2433.

Leopold DA and Logothetis NK (1999) Multistable phenomena: Changing views in perception. Trends in Cognitive Sciences 3: 254–264.

Leopold DA, Maier A, Wilke M, and Logothetis NK (2005) Binocular rivalry and the illusion of monocular vision. In: Alais D and Blake R (eds.) Binocular Rivalry, pp. 231–258. Cambridge, MA: MIT Press.

Long GM and Toppino TC (2004) Enduring interest in perceptual ambiguity: Alternating views of reversible figures. Psychological Bulletin 130: 748–768.

Lumer ED, Friston KJ, and Rees G (1998) Neural correlates of perceptual rivalry in the human brain. Science 280: 1930–1933.

Parker AJ and Krug K (2003) Neuronal mechanisms for the perception of ambiguous stimuli. Current Opinion in Neurobiology 13: 433–439.

Rees G (2007) Neural correlates of the contents of visual awareness in humans. Philosophical Transactions of the Royal Society of London. Series B, Biological Science 362: 877–886.

Sheinberg DL and Logothetis NK (1997) The role of temporal cortical areas in perceptual organization. Proceedings of the National Academy of Sciences of the United States of America 94: 3408–3413.

Sterzer P and Kleinschmidt A (2007) A neural basis for inference in perceptual ambiguity. Proceedings of the National Academy of Sciences of the United States of America 104: 323–328.

Tong F, Meng M, and Blake R (2006) Neural bases of binocular rivalry. Trends in Cognitive Sciences 10: 502–511.

Tong F, Nakayama K, Vaughan JT, and Kanwisher N (1998) Binocular rivalry and visual awareness in human extrastriate cortex. Neuron 21: 753–759.

Tononi G, Srinivasan R, Russell DP, and Edelman GM (1998) Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proceedings of the National Academy of Sciences of the United States of America 95: 3198–3203.

Biographical Sketch

Philipp Sterzer (born 1970, Germany) is a supervising psychiatrist and the head of the Visual Perception Laboratory at the Department of Psychiatry, Charité University Hospital, Berlin. After medical training at Ludwig Maximilian University in Munich and Harvard Medical School in Boston, USA, he obtained his MD under the supervision of Florian Holsboer at the Max-Planck Institute of Psychiatry in Munich. He completed his training in clinical neurology at Johann-Wolfgang-Goethe University in Frankfurt am Main, where he also worked as a postdoctoral research fellow in the laboratory of Andreas Kleinschmidt. He then worked as a postdoctoral fellow in Geraint Rees's group at the Wellcome Trust Centre for Neuroimaging, University College London, before moving to Berlin in 2006 to work as a clinical psychiatrist. In 2008, he became a group leader with the award of an Emmy Noether junior research group from Deutsche Forschungsgemeinschaft. His main research interests include the neural basis of visual awareness, the interactions of emotion and motivation with visual awareness, and alterations of these processes in mental diseases.

Geraint Rees (born 1967, the United Kingdom) is a professor of cognitive neurology and Wellcome Trust Senior Clinical Fellow at the Institute of Cognitive Neuroscience and Wellcome Trust Centre for Neuroimaging, University College London. After medical training in Cambridge, Oxford, and London, he completed his PhD under the supervision of Chris Frith at University College London's Functional Imaging Laboratory. He then worked as a postdoctoral fellow in Christof Koch's laboratory at the California Institute of Technology for two years before returning to the Institute of Cognitive Neuroscience at University College London in 2001. In 2002, he became a group leader with the award of a senior fellowship from the Wellcome Trust. His work has been internationally recognized with the award in 2003 of the Young Investigator Medal of The Organization for Human Brain Mapping. In 2007, he was awarded the Experimental Psychology Society prize and gave the Royal Society Francis Crick lecture. His research seeks to understand the neural basis of human consciousness.

Blindsight

C T Trevethan and A Sahraie, University of Aberdeen, Aberdeen, UK

Glossary

Contrast sensitivity function – A function which relates the perception of image components to the reciprocal of minimum contrast energy needed for their detection. The contrast sensitivity function of a visual system represents a combination of both optical and neuronal sensitivity.

Perimetry – Methods for quantifying visual sensitivity. In an automated perimeter, visual targets of specific size or contrast are presented at multiple retinal locations. Visual sensitivity is then determined by subjective reports of the observer on whether s/he detected the target.

Psychophysics – Study of the relationship between physical parameters of a stimulus and its resultant psychological percept. The discipline was founded by Fechner in eighteenth century.

Pupillometry – Determining the size of the pupil aperture. Various techniques have been developed to determine the steady state pupil diameter and its fluctuations over time.

Retinotopic – Related to the retinal image. A topographic retinotopic representation means that each point on the retinal image is represented in a systematic fashion.

Scotoma – Area of total or relative visual loss within someone's field of vision.

Signal detection theory – An approach originally developed for engineering applications, and subsequently applied to human behavior which allows determination of subjective false alarms and bias in addition to sensitivity, not addressed in Fechnerian psychophysics.

Spatial frequency – Number of cyclic variations of light and dark per unit measure of space.

Two alternative forced choice paradigm – A psychophysical technique for determining sensitivity. In spatial version, two targets are presented simultaneously in separate spatial locations and the observer is asked to choose between the two. In the temporal version, targets are separated in time, and the time intervals often signaled via a separate modality. For example, the temporal intervals for presentation of visual targets may be separated by auditory bleeps.

Visual field defect – Areas of relative or absolute blindness (scotoma).

Introduction

Put simply, blindsight refers to the ability to process certain visual stimuli presented within an area of visual field which is clinically blind, following occipital brain injury. However, the concept, theory, and practicalities of blindsight are far from simple! This area of research is exciting, complex and often controversial. Investigations into blindsight have important implications for our understanding of normal visual processing, visual awareness and consciousness, as well as plasticity and recovery of function following brain injury. Blindsight has an interesting history in terms of the links between animal and human research. It is also an area in which objective science and the study of subjective, internal phenomena meet, raising many interesting theoretical and methodological challenges.

Stroke, closed head injury, congenital abnormalities, and surgery can all cause damage to striate cortex (also know as primary visual cortex, V1 and Brodmann area 17) and result in an area of cortical blindness affecting the corresponding region of visual field. When areas of cortical blindness

(also known as visual field defects (VFDs)) are suspected, vision is tested clinically using a method called perimetry. During perimetric testing, a patient fixates (looks steadily at) a central point and pushes a buzzer each time they are aware of a small light flash presented throughout their visual field. Because lights presented within a VFD are not consciously perceived, the person will fail to push the buzzer when these lights are presented, creating a 'map' of the subjectively blind area of vision. Blindsight refers to the ability to process certain visual stimuli presented within these perimetrically blind areas of visual field under certain conditions.

Early experimental research into the impact of removal of striate cortex in nonhuman primates initially suggested survival of some visual abilities. Although some later research raised uncertainties about the limits of the lesions in some of these studies, it now seems clear that monkeys without primary visual cortex retain a range of visual functions. In contrast, initial observations suggested that human patients without primary visual cortex were phenomenally blind (i.e., they were not aware of perceiving any visual stimuli). The origins of blindsight lie in investigations of this apparent contrast between human and nonhuman primate visual systems. Unexpectedly, the use of animal methodologies to test human patients revealed that although patients stated that they were using 'guesswork' they demonstrated a range of visual abilities in the absence of striate cortex.

Although the concept of blindsight is now established, from the outset, general theory, practical issues, and criticisms have been raised and continue to be debated. Methodological issues remain vital for preventing potential experimental artifacts, selecting appropriate stimuli and experimental paradigms in order to elicit blindsight when present. Blindsight, for itself and its associated possibilities, such as insight into normal visual processing, is an area of lively activity.

One of the most intensively studied blindsight patients (DB), demonstrated detection and localization of specific visual stimuli, as well as discrimination of orientation, movement and simple form. Other visual attributes discriminated in the absence of conscious experience include; color, orientation, simple shapes, motion and the onset and offset of visual events. Implicit methods, for example, measurement of minute changes in pupil diameter (pupillometry, see also the later section on pupillometry) have also revealed evidence for processing of visual stimuli within blind areas of visual field. Physiological measures that do not rely on asking about a person's subjective experience of visual events hold promise as a potential screening tool for blindsight. To date, pupillometry has demonstrated evidence for processing of color, movement and grating stimuli in the absence of conscious visual experience. Although much of the research evidence for blindsight in humans comes from single cases or small group studies, there is now increasing evidence that blindsight may not be a rare phenomenon. The possibility that blindsight may not be as rare as once thought, raises important questions about the development of rehabilitation strategies with the aim of potentially recovering some degree of visual function.

Early Experimental Evidence from Nonhuman Primate and Human Cases

Early Human Evidence

Much of the early evidence relating to VFDs was from cases of war veterans who suffered gunshot wounds to the head during the Russo-Japanese, World War I (1914–1918) and World War II (1939–1945). As the lesions caused by gunshot wounds were often fairly specific, they provided the opportunity to investigate the relationship between occipital lobe damage and resulting VFDs. Even at this time, there was disagreement about whether any visual abilities remained following destruction of striate cortex. In 1918, Gordon Holmes, a prominent ophthalmologist, published a summary of 15 cases of visual cortex and optic radiation lesions and resulting VFDs. Holmes reported a systematic, topographical representation of the contralateral hemifield of vision in the striate cortex (see Figure 1). Holmes concluded that scotoma were areas of total, permanent blindness. The view that destruction of unilateral striate cortex resulted in total and permanent blindness affecting the contralateral visual field and the view that bilateral destruction resulted in

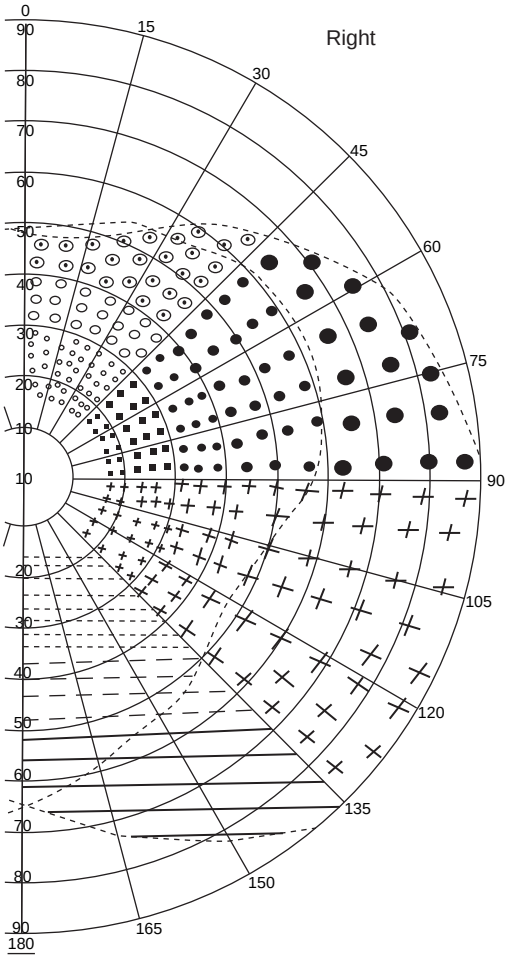


Figure 1 On the left the striate area of the left hemisphere of the brain is shown, the calcarine fissure being opened up to reveal that portion of it which lines its walls. On the right is right half of one visual field. The correspondence of the markings indicates the representation of different segments of the visual field on the cortex. Reproduced from Holmes G (1945) Ferrier lecture: The organization of the visual cortex in man. Proceedings of the Royal Society of London. Series B, Biological Sciences 132(869): 348–361.

complete blindness were widely held, particularly by the medical establishment. Poppelreuter claimed that no scotoma was 'absolutely' blind. He maintained that some form of rudimentary function was present and could be elicited by increasing the size or luminous intensity of the stimulus. Poppelreuter ordered different levels of visual function from: amorphous light sensitivity, size perception without definite form, amorphous form perception, perception of discrete objects, mild amblyopia, and normal vision. An investigation of gunshot cases from the World War II by Teuber revealed some discrimination between light and dark in VFDs. However, the possibility of light scattering into the intact area of visual field is difficult to rule out. Following the investigations into visual deficits resulting from gunshot wounds during the two world wars there was relatively little interest in residual visual abilities in humans as the focus shifted to nonhuman primate research.

Early Nonhuman Primate Evidence

Perimetry is the systematic measurement of the visual field by detection of targets (usually light stimuli) presented throughout the visual field. The use of perimetry for mapping VFDs was adapted for nonhuman primates by Alan Cowey in 1963. Until this point, there were mixed reports about the extent of the impact of removal of striate

cortex on vision in monkeys and progress was limited by the lack of a standard method of 'mapping' VFDs. The ability to study VFDs perimetrically meant it was possible to precisely define the characteristics of the deficit and relate these findings to visual organization in the monkey. Monkeys were trained to press a lever on their left when a buzzer was presented alone and a button on their right when a buzzer was presented simultaneously with a light flash from a bulb. The animal's responses could not be based on expectation as they had no means of knowing on which trial the bulb would be illuminated. If they were simply guessing there would be a high proportion of false positive responses. It was concluded that monkeys can conceal large VFDs by scanning to use intact areas of vision. One problem with this research was that it did not adequately control for light scattering from the bulbs into intact areas of visual field. This research was later extended by comparing the behavioral effects of retinal lesions and cortical damage by using the retinal lesions and natural blind spot as controls. Overall, it was demonstrated that VFDs following unilateral striate cortex removal are not absolute and are not as severe as the natural blind spot or retinal lesions. It had also been established that, with systematic practice, the VFD resulting from a striate cortex lesion could actually reduce in size. In contrast, when a striate cortex lesion was followed by an additional retinal lesion, the VFD no longer shrank with practice. Targets that were detectable following striate damage were not detectable following only a retinal lesion. Comparing the detection abilities within the natural blind spot as well as in VFDs associated with retinal lesions and the detection ability within VFDs associated with brain damage, demonstrated that there are clear distinctions between the two types of blindness. This research was also important for highlighting the possibility of improvement within VFDs associated with striate cortex damage with practice.

Following the above developments, was a case study of 'Helen,' a monkey who was extensively observed and tested over 8 years after bilateral removal of striate cortex.

Helen demonstrated localization of moving stimuli, size discrimination, brightness discrimination, rudimentary figure-ground differentiation, and the ability to move around space avoiding large obstacles. However, evidence from formal testing and behavioral observations suggested that Helen consistently failed to recognize familiar objects, for example, fruit. Interestingly, her performance also deteriorated markedly if she was scared or upset.

At this stage, some of the main research findings suggested that discrete lesions of visual cortex produce discrete VFDs. Nevertheless information of specific types, could still be detected within the VFD. Monkeys had superior visual abilities following striate cortex lesions compared to comparably sized retinal lesions. Following striate cortex lesions, the defective area gradually shrank in size and became increasingly sensitive, suggesting at least a degree of plasticity. Even when striate cortex is removed completely, visual information can still be processed and vision may be no worse than amblyopic. It was noted that these residual visual abilities in destriated monkeys were likely to be mediated by the posterior association areas as, if these were removed all but extremely crude discriminations were possible. The importance of attention–provoking stimuli was also highlighted as the monkeys seemed able to notice and fixate these stimuli but unable to examine them further once they were detected.

The full extent of residual visual abilities following destriation in monkeys became clearer when a detailed account of a number of experiments carried out on 82 rhesus monkeys with total, bilateral ablation of striate cortex was published in 1982. Combined with other research reports, these findings supported the view that monkeys without visual cortex retained or recovered visual capacities, which exceeded classical expectations. Even immediately after removal of striate cortex, two mechanisms mediated by the midbrain remained. These were the pupillary light reflex response and the blink reflex in response to increases in illumination. Over a period of weeks visually evoked nystagmus (rapid, rhythmic, repetitious, involuntary eye movements) recovered, then, over time, the monkeys gradually appeared to relearn a range of visual abilities. These recovered skills included: accurate reaching, discrimination between figures on the basis of differences in total light flux (amount of perceived light power) and discrimination of a variety of flux-equated (where targets are

equated for perceived light power) targets on the basis of shape (triangle vs. circle), orientation, spatial frequency, contrast and wavelength. Despite the recovery of these abilities, the monkeys' visual sensitivity was still greatly decreased compared to preoperative levels. For example, preoperatively monkeys could discriminate all spatial frequencies between 0.5 and 32 cycles per degree compared to 0.5 and 4 cycles per degree postoperatively.

The potential importance of subcortical processing in blindsight was highlighted by comparing the effects of lesions of striate cortex and superior colliculus on detection of visual stimuli and accuracy of saccadic eye movements. Following unilateral ablation of striate cortex, monkeys gradually improved with practice at detection of light stimuli. This improvement was clearly a practice effect, as trained areas of visual field improved more than untrained areas. Following superior colliculus lesions, a midbrain nuclei with extensive multimodal processing capability and connectivity (in the absence of cortical damage), there was no evidence of detection or significant saccadic deficits. Importantly, following joint lesions of striate cortex and superior colliculus, no recovery was observed, even after 15 weeks of testing. It seemed that either striate cortex or superior colliculus was 'sufficient' for visual detection and visually guided saccades and that one or a combination of these two structures are necessary. Lesions in both (in either order) appear to result in permanent deficits in visual detection and saccade accuracy. These findings were strengthened by further research demonstrating that monkeys were able to reach accurately toward targets that were illuminated very briefly (110 ms), controlling for the possibility of discrimination on the basis of eye or head movements.

Recent Nonhuman Primate Evidence

More recently, the fascinating question of whether destriated monkeys experience phenomenal blindness within their VFDs in the same way as human patients, or whether they subjectively perceive stimuli has been investigated. Monkeys were trained to reach out and touch light stimuli when they were presented on screens in front of them. They were also trained to touch a different area when it was a 'blank' trial (i.e., no stimulus presented). In 10% of trials in which stimuli were presented within the monkey's blind field, the monkey almost always categorized the trial as a 'blank' by touching the 'blank area.' This finding suggested that monkeys were experiencing blindsight in the same way as human cases. That is, although they were capable of reaching for these stimuli when they were presented within their blind field, they classified them as eliciting no visual experience.

The extent to which monkey residual visual abilities parallel those found in humans has been further investigated. The ability of monkeys with unilateral removal of striate cortex to make accurate saccades to visual targets under two different conditions has been compared. When monkeys fixated a central point and targets were presented throughout their visual field (analogous to clinical perimetry), they failed to initiate saccades to targets in their blind field, appearing blind. Yet, when the central fixation point was extinguished simultaneously with the onset of the target, they were able to successfully localize targets presented within the blind region of their visual field. There are two possible explanations for these findings. First, that the fixation offset facilitated residual visual performance by cueing the monkey to make the eye movement or, that the fixation offset disinhibited the comparatively weaker signals from the target within the VFD. Overall, these findings suggest that similar to humans, residual vision in destriated monkeys is often too weak to result in explicit responses to blind field stimulation. As a result, an external signal and/or release from contralateral inhibition may be necessary to demonstrate the spared capacities. These conclusions are consistent with human evidence showing that patients could accurately localize visual stimuli with eye movements within their blind field, yet when they were asked to indicate the presence of stimuli in their blind field during perimetry, they were unable to do so and appeared blind. When asked to guess the location of the stimuli in the presence of a signal to indicate target onset, they made accurate eye movements to the targets. So, it seems that under certain conditions, residual vision in man and in monkey can be remarkably similar. Particularly, both monkey and humans

without striate cortex seem unable to report visual sensations despite being able to localize targets under forced choice (FC) conditions.

Recent Human Evidence

The first clear evidence that visual stimuli presented within VFDs could influence eye movements was reported in four patients who were capable of looking in the direction of a light flash in their VFD despite remaining unaware of the visual event. In contrast, patients with VFDs caused by retinal lesions did not show any association between target position and eye position.

At about the same time as this observation was published, Weiskrantz and colleagues were beginning to investigate the case of DB, a particularly interesting blindsight case. The term 'blindsight' was coined by Larry Weiskrantz for the title of a seminar presentation!

There were two crucial aspects of DB's case which acted as a catalyst in the research into residual vision following striate damage. The first was that his VFD was caused by surgical removal of a small tumor in the calcarine sulcus. So there were notes detailing the lesion dimensions, including removal of the major portion of the calcarine cortex (in which striate cortex is situated) on the medial surface of the right hemisphere. Following surgery, DB's symptoms (debilitating migraines) were much relieved, but, the majority of his left visual field was blind when tested by clinical perimetry (Figure 2). The special interest in DB's case came when it was noticed that DB appeared to be able to locate some objects within his blind field despite denying any conscious visual experience. The second crucial aspect of DB's case was the decision to test him using the same FC methodology as used for the previous animal experiments. Rather than reliance on the usual clinical methods, which required DB's subjective experience of his vision, for example perimetry, the use of FC methods which 'forced' DB to make decisions about visual stimuli presented within his VFD revealed a range of fascinating visual abilities.

The adoption of FC methodologies previously used in animal experiments proved crucial in uncovering DB's residual visual abilities. For example, in a temporal two alternative FC paradigm (Figure 3), DB was required to choose, if necessary by guessing, in which of the two time intervals a visual stimulus was presented within his VFD. Although subjectively he did not experience any visual presentations in either time interval he was 'forced' to make a choice. Over time, with sufficient numbers of trials it was possible to ascertain whether his 'guesses' were statistically different to performance on the basis of chance.

Extensive testing revealed DB's ability to successfully detect and localize stimuli within his blind field by pointing, or less accurately by eye movements. He was able to discriminate an orientation difference of 10 at a position of 45 from the fixation point. He could discriminate moving from stationary stimuli and was still able to detect stimuli which appeared gradually rather than with a sudden onset. Although DB's ability to discriminate between an 'X' versus 'O' initially appeared to suggest rudimentary form discrimination, further testing showed that what appeared to be an ability to discriminate form was actually based on orientation discrimination.

Figure 2 DB's field defect, a left hemianopia, measured using a Humphreys perimeter, 30-2 Full Threshold program. Locations marked at '<0' represent locations where the brightest stimulus (10 000 apostilbs) did not elicit a response. Reproduced from Trevethan et al. (2007) Cognition.

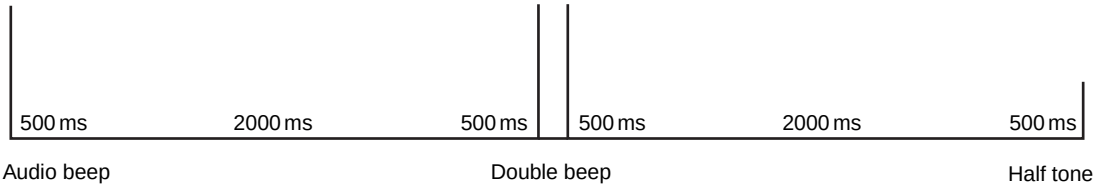


Figure 3 Example of temporal two-alternative-FC paradigm in which stimulus was presented in the second time interval. The patient fixates the fixation cross to ensure that the stimulus is presented entirely within the blind visual field. Audio beeps signal the start and end of each time interval. Usually fixation is continually monitored using a video camera or eyetracker.

This example highlights some of the complexities of testing for blindsight. When he was tested with stimuli with reduced orientation cues for example, triangle versus 'X,' performance deteriorated to chance levels. He was also unable to make a 'matching' (same/different) judgment when both stimuli were presented within his blind field despite his ability to make the judgment successfully when one stimulus was presented within his blind field and one within sighted field and despite his ability to distinguish between the stimulus in his blind field when they were presented singly. Overall, the profile of DB's residual abilities was broadly consistent to that reported in monkeys without striate cortex, that is, detection and localization of certain stimuli in the absence of clear form discrimination or recognition.

Throughout testing, in response to certain stimuli, for example, stimuli with an abrupt onset, DB reported some experiences of subjective awareness, such as, 'something comes in from the screen.' This subjective awareness was not experienced as 'seeing' but as a 'feeling' that something was presented. In order to investigate DB's subjective experience he was asked to report 'aware' or 'unaware' after each experimental trial, called the 'commentary key paradigm.' For example, in a temporal two alternative FC paradigm, DB would be informed that a target would be presented in either the first or second time interval. At the end of each trial he reported in which interval he guessed the target was presented and whether he was aware or unaware, for example, 'first, unaware.' Following this approach, Weiskrantz defined two types or modes of blindsight performance. Type II performance – detection with some awareness, in the absence of 'seeing' and Type I performance – detection in the absence of acknowledged awareness. The verbal commentaries revealed that DB was more likely to report subjective experience of the stimulus when it had an abrupt onset. However, this experience was usually nonveridical and his performance actually tended to improve when he reported that he was not subjectively aware of stimuli. Interestingly, his performance was often best at times when he was tired and felt that he was not performing well.

Recent Investigations into Blindsight in DB

More recently, DB was tested after a gap of 17 years. In addition to confirmation of his original abilities, he also reported his subjective experience of conscious visual negative afterimages in response to a range of stimuli which he was unaware of at the time of presentation, that is, unconscious stimuli producing conscious afterimages. The complementary colors and contrasts of the negative afterimages match the shapes and contrast of the original stimuli and obey Emmert's law (i.e., change size directly with projected distance). Event-related potential analysis suggested a strong anterior, left focus for the blind field presentations compared to an intact field posterior focus. The differential patterns of activity were not associated with hemispheric differences

per se as they were not found in an age-matched control participant. Frontal activation has been associated with stimuli which provoked conscious awareness in another extensively studied blindsight case, GY. In the case of GY, a functional magnetic resonance imaging investigation revealed predominantly right frontal activation associated with subjective awareness of certain stimuli. Frontal activation has also been implicated in comparisons of conscious and unconscious events in unilateral neglect and change blindness. Investigations of visual awareness in blindsight support ongoing research suggesting that visual awareness comprises and requires further processing in addition to the posterior input stage.

Further recent investigations with DB have revealed unexpected and fascinating aspects to his blindsight abilities. DB has demonstrated blindsight detection that can be superior to normal sighted vision, although still in the absence of conscious awareness. He was able to reliably detect extremely low-contrast, static grating stimuli within his blind field that he was unable to detect reliably within his sighted field. A group of age-matched control participants with normal vision were also unable to reliably detect these. DB also demonstrated form discrimination within his blind field. He reliably identified low contrast outline shapes, judging whether stimuli were the same or different and identified complex images (photographs). Although only a single case, who may well not reflect the blindsight abilities of many cases, the dramatic improvement in DB's blindsight abilities is striking. The apparent improvement in his abilities is particularly interesting as although he has participated in considerable amounts of experimental testing over the years, he has not taken part in a specific visual rehabilitation program.

Implicit Processing

Pupillometry

Psychophysical techniques have been called the 'heroic method' of testing for blindsight because of the reliance on considerable effort, cooperation and time on the part of both the participant and the experimenter. Although the results of psychophysical investigations have brought great advances in the investigation of blindsight, the value of indirect and objective testing methods is clear. Pupillometry is the measurement of minute fluctuations in pupil diameter in response to a stimulus. It is an objective method which can be used in conjunction with subjective methods which may also offer greater insight into the mechanisms mediating blindsight. The retina, optic nerves, optic chiasm and optic tracts are all shared by both visual processing and pupillary control. Consequently damage to these areas affects both vision and pupil responses, enabling the use of pupil responses as an objective indicator of visual pathway function. It has been demonstrated that pupil responses to sinusoidal gratings resemble contrast sensitivity functions for foveal (central) and peripheral presentations. This suggests that pupil responses could be used as an objective measure of visual acuity. Measurable blindfield pupil grating responses (PGRs) have been reported in GY which were similar to those measured in two monkeys without striate cortex. The blindfield PGRs suggested a narrowly tuned channel for spatial vision with a peak sensitivity at 1 cycle per degree and a cutoff around 7–8 cycles per degree. There was close correspondence between the residual spatial channel in GY, measured objectively by pupillometry, and the psychophysically determined data (Figure 4). Blind field pupil responses to gratings of equal space-averaged luminance (i.e., no difference in illumination) and colored stimuli in GY have been demonstrated in the absence of reported awareness. The finding that pupil responses can be measured in the absence of acknowledged awareness raises possibilities for the further development of pupillometry as a potential screening technique for blindsight. Measurement of reliable pupil responses in DB as well as in a group of blindsight cases (unpublished data) supports this potential development. In terms of the mechanisms mediating pupil responses, the narrowing of the response profile following a cortical lesion can be viewed as direct evidence of cortical involvement in the generation of stimulus specific pupil responses.

Other Examples of Implicit Processing

Implicit processing within a VFD occurs when a stimulus presented in a blind area of visual field

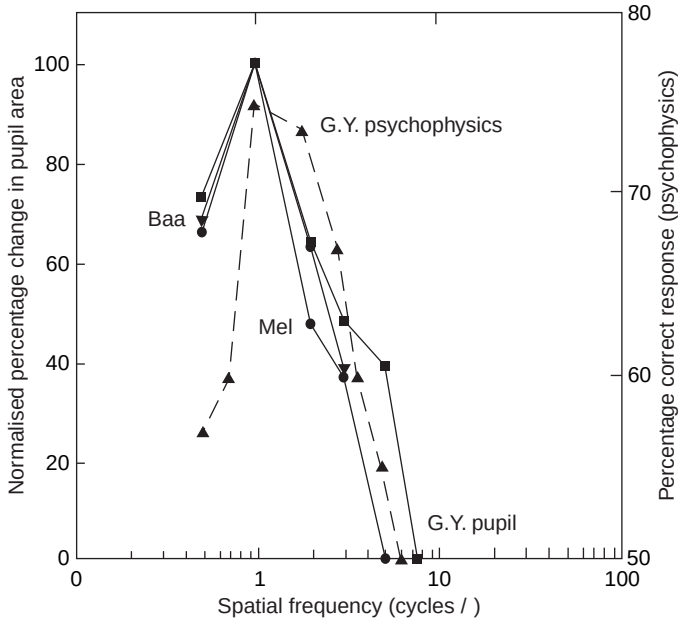


Figure 4 Psychophysical and pupillometric data for blind field stimulus presentation in a hemianope (GY) and the pupil responses in two Monkeys with unilateral striate lesions. Reproduced from Weiskrantz et al. (1998) Brain.

can be shown to exert an effect on the response of the observer despite their failure to 'see' or be aware of it. There are several different methods of investigating implicit processing including, 'completion' and the 'redundant target effect.' The 'completion' phenomenon occurs when a patient with a hemianopia is shown a stimulus, such as an outline circle, with its center in line with the fixation point (so half of the circle is within their blind field) yet they report seeing a full circle. If the half-circle within the blind field was shown alone, no stimulus was reported. In the 'redundant target effect,' simultaneous or prior presentation of an unseen stimulus influences reaction times to a seen target.

How Common is Blindsight?

Initially, it seemed that across a random sample of patients with VFDs resulting from cortical damage, blindsight would only be found in a minority. An early study reported evidence of residual function (responses to movement and localization by pointing or eye movements) in 5 out of 25 patients. One of these patients was GY, who is another extensively studied blindsight case.

There are several reasons why it was thought that evidence of blindsight would be unusual or difficult to clearly demonstrate. First, lesions affecting occipital cortex are typically variable and as V1 is buried deep in the medial aspect of the brain, lesions restricted to V1 alone (rather than including visual association areas) are likely to be fairly uncommon. The age at which damage occurs may also influence the likelihood of demonstrating blindsight as this has already been shown in animal research. It is also important to consider that many of the ways of testing for blindsight can seem strange! Asking someone to repeatedly guess about something that they cannot consciously see can be very frustrating and tiring and not all participants are willing and able to do this. Another absolutely crucial consideration is the choice of stimulus parameters as parameters that may be suitable for testing normal vision may not apply to the blind field. A more recent larger-scale investigation found eight out of ten cases tested in a psychophysical investigation of spatial frequency sensitivity showed evidence for blindsight. The results of an ongoing study into visual training following cortical blindness has demonstrated evidence for blindsight in 20 out of 23 cases tested to date.

The Importance of Specific Stimulus Parameters

Closely related to the issue of incidence of blindsight is the issue of stimulus parameters. Systematic investigation into the factors that affected visual sensitivity in the case of GY began to reveal the importance of stimulus parameters in determining whether or not blindsight may be demonstrated. Despite previous demonstrations of blindsight in GY, when he was retested by a different group of researchers, no evidence of blindsight was reported. It later emerged that the choice of stimulus parameters for the experiments was crucial. In the negative study, the diameter of the visual stimulus was 8.4 , whereas an increase in size to 13.2 resulted in near perfect performance. Similar findings were also found with the temporal characteristics of the stimuli as well as stimulus contrast (e.g., chance levels of detection at a contrast of 15% compared to 98% correct performance at a contrast of 84%). These findings suggested that in GY, these changes to stimulus parameters resulted in a dramatic improvement in performance. These results raise important questions about the incidence of blindsight and detection/screening methods as they clearly demonstrate how such abilities could be overlooked during testing. As mentioned previously, a more recent, larger-scale, systematic investigation found evidence for blindsight in 20 out of 23 cases tested. In these cases, blindsight was characterized by sensitivity to a narrow-range of low spatial frequencies (<4 cycles per degree) and temporal frequencies between 5 and 20 Hz. Stimulus size and contrast were also important for successful detection.

Criticism and Controls – Does Blindsight Exist and is it just Normal Degraded Vision?

From the outset, once details of blindsight within clinically (perimetrically) blind scotoma were reported, a number of criticisms and questions about experimental controls were raised and continue to be discussed. Indeed, some researchers concluded that an adequate case for blindsight had not been made, and some still resolutely hold this position. In the light of continued research, in part generated in response to criticism, this view now seems extreme. However, a number of important methodological issues require discussion for a full understanding of the research evidence for blindsight. At times, a range of criticisms has been leveled against research into blindsight, both in general and in the case of specific experimental findings. In terms of experimental methodology, artifacts as a result of poor fixation and stray light either from within the eye (intraocular) or from visual stimuli (extraocular) require consideration. In terms of the theoretical explanations for experimental findings a number of issues have been highlighted:

Is blindsight simply degraded/near threshold normal vision?
Is blindsight only the result of differences in decision criteria?
Is blindsight mediated by residual 'islands of V1'?

A number of control methods have been used by experimenters which include fixation monitoring, stimuli of brief duration, use of the optic disk as a control presentation position, use of various luminance masking methods, and comparison with patients with pregeniculate lesions.

Experimental Methodology

One of the first criticisms raised in discussions of blindsight is the possibility that the patient had faulty eye fixation and/or made inadvertent, unmonitored saccades toward the visual target. If a patient moves his/her eyes, visual performance loses its scientific interest as intact areas of visual field are used. Clearly, rigorous controls are required to ensure accurate fixation. Apart from the early investigations, eye fixation is usually controlled throughout testing by an infrared camera. Stimuli presented for brief durations shorter than the saccadic onset latency (minimum time needed to initiate an eye movement to a visual target) can also rule out the possibility of eye movements accounting for performance.

Clearly, if blindsight is shown to be attributed to stray light scattering into intact areas of visual field or within surfaces of the eye, the results of blindsight experiments would not reveal anything novel about visual processing and would not be of

particular scientific interest. For this reason, the question of whether discriminable light energy has been scattered into intact parts of visual field or within surfaces of the eye is extremely important. A number of experiments have been carried out to investigate the potential effect of light scatter on performance. Taken as a whole, the implications of the investigations into light scatter in blindsight suggest light scatter can be successfully controlled experimentally and cannot be used as an adequate explanation for many cases of residual visual processing following striate cortex damage.

Is Blindsight Mediated by Residual 'Islands of V1'?

It follows that, if blindsight is mediated by residual 'islands' or 'tags' of striate cortex, the study of blindsight would be limited to the study of degraded vision and would not advance knowledge of visual function or consciousness greatly. This point has been raised by some researchers and there is experimental evidence suggesting that in some cases, visual abilities following striate cortex damage may be mediated by some residual islands of striate cortex. More recently it has been acknowledged that variation between cases is likely and some cases of blindsight may be mediated by such remnants of functioning striate cortex. Despite this caveat, there are a number of reasons why this explanation is unlikely to apply to all cases of blindsight. The idea that blindsight abilities are mediated exclusively by residual 'islands' of remaining striate cortex cannot apply to the considerable body of animal evidence in which the completeness of the lesion and the absence of striate cortex can be confirmed as well as the cases in which the entire hemisphere was removed. The role of subcortical mechanisms was also underlined by evidence that, in two monkeys, following unilateral striate cortex lesions, the ability to detect and saccade toward light targets gradually recovered unless there was an additional superior colliculus lesion. Although it is not appropriate to assume a complete 'mapping' of the monkey visual system onto the human visual system, the close similarities between these two systems, which have been demonstrated anatomically and behaviorally, suggests the applicability of the animal evidence in relation to investigations of blindsight in humans.

Evidence from patients who have undergone complete removal of all cortex from one hemisphere (hemispherectomy), usually as treatment for severe epilepsy, is very relevant as it can help to determine whether subcortical pathways alone can sustain blindsight in the absence of cortical input. Evidence for successful localization, pattern discrimination and motion detection has been reported. Some researchers have presented evidence suggesting that light scatter may have been a factor in some of these experiments. Other evidence, reporting interactions between the intact and impaired hemifield cannot be explained on the basis of light scatter and instead point toward potential subcortical mediation in some cases.

The range of blindsight abilities demonstrated in the case of GY, for example, his ability to follow the path of motion of a small spot moved through a range of trajectories throughout different areas of his VFD, cannot be explained based on small islands of vision within the damaged striate cortex (these capabilities would require a considerable number of islands!) There is also substantial anatomical evidence from a high-resolution computed tomography scan, magnetic resonance imaging (MRI) scan as well as functional imaging (PET and functional magnetic resonance imaging (fMRI)) which do not show any residual 'tags' of striate cortex, only an area of spared tissue situated at the occipital pole which is consistent with GY's area of macular sparing shown by perimetry. Functional imaging data from another frequently tested case, FS, suggests that blindsight is not dependent on islands of preserved tissue in striate cortex. When a large, flickering stimulus was presented to assess visual responsiveness in V1, activation was found in the contralesional (opposite side to lesion) visual cortex and ipsilesional (same side as lesion) extrastriate cortex. Crucially, no stimulus-related MRI changes were shown in striate cortex on the side of FS's brain injury.

Decision Criterion

Blindsight can be demonstrated in the form of a dissociation between visual performance in two different paradigms/tasks, namely clinical perimetry and FC tasks. In humans, the apparent discrepancy between an area of clinically blind visual field and the ability to make some form of

visual discrimination was only revealed by the implementation of 'animal type' FC methodologies. The basis of clinical perimetry is a 'yes-no' (yn) task in which one of two possible stimuli (target or blank) is presented on each trial and the participant's task is to judge which one was presented. This task allows the participant the freedom to say 'blank' or 'no stimulus' on every trial presentation if they wish to do so. Consistently replying with 'no' is what one would expect if they are subjectively unaware of visual stimuli throughout the task. In an m Alternative FC task (mAFC), each of m different stimuli is presented at every trial and the participant has to judge which of m intervals contained a specified stimulus, either in a temporal or spatial (i.e., localization) interval. This paradigm ensures that the participant is effectively forced to make a judgment, for example, between two temporal intervals, as the option to judge 'no stimulus' simply does not exist. This facilitates the revelation of above-chance detection or discrimination performance where it exists, in the absence of subjective awareness (generally required for yn task). Blindsight is characterized by the dissociation in response between the two paradigms, for example, performance levels of $c.$ 0% detection in yn paradigm compared to >90% correct discrimination in the FC task (despite the denial of subjective awareness throughout testing). One of the criticisms of blindsight theory is that the apparent dissociations between yn responding in perimetry and FC performance could result from the use of different decision/response criteria in the two tasks. The implication of this assertion is that the distinction between blindsight and normal, near-threshold vision would not be clear, which implies that the study of blindsight would not add anything to the understanding of mechanisms of visual awareness that could not be drawn from the investigation of normal participants operating at the lower limits of their vision.

According to signal detection theory (SDT) the judgment of a participant in a detection or discrimination task depends not only on his/her sensitivity (d') but also on his/her response criterion/bias (the tendency to select one or other of the stimuli, irrespective of sensitivity). In SDT, sensitivity is calculated independently of response bias and vice versa. However, in the majority of blindsight research percent correct is used to denote performance which, it can be argued, represents performance accurately only in the absence of response bias. In contrast, 2AFC tasks are criterion free as any bias reflects a bias to one or other interval rather than to one or other stimulus. This raises the potential issue of whether the dissociation reported in blindsight between perimetry and FC tasks is due to a difference in response criteria between the two tasks. This has been tested directly in GY during yn and FC detection of static and moving stimuli. GY's response criterion differed significantly between yn and FC responding, and the difference was sufficient to result in a blindsight-type dissociation with bias-sensitive measures of performance. When measured independently of bias, GY's sensitivity to static targets was greater in the FC compared to the yn task (in contrast to normal control participants), but GY's sensitivity to moving targets did not differ. These results suggested that differences in response criterion could account for dissociations between yn and FC detection of motion stimuli, but not for static target presentations. This may explain the trend for blindsight cases to report increased awareness in response to motion stimuli. Importantly, these results also suggest that blindsight is not qualitatively the same as normal, near-threshold vision and that the neural mechanisms for pattern and motion-detection in blindsight may differ.

The question of decision criterion/response bias has also been addressed by altering the proportion of stimuli to blank trials in perimetry. Despite variations in response criterion, the patient's ability to detect stimuli remained essentially impervious to such variations, again suggesting that the discrepancy between subjective awareness and detection ability cannot be attributed to a difference in response criterion.

Is Blindsight Normal, Degraded Vision?

One of the objections to blindsight as a phenomenon is the argument that it is qualitatively the same as normal vision but quantitatively weaker. This issue is important in relation to the mechanisms mediating blindsight as well as potentially providing an insight into the mechanisms associated

with visual awareness. Participants have been reported to localize very briefly presented targets by apparently guessing as they reported being unaware of the stimuli. It has also been reported that when participants are asked to report/guess the position of a mixture of subjectively visible or invisible targets (choice of four locations), they accurately report the position of the targets despite remaining subjectively unaware of it and rating their confidence as poor. Despite these reports of localization in the absence of awareness in participants with intact vision, these findings have been difficult to replicate. Other studies have reported high correlations between localization and confidence. Although the investigation of perception in the apparent absence of awareness in participants with intact vision may have a role in blindsight research, the claim that blindsight is normal, degraded vision does not appear to be supported.

What is Mediating Blindsight?

The geniculostriate pathway (retina – dorsal lateral geniculate nucleus – primary visual cortex), the pathway disrupted by damage to primary visual cortex appears to make the largest contribution to conscious vision. To date, around ten different pathways from the eye to different regions of the brain have been discovered (Figure 5), providing a number of possible alternative routes for the processing of visual information in the absence of the geniculostriate pathway. A number of neuroimaging studies have been carried out, some highlighting the role of the superior colliculus in mediating the unconscious (Type I) visual capacity. In addition to the potential role of other visual pathways, there is also evidence for the contribution of primary visual cortex from the intact hemisphere is some cases. The potential role of extrastriate cortex on the same side of the brain as the lesion has also been highlighted. Recently, the use of transcranial magnetic stimulation to simulate blindsight in observers with normal vision by temporarily disrupting visual cortical electrical activity has provided further insights into the processes underlying unconscious visual perception in blindsight.

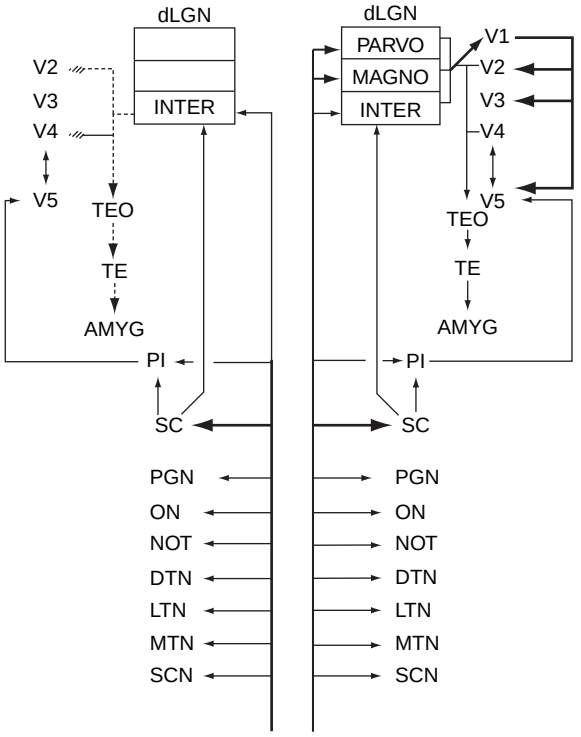


Figure 5 Schematic diagram of the ten established pathways from the eye to their retinorecipient targets in the brain. Only some of the forward projections from there are shown. The right side of the diagram shows the normal arrangement, and the left side the effect of removing striate cortex, V1. Thicker lines and arrows indicate heavier projections. For simplicity the many onward cortical pathways from V2, V3, V4, and so on are not shown, except those destined for the amygdala. Note especially on the degenerated side that the visual input to the ventral processing stream is severely impoverished but that the pregeniculate nucleus has expanded whereas the olivary nucleus has contracted. Labeling from bottom upward: SCN, suprachiasmatic nucleus; MTN, LTN, and DTN, medial, lateral, and dorsal terminal accessory optic nuclei; NOT, nucleus of the optic tract; PGN, pregeniculate nucleus; ON, olivary nucleus; SC, superior colliculus; PI, inferior pulvinar; dLGN, dorsal lateral geniculate nucleus. Reproduced from Cowey A (2004) The 30th Sir Frederick Bartlett lecture. Fact, artefact, and myth about blindsight. Quarterly Journal of Experimental Psychology, A 57(4): 577–609.

What Does Blindsight Tell us about Conscious Awareness?

Blindsight has highlighted that although primary visual cortex is important for conscious visual awareness, consciousness itself cannot be localized

to one brain area. Neuroimaging evidence highlights the likely interaction between several different brain regions, particularly several frontal areas. fMRI has been used to compare brain activity in GY when performing visual discriminations with and without conscious awareness. Comparison between 'aware' and 'unaware' modes of processing revealed that rather than an isolated 'center' for visual awareness there was a shift in neuronal activity from cortical areas, particularly dorsolateral prefrontal areas and extrastriate area 18 (both hemispheres) when GY was aware of the stimuli; to subcortical activity including activation of the superior colliculus and right medial and orbital areas when he was unaware of the stimulus. So, the areas that appear to be differentially involved in visual awareness are located far from the classic areas of visual cortex, supporting the view that visual consciousness involves frontal sites of activity. Imaging evidence from participants with normal vision reveals activation associated with visual awareness in multiple extrastriate ventral, parietal, and prefrontal cortical areas. There is also interest in the potential importance of connectivity between regions involved in visual awareness.

Visual Rehabilitation and Blindsight

As already described, early expectations for recovery of vision following V1 damage in humans were limited. Rehabilitation strategies have focused on compensatory techniques rather than attempting the restoration of vision within the blind field: for example, use of prisms in spectacles in order to superimpose a limited part of blind field image onto the intact field. Another way that patients can benefit from saccadic training involves learning to scan more efficiently and regularly into the blind field. This technique has been used on computer screens as well as large scale patterns presented in the immediate environment and has shown some success in a number of cases. Following the success of saccadic training in nonhuman primates, extensive research has been carried out with the aim of shrinking the field defect by training patients to saccade to light stimuli presented on the sighted/blind field boundaries. Repeated stimulation of such boundaries using light targets (without the saccadic task) has also been shown to result in improved visual sensitivity. Nevertheless, intrinsic to all studies that rely on stimulation of sighted/blind field borders, there are some methodological concerns that small gaze shifts or eccentric fixation may contaminate some of the data, exaggerating the extent of the recovery. More recent studies have shown that visual stimulation deep within the field defect, using stimuli which are optimally configured to elicit blindsight performance (low spatial and high temporal frequency structures), can lead to significant increases in visual sensitivity. In some cases after months of stimulation, blindsight performance may be manifest. The rate of recovery increases if positive feedback is provided after each successful detection. Research is continuing into these recent, exciting developments. It will be interesting to continue to explore which factors are important in facilitating this recovery of function.

It is thought that the recovery of function following repeated stimulation may be mediated via two possible channels which may not be mutually exclusive. The signals within surviving neurones after brain injury may be strengthened and therefore their response rises to above threshold levels for further processing in other brain regions. Alternatively the signals along alternative pathways (outlined above) may be utilized, leading to successful detection. Further research is needed to establish which one (or both) of the above-postulated mechanisms is likely to mediate the improvement.

Blindsight is an area which appears to attract strong views and opinions. It is an area where theories and practices of science are proposed, refuted and modified. The focus of much of the human research into blindsight on single case studies or small number of cases provides important examples of the possibilities, but also raises questions about the extent to which these findings can be generalized. There is much more to be discovered, confirmed, disproved and understood fully. It is hoped that the reader will wish to be critical, curious and energetic in discovering more about blindsight!

See also: The Neural Basis of Perceptual Awareness; Perception: Subliminal and Implicit.

Suggested Readings

Azzopardi P and Cowey A (1998) Blindsight and visual awareness. Consciousness and Cognition 7(3): 292–311.

Campion J, Latto R, and Smith YM (1983) Is blindsight an effect of scattered light, spared cortex and near-threshold vision. Behavioural Brain Science 6: 423–428.

Cowey A (2004) The 30th Sir Frederick Bartlett lecture. Fact, artefact, and myth about blindsight. Quarterly Journal of Experimental Psychology, A 57(4): 577–609.

Cowey A, Stoerig P, and Le Mare C (1998) Effects of unseen stimuli on reaction times to seen stimuli in monkeys with blindsight. Consciousness and Cognition 7(3): 312–323.

de Gelder B, Vroomen J, Pourtois G, and Weiskrantz L (1999) Non-conscious recognition of affect in the absence of striate cortex. Neuroreport 10(18): 3759–3763.

Fendrich R, Wessinger CM, and Gazzaniga M (1992) Residual vision in a scotoma: Implications for blindsight. Science 258(5087): 1489–1491.

Ro T, Shelton D, Lee OL, and Chang E (2004) Extrageniculate mediation of unconscious vision in transcranial magnetic stimulation-induced Blindsight. Proceedings of the National Academy of Sciences USA 101(26): 9933–9935.

Sahraie A (2007) Induced visual sensitivity changes in chronic hemianopia. Current Opinion in Neurology 20: 661–666.

Sahraie A, Weiskrantz L, Barbur JL, et al. (1997) Pattern of neuronal activity associated with conscious and unconscious processing of visual signals. Proceedings of the National Academy of Sciences of the United States of America 94(17): 9406–9411.

Sanders M, Warrington EK, Marshall J, and Weiskrantz L (1974) ''Blindsight'': Vision in a field defect. Lancet April 20: 707–708.

Stoerig P and Cowey A (1997) Blindsight in man and monkey. Brain 120(3): 535–559.

Weiskrantz L (1986) Blindsight: A Case Study and Implications. Oxford: Oxford University Press.

Weiskrantz L (1997) Consciousness Lost and Found. Oxford: Oxford University Press.

Biographical Sketch

Ceri T Trevethan graduated with a first class honours MA degree in psychology from the University of Aberdeen in 2000. Ceri subsequently obtained a PhD in psychology in 2004, investigating the spatial and temporal characteristics of blindsight performance. During this time she also worked as a research assistant with Arash Sahraie on a project investigating rehabilitation of visual deficits following brain injury. This was followed by a 2-year BBSRC funded postdoctoral fellowship. Ceri is currently a trainee clinical psychologist with National Health Service, Grampian and an honorary research fellow at athe University of Aberdeen.

Arash Sahraie obtained a PhD in optics and vision sciences from City University, London in 1993 under supervision of Prof. John Barbur. He then continued with 5-year postdoctoral assistantship to Profs. John Barbur and Larry Weiskrantz before moving to University of Aberdeen, Scotland in 1998. He is currently professor in vision sciences at College of Life Sciences and Medicine, University of Aberdeen. Arash Sahraie has a long-standing research interest in neuropsychology of vision and rehabilitation of visual deficits following brain injury.

Brain Basis of Voluntary Control

S Pockett, University of Auckland, Auckland, New Zealand

Introduction

Philosophical arguments notwithstanding, an ordinary dictionary is usually a pretty good indicator of how words are commonly used. *The Concise Oxford Dictionary* defines a voluntary movement as one that is controlled by the will.

Such a definition suggests that an article on the brain basis of voluntary control should discuss not only the brain basis of movement *per se*, but also (and perhaps more importantly) the brain basis of the will. This makes the whole exercise both more interesting and less straightforward. What is the will, anyway? Does it even have a basis in brain function? Our everyday intuition tends to be that the will may be some sort of dualist entity, capable of acting on the brain, but not itself generated by the brain. More radically, does the will really exist at all? Perhaps will is nothing more than an inference, a mythical will o' the wisp created by conscious selves desperate to believe that they and not their brains control their bodies.

Returning to earth, we might profitably adopt the same strategy with regard to the noun will as we did for the adjective voluntary – ask the dictionary. *The Concise Oxford Dictionary* says the noun will means ''faculty by which one decides what to do; fixed desire or intention.'' This definition implies (although to be fair, it does not explicitly require) that the will is a faculty of consciousness. Certainly, thanks largely to Sigmund Freud it does seem possible nowadays to have desires that are unconscious. Perhaps we might even allow the existence of unconscious intentions, though this at least initially seems more like a contradiction in terms. But the faculty by which one decides what to do must surely be a faculty of consciousness. We may not be conscious of all the factors influencing a given decision, but we do consciously make decisions. Don't we?

Maybe not. The first professor of psychology at Harvard, William James, spent a whole chapter of his 1890 magnum opus *The Principles of Psychology* arguing against his contemporary Thomas Huxley's theory that humans do not consciously control their actions, but are basically automata. In contrast, the current professor of psychology at Harvard, Daniel Wegner, has summarized recent empirical evidence on the matter in a book entitled *The Illusion of Conscious Will*. Contrary to our everyday intuitions, what relatively little science has been done in this area tends to show that not only the neural events controlling bodily movements, but arguably even the neural events initiating bodily movements, are inaccessible to consciousness. Experiments show that we are not very good even at knowing whether we caused any given occurrence ourselves, or whether somebody or something else did. But then, of course, given the tenacious nature of everyday intuitions, it is not surprising that philosophers (specifically, all the philosophers represented in a recent multiauthor book called *Does Consciousness Cause Behavior?*) vehemently disagree with Wegner's conclusions on this matter (although not with the empirical evidence on which those conclusions are based). The century-old argument continues. A pessimist might conclude that a dozen decades of hard work have brought us no closer to any definitive understanding of the innocent-looking word 'voluntary.'

But, such pessimism is rarely justified. We have learned some things over the last 120 years. To ease the reader gently into these deep and treacherous waters, the present article first provides a short and relatively uncontroversial account of current knowledge about the functioning of those brain areas known to be involved in decisions, intentions, and movement. It then strikes out strongly toward the center of the issue, describing a number of lines of experimental evidence that suggest that, at least much of the time, the functioning of these brain areas does not give rise to any conscious sensations or experiences at all. Finally, after skirting delicately around the whirlpool of philosophical

controversy whipped up by these experimental results, we retreat to the far shore to consider briefly the possible consequences of our findings for the legal system, which presently regards humans as conscious agents.

The Neuroscience of Voluntary Control

Nearly 150 years ago, the same Huxley whose automaton theory was so vigorously opposed by James famously remarked that "the great end of life is not knowledge but action." The enduring truth of this aphorism is suggested by the fact that a large part of the brain is concerned in one way or another with the production of actions. Figure 1 shows the physical locations in the brain of the areas known to be involved in the initiation and control of voluntary movements. A list of these areas includes the primary motor cortex, the supplementary and presupplementary motor areas (pre-SMAs), the premotor cortex, the frontal eye fields, the cingulate cortex, the posterior parietal cortex, the dorsolateral prefrontal cortex (DLPFC), the basal ganglia, the thalamus, the cerebellum – and of course, much of the spinal cord (not shown).

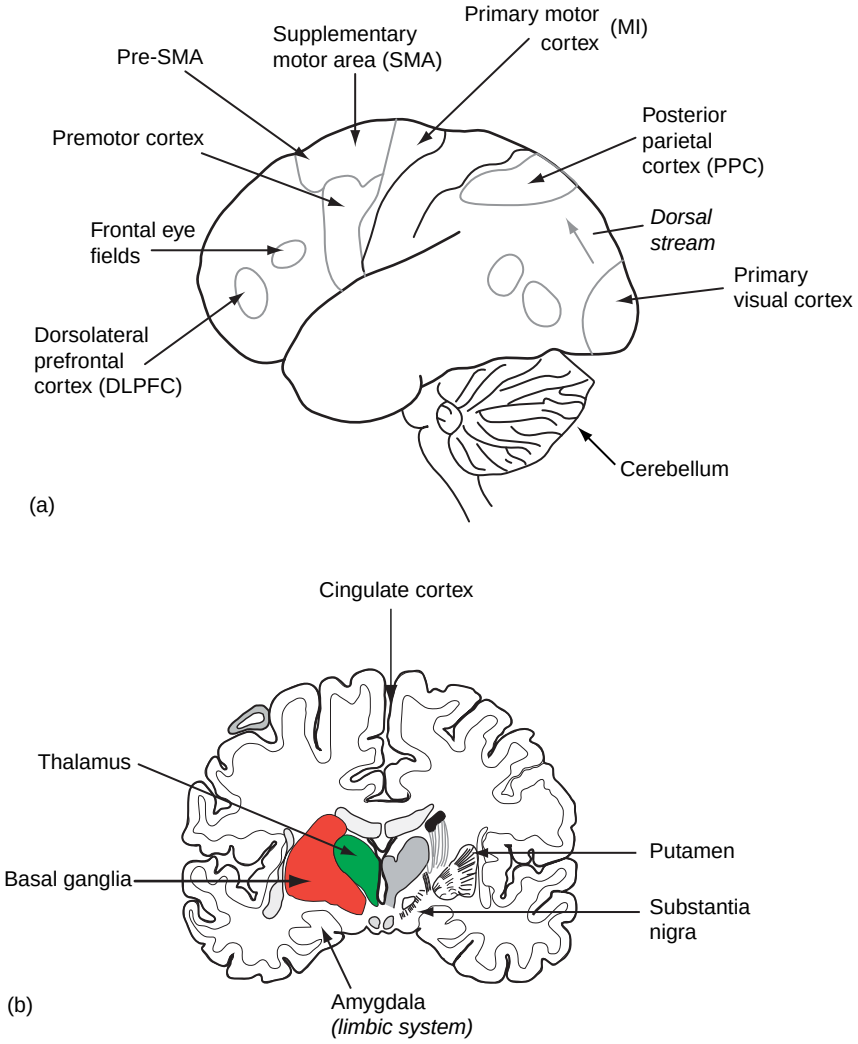


Figure 1 Location of brain areas involved in movement. (a) Surface of left hemisphere. (b) Transverse section of brain in the region of SMA. The putamen and substantia nigra are two of the many components of the basal ganglia. The amygdala is part of the limbic system, which mediates emotion and motivation.

Figure 2 shows a schematic model that attempts to pull together what is currently known about how these brain areas interact to implement voluntary control. At this stage our knowledge about the neuroscience of motor control is far from complete, and so like most (if not all) scientific models, this one must be considered a heuristic rather than a final construction.

According to the model in Figure 2, a willed movement is (not unreasonably) conceptualized as starting with the will. For the moment we can sidestep the whole controversy about the nature and even existence of the will and define that entity operationally as a specific intention.

Neuroscientific evidence suggests that there are two different kinds of intention, subserved by two widely separated areas of brain. Intentions of the first kind are called willed intentions. Willed intentions are abstract, early plans for movement. They specify the goal and type of movement, but not the detail of how the movement will be carried out. Given the general accuracy of the lay notion that the frontal lobes of the brain are involved in thinking, it is not surprising that willed intentions are generated in frontal areas, specifically the DLPFC and its adjacent pre-SMA. One important feature of willed intentions is that they need not necessarily be acted upon – indeed the road to Hell is said to be paved with good ones.

The second kind of intention is called a sensorimotor intention (or motor representation, or stimulus intention, depending on the author). This kind of intention specifies the detail of how an intended movement is to be carried out. Sensorimotor intentions are located not in the frontal areas, but toward the back of the brain in the posterior parietal cortex. Such a location is advantageous because visual input is very important in the construction of specific plans for interacting with the world, and the posterior parietal cortex is directly connected to the dorsal stream of the visual system. (Visual input from the outside

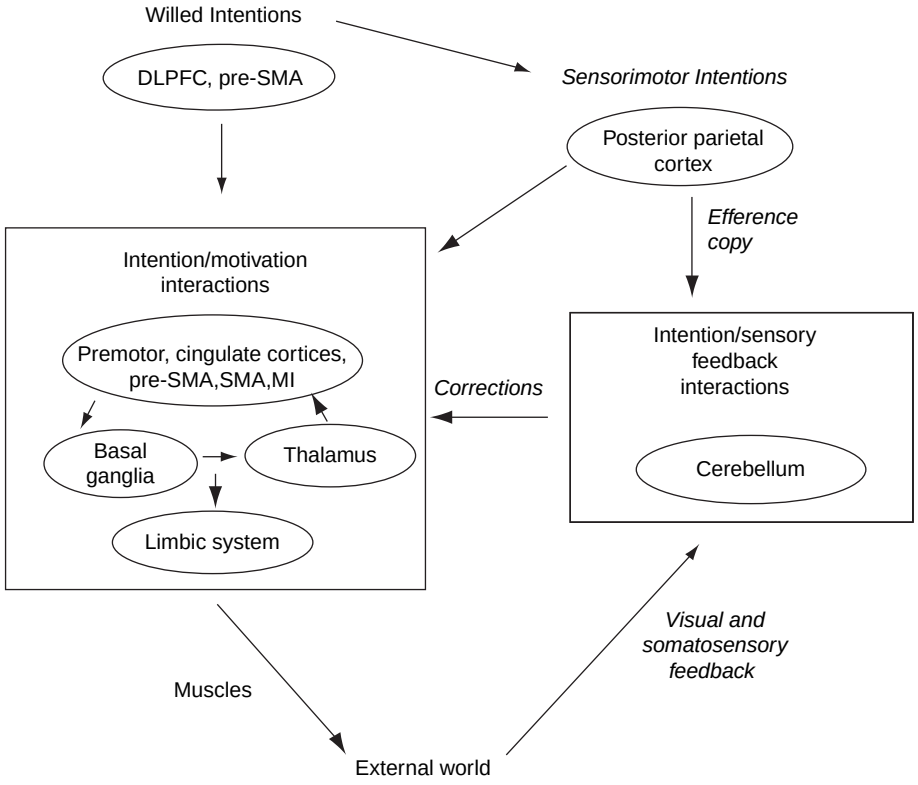


Figure 2 Model of anatomical and functional relationships in the motor system. DLPFC, dorsolateral prefrontal cortex; pre-SMA, presupplementary motor area; SMA, supplementary motor area; MI, primary motor cortex. See Figure 1 for physical locations of these areas.

world enters at the eyes and immediately flows caudally through the large fiber tract of the optic radiation to the primary visual cortex, at the very back of the brain. After traversing through the secondary visual areas slightly rostral to (in front of) the primary visual cortex, visual information then flows forward toward the nose in two streams. The ventral stream is believed by many to culminate in visual perception in the infratemporal area just above the ears. The dorsal stream is not associated with consciousness, leading instead to the sensorimotor intention areas in the posterior parietal cortex.)

Unlike willed intentions, it is not clear whether sensorimotor intentions can be formed and then not acted upon. If sensorimotor intentions can be formed but not acted upon, the initiation of an action must logically be taken as occurring after the formation of its sensorimotor intention. If sensorimotor intentions are necessarily carried out once they are formed, the initiation of action could be considered to occur before the formation of sensorimotor intentions.

This basic uncertainty about the timing of the all-important action-initiation step is reflected by the lack of a box in Figure 2 labeled action initiation. At this stage, so little work has been done on the neuroscience of action initiation that almost nothing is known about its neural basis. It is likely that the basal ganglia are involved, as shown both by (1) functional magnetic resonance imaging (fMRI) studies and (2) the fact that Parkinson's patients often find it difficult to initiate actions. (Probably the primary lesion in Parkinsonism is in the substantia nigra of the basal ganglia) (see Figure 1). However, the SMA has also been suggested as a primary site of action initiation. The truth is that the complex loops sketched in the 'Intention/Motivation Interaction' box in Figure 2 are presently little understood and at this stage no particular brain area or areas can definitely be credited with the function of initiating voluntary actions.

However, whether or not we understand when and where it happens, at some point any given action must be initiated, or launched. The action then needs to be seen through to its completion. This means that the movement must be controlled and if necessary corrected. With very fast, so-called ballistic movements, there is no time for correction during the movement itself, but if the end result of the movement is not as planned, at least the next such movement can be adjusted. Slower movements can be adjusted on the basis of visual and other feedback during the course of the movement.

Much of the function of control and adjustment of voluntary movements is subserved by the cerebellum (see Figure 1). The cerebellum is an evolutionarily ancient structure with a unique anatomical structure. More functions for it are being discovered all the time, but probably its main task is to compare a copy of the original motor instructions from the posterior parietal cortex (a so-called efferent copy) with feedback from the environment, in order to compute error and error correction messages. These error correction messages are then fed back to the posterior parietal cortex and/or forward to the frontal motor areas, the spinal cord, and eventually the muscles.

Consciousness and Voluntary Control

The brain activity described in the 'Introduction' section serves to explain the generation of first general and then specific intentions to act, and the control of movements once they have been initiated. But so far, no mention has been made of where consciousness does or does not fit into this essentially deterministic chain of events. Our everyday intuitions would probably say that the formation of a specific intention to act is conscious. The initiation of a voluntary act or movement would also generally be considered to be conscious. The dictionary definition of 'voluntary' might suggest that the control of voluntary movements is conscious as well. But intuitions and definitions are always trumped by empirical evidence. What does the empirical evidence say?

Control of Movements

It is a truism that many voluntary movements, particularly those which could be called overlearned, are carried out without the engagement of very much at all in the way of consciousness.

A skilled typist can produce a manuscript without any awareness of the multiple complex motor decisions made during the act of typing. A skilled driver may navigate the entire trip from work to home without ever thinking consciously about the act of driving, and without remembering any of the particulars of the journey on arrival. These are not acts that could be described as automatisms, such as sleep-walking, or reactions, such as withdrawal of the hand from a painful stimulus. They are complex sequences of deliberate movements, directed by both internal and external stimuli. When performing such acts, we can become very conscious of what we are doing if something interrupts the flow of events. But even then, if a sudden reaction becomes necessary – a child runs out in front of the car, for example – we act first and become conscious of the stimulus only after the action is completed. We find our foot on the brake pedal and then become aware of the child. Such anecdotal experiences of the unconscious control of ongoing movements have been amply confirmed by experimental data, as described in the article by Marc Jeannerod elsewhere in this volume.

What is the brain basis of this kind of relatively unconscious motor activity? The key to answering this question is to examine the differences between neuroimaging data obtained during the performance of novel actions and neuroimaging data obtained during performance of the same actions after they have been well learned. When learning a new action, we are conscious of every little decision. As the skill is learned, the action becomes less and less conscious. Karl Friston and Richard Frackowiak at University College London and Richard Passingham at Oxford have shown that the brain areas that become less active as a motor skill is learned are the prefrontal cortex and the SMA. This suggests that the prefrontal cortex and the SMA are somehow involved in the production of consciousness. It will be recalled from the discussion in 'The neuroscience of voluntary control' section that the prefrontal cortex is implicated in decision making and the SMA has been suggested as one of the sites involved in action initiation.

In contrast, the brain sites concerned with the unconscious control of voluntary actions are very likely to be the posterior parietal cortex and the cerebellum. As mentioned earlier, these are the sites involved in producing sensorimotor intentions (the specific plans for how a movement will be carried out) and fine-grained motor control, which the article by Jeannerod shows to be largely unconscious. This seems eminently reasonable from a biological point of view. If it were not so, we would not be able to walk and carry on a conversation at the same time. Consciousness has a very limited capacity. It would not do to be conscious of every little calculation involved in keeping one's balance or placing one's feet. We need to be able to think about other concerns while still moving efficiently about the world. Babies and people relearning the arts of movement after a neurological injury have to think about the details of walking, but the rest of us do not.

But surely we do need to have conscious control over the initiation of sequences of motor actions. We have to decide consciously to get into the car and start driving. If all our acts were automatic at that level, we could not count ourselves as conscious selves – could we?

Initiation of Movements

As mentioned earlier, considerable confusion presently surrounds the initiation of voluntary movements. One of the few experiments done so far that has examined the act of initiating a previously willed intention was first carried out a quarter of a century ago by Benjamin Libet. Libet (who died in 2007 at the age of 91) asked his subjects to watch a spot of light rotating on a clock face while they made a series of spontaneously generated finger movements. After each movement, the subject had to report exactly where the spot was on the clock face at the instant they had felt the urge or 'wanting' to move their finger. Libet called this reported instant time W. The clock method of measuring the objective time of a subjective event is now generally called the Libet clock, but it is actually a modification of a general method invented by Wilhelm Wundt almost a century previously. Libet's conceptual breakthrough was to compare the mean W times from groups of 40 movements with the electroencephalography (EEG) event-related potential extracted by back-averaging the EEG immediately preceding those 40 movements. (It was necessary to average 40 movements because

event-related potentials are so small that they are buried in the biological noise and cannot be seen for individual trials. Averaging is a standard technique for pulling small signals out of noise. It works because at any given instant after (or in this case before) the event in question (in this case the movement) random noise is equally likely to have a positive or negative voltage. This means that in the mean of many trials, the noise averages out to close to zero. In contrast, the signal is always the same at any given instant before (or after) the movement, and so averaging does not change its amplitude.) The event-related potential in question is a slow, negative-going waveform that had been discovered 20 years before Libet's experiment by Kornhuber and Deecke, who named it, in German, the Bereitschaftspotential. The Bereitschaftspotential is now generally known by its English name of readiness potential (RP). An example of an RP is shown in Figure 3.

The outcome of Libet's experiment was reported in 1983 in the journal Brain, and has been the subject of vigorous debate ever since. The main finding was that the RP started about 350 ms before time W (see Figure 3). The startling implication of this is that the brain initiates voluntary movements before the subject is conscious of having willed them. In other words, the brain does all the work of starting a movement un- or preconsciously, and keeps consciousness informed of what it has done only as a sort of professional courtesy. On this interpretation, it seems that consciousness has no function when it comes to voluntary movement. If there is such a thing as a will that causes bodily movements, this will cannot be a conscious entity.

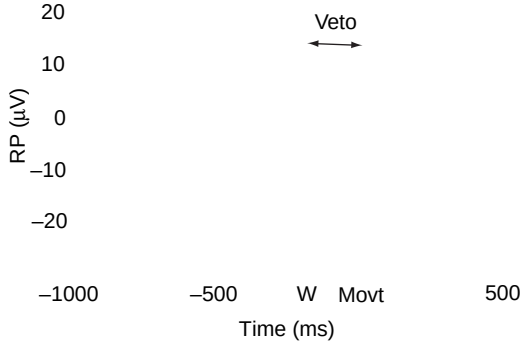


Figure 3 Relationship of Libet's time W and veto window to the EEG RP. RP is evident in average of 50 trials recorded from the vertex, presented with negative voltage shown in the downward direction (in contrast to the earlier convention of presenting negative as up). In Libet's terminology this is a Type II RP, produced by spontaneous rather than preplanned movements. Type II RPs start about 500 ms before the movement.

Libet himself believed that his results did demonstrate the unconscious initiation of voluntary acts. However, he remained convinced that consciousness must have a function. To reconcile these two opposing beliefs, he proposed that consciousness has the opportunity to veto the action initiated by the brain, in the approximately two-tenths of a second between time W and the actual movement (see Figure 3). This idea presently remains hypothetical, in part because it is not possible to study the neural activity associated with a veto by back-averaging off vetoed movements, since a vetoed movement by definition never takes place.

Other workers have been less convinced that Libet's original result does demonstrate the unconscious initiation of voluntary acts. The result itself is not in question – the main experimental finding has now been replicated in a number of different laboratories. RPs do, undeniably, start before time W. But controversy still surrounds the interpretation of this fact. The basic problems are

1. The relationship between RPs and the neural activity leading up to a movement remains to be elucidated. Specifically, it is not clear whether the start of the RP does represent the initiation of the movement, or whether it simply indicates a readiness for the movement to be initiated at some later point in time. There do exist at least two types of expectation-related waveform that resemble RPs, but precede events that are not movements.
2. It is not entirely clear that subjects do actually experience at time W subjective events that could be called urges or wantings. Recent experiments show that reports of W can be influenced by external events that take place after the movement. This suggests that W reports may represent something more akin to

a cognitive reconstruction after the event of what the subject infers must have happened. Perhaps the subject never actually experienced an urge to move, but in an effort to please the experimenter simply indicated that he or she had done so, at a time subjectively perceived as being a little before the movement itself.

Work is actively proceeding on these issues, but at this point it is not far enough advanced for its conclusions to be enshrined in an encyclopedia.

One obvious question in this area that has not so far been adequately addressed concerns the nature of the neural activity that is going on at time W. If there is a particular neural event that always coincides with the subjectively reported urge, it seems a reasonable bet that the urge might be a real subjective event, generated by (or identical with, if you espouse the neural identity theory of consciousness) the neural event. It might seem that it should be fairly simple to ascertain exactly what is happening in the brain at time W. Unfortunately, technical considerations mean that this is far from the case.

The problems with measuring the nature and location of the brain activity occurring at any particular time arise from the characteristics of the three noninvasive techniques that are currently used to measure brain activity in humans. The newest of these noninvasive techniques is fMRI. This measures blood oxygen level dependent (BOLD) signals, which essentially means it measures the increased blood flow that accompanies increased neural activity. The spatial resolution of BOLD is of the order of millimeters, but the temporal resolution is very bad, owing to the variable and indeterminate period of time (on the order of a couple of seconds) necessary for blood flow to a given brain area to increase once that area becomes active. The oldest of the noninvasive techniques is EEG. This measures the voltage difference between chosen points on the scalp. The temporal resolution of EEG can be of the order of microseconds (although millisecond resolution is more usual), but the spatial resolution is of the order of tens of millimeters, simply because the scalp is so far away from the brain. It is often thought that the main problem contributing to the poor spatial resolution of EEG is the smearing action of the skull, but in fact the real difficulty is the point spread function due purely to the distance between the site of waveform generation in the cortex and the measuring electrodes on the scalp. It is not widely appreciated that this distance is 15–20 mm, while the width of cortex generating most EEG waveforms is only 2–3 mm. The third noninvasive technique in our arsenal is magnetoencephalography (MEG). This uses superconducting quantum interference device (SQUID) sensors to measure directly the magnetic component of the electromagnetic fields generated by the brain. Since magnetic fields are unaffected by the skull, it is widely believed that the spatial resolution of MEG is substantially better than that of EEG. In fact it is about the same, because the main smearing influence in any electromagnetic measurement is the distance between the field generation site and the sensors, and this distance is actually slightly increased in MEG because the necessity to cool SQUIDs with liquid helium means that MEG sensors can not be placed directly on the scalp. MEG does however have the advantage that its measurements are reference-free. The requirement to measure EEG signals between recording sites and a reference site is the source of many problems. The spatial resolution of both EEG and MEG can theoretically be increased by mathematical solution of the inverse problem, which allows localization of the source of the electromagnetic fields in the brain. But the inverse problem is notoriously underdetermined, which means that in principle there are an infinite number of solutions that fit any given dataset.

Returning to our question, it can be seen that none of these noninvasive techniques is optimal for determining exactly what brain areas are active at time W. Either the spatial resolution is good but the temporal resolution is too low (fMRI) or the temporal resolution is good but the spatial resolution is too low (EEG and MEG). More precise spatial localization with preserved temporal resolution could potentially be achieved using electrocorticography, which is essentially the recording of EEG data from the surface of the brain instead of from the scalp. This is occasionally done for the clinical purpose of localizing epileptic foci prior to their excision, but to date published data from this

technique do not contain enough detail to show exactly what brain areas are active 150–200 ms before a voluntary movement.

What data are available from the noninvasive techniques suggest that at time W activity is occurring in the SMA and/or the primary motor area (MI). The suggestion that SMA activity might be the neural correlate of an urge to move is supported by the finding that low-level electrical stimulation of the SMA in awake human patients does sometimes elicit verbal reports of an urge to move. However, higher intensity stimulation of the same areas invariably causes actual movement, and so it is possible that downstream activation of the primary motor area by the low-level stimulation might be the real correlate of the reported urges, or even that very small actual movements might be misinterpreted by the patients as urges.

But whatever the eventual outcome here, the neural activity underlying (or at least occurring at the same time as) the urges in Libet's experiments can only be related to the question of when a predetermined movement should be made. This sort of putative urge has nothing to do with the preceding decisions about what movement should be made, or that a movement should be made at all. In the case of Libet's experiments, those decisions were made many hours before any of the particular movements was initiated. The desires (to please the experimenter), choices (to participate in the experiment), and intentions (to follow the experimental instructions) occurred anything up to several weeks before the 'spontaneous' urge to make any particular movement. What is the relationship of consciousness to these earlier events?

The Will (Desires, Choices, and Intentions) – Experienced or Inferred?

Given that it is lamentably common to form a willed intention to do something (one's tax return springs to mind) – but then somehow never to get around to actually doing it – it seems clear that action initiation is not the same thing as willed intention. Libet-style experiments study action initiation, but they do not study the willed intention that preceded it (in this case the intention to do as the experimenter asks and move one finger repeatedly while watching a clock).

Wegner carried out a different set of experiments, aimed at finding out how conscious we are of our willed intentions. His thesis was that the experience of having caused an action is not a direct introspection of any particular brain events, but rather an inference like any other inference of cause and effect. The suggestion is that we think that A causes B if and only if

1. A occurs just before B,
2. A is consistent with B, and
3. there is no other apparent cause of B.

Likewise, we believe we have caused a given event if and only if

i. we think about the event just before it happens,
ii. our thought is consistent with what happens, and
iii. there is no obvious external cause for what happens.

In a test of condition (i), Wegner and colleagues showed that subjects could readily be fooled into thinking they had caused a computer cursor to stop over a particular object (when in fact the experimenter had caused the cursor to stop) if the name of the object was played into the subject's earphones just before the cursor stopped. Actually the subjects were only 56% sure that they had caused the stop both when the trick above was played and when they themselves actually had caused the stop, which again suggests that we are quite bad at introspecting our own intentions. In a test of condition (ii), subjects who viewed the experimenter's gloved hands in the position where their own would normally be could be fooled into thinking they were controlling the gloved hands, provided the gloved hands moved in accordance with a set of instructions played into the subject's earphones. Again, people were not very good at knowing whether or not they were controlling the hands even in the baseline condition. But when the subjects were led to think about what the hands were doing just before they did it, the perception of control increased significantly (although still only to about 3 on a scale from 1 = no control to 7 = complete control).

These and other experiments have led Wegner to propose the model in Figure 4. The suggestion is that actions and thoughts about actions are

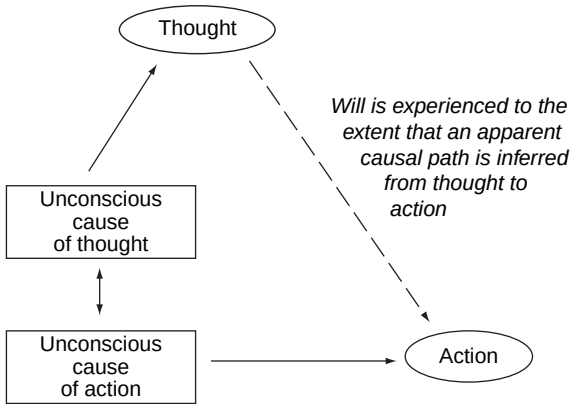


Figure 4 Wegner's model of the generation of voluntary actions. Unconscious events are shown in rectangular boxes. Conscious events are shown in ellipses.

generated by related but separate unconscious processes that act in parallel. The sensation of having caused the action by an effort of will is then generated by a third process, which infers causation of an action by a thought in the same way it might infer causation of tree movement by the wind.

Ironically, if Wegner's model is accurate, it has the contradictory effects of

1. suggesting that conscious intentions are often inferred rather than introspected, and thus that consciousness does not cause actions, and
2. providing at least indirect support for the idea ('Initiation of movements,' section point 2) that Libetian urges to move are also inferred rather than introspected, and thus throwing doubt on Libet's conclusion that consciousness does not cause actions.

This is all very confusing! Perhaps the safest conclusions at this point are (1) it is doubtful whether we really do have conscious access to either our urges or our intentions, and thus (2) it is doubtful whether consciousness plays any vital role at all in voluntary movement.

Philosophical Arguments

Although this is primarily an article about the brain basis of voluntary movement, it is instructive to consider briefly some of the philosophical arguments that have been put forward in opposition to the empirically derived conclusions in the preceding sections – not least because at root, these philosophical complaints constitute arguments against ever accepting any conclusions based on empirical data.

For example, one contributor to the book Does Consciousness Cause Behavior? says that we should not conclude anything about whether or not consciousness causes behavior because we do not yet know enough of the physiological facts. This would be fair enough if he did not himself then conclude that therefore there is no reason to throw out our prescientific intuitions that we do consciously cause our own acts. Another argument put forward in the same book is that Wegner's experiments are irrelevant because they put the subjects in abnormal situations. But all scientific experiments put their subjects in abnormal situations. Other contributors complain that while the experimental data that are available clearly show that some voluntary movements are not caused by consciousness, they do not show that all voluntary movements are not caused by consciousness. Such complaints raise the ancient specter of induction. The problem with induction is that one can never prove by induction (i.e., by making a finite number of experimental observations) that, for example, all swans are white, because there is always the chance that tomorrow one will meet a black swan. (This famous northern-hemisphere-grown example is particularly enjoyed by philosophers who hail from the southern hemisphere, where black swans are endemic). The attentive reader will by now have noted that the problem with all of these arguments is that science in general proceeds exactly by induction and experimentation, and so far science has proven itself a remarkably effective method of learning about the natural world. In fact, scientific experiments are the source of much, if not most, of our current understanding of how things work.

That being said, it is certainly fair to point out that science is never finished, and in the present case more work needs to be done before it is reasonable to conclude even on a heuristic basis that consciousness never has any role in voluntary movement. Before we can make such a sweeping

conclusion, we need to know more about the brain basis of movement initiation. We need to know what causes the transition from willed intention to movement initiation. And we need to know whether or not it is possible in principle accurately to introspect our own urges, decisions, and intentions.

At present, the most we can say for sure is that the role of consciousness in voluntary movement is in question, and that this role is certainly smaller than that previously thought.

Legal Implications

The machinations of the Western legal system might seem remote from the day-to-day activities of brain scientists, but knowledge about the brain basis of the will is actually very important for the law. At present, most legal jurisdictions require the perpetrator of a crime to have consciously intended to do whatever he or she did in order for the act to be regarded as culpable. People are not generally put in jail because of their involvement in wholly accidental occurrences, when they did not consciously even put themselves in a position that could reasonably have been expected to lead to the accident (e.g., get into the car in an intoxicated state in the first place and start driving down the road).

This requirement for conscious intent makes the relationship of consciousness to action vitally important to the legal system. If it were to be accepted as a scientific fact that we never have conscious access to our intentions and/or to whatever neural activity initiates our actions, but simply have to infer these after the fact in the same way that we infer intent on the part of other people, then either the law would have to be changed, or nobody could ever be found guilty of anything.

So the question becomes, how sure are we that we do not have immediate access to our own intentions, or to the decisions that initiate our acts? As mentioned earlier, we are certainly not completely sure. But then no scientific conclusion is ever completely secure. Are we sure enough to recommend that the law be changed?

There are two points at issue here, which have different legal implications and thus should probably be considered separately:

1. Do we have conscious access to our long-term intentions?
2. Do we have conscious access to the events that initiate a given voluntary act?

Conscious Access to Long-Term Intentions?

How sure are we that we do (or do not) have conscious access to our long-term intentions? It is true that Wegner's experiments are of limited scope. But they build on a long tradition of research indicating that introspection of one's motives, intentions, and desires is significantly unreliable. People readily answer questions about why they did things, but as often as not their answers indicate that they are actually inferring rather than experiencing their own motives – indeed inferring them with little more accuracy than they could infer the motives of other people. Certainly, we are sometimes accurately aware of our own intentions and motives – but then we are sometimes accurate about other people's intentions and motives, too. The critical point is that we seem to have little direct introspective access to the thought processes involved in our own evaluations, judgments, and problem solving. We often do not know why we do what we do, or even that we intended to do it.

There may by now be enough data on this to render prudent a removal of the word conscious from the law relating to intent.

Conscious Access to Action Initiation

On our present scientific understanding it is conceivable that we also lack direct introspective access to the initiation of actions. However, less experimental evidence is available on this. More work needs to be done before it can justifiably be concluded that action initiation is not under direct conscious control.

See also: Free Will; Intentionality and Consciousness; Neuroscience of Volition and Action; Perception, Action, and Consciousness.

Suggested Readings

Huxley TH (1874) On the hypothesis that animals are automata, and its history. The Fortnightly Review 16: 555–580. Reprinted in Method and Results: Essays by Thomas H. Huxley. New York: D. Appleton and Company, 1898.

James W (1890) The Principles of Psychology. New York: Henry Holt & Co.

Libet B, Gleason CA, Wright EW, and Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). Brain 106: 623–642.

Nadel L and Sinnott-Armstrong W (in press) Libet, Free Will and Responsibility. Oxford: Oxford University Press.

Nisbett RE and Wilson TDeC (1977) Telling more than we can know: Verbal reports on mental processes. Physiological Review 84(3): 231–259.

Pockett S, Banks WP, and Gallagher S (2006) Does Consciousness Cause Behavior? Cambridge, MA: MIT Press.

Wegner DM (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

Biographical Sketch

Susan Pockett earned her PhD in neurophysiology from the University of Otago (New Zealand) and subsequently worked on synaptic plasticity in the physiology departments of the University of Oslo (Norway), the University of Auckland (New Zealand), University College London (England), the University of New South Wales (Australia), and the University of Manitoba (Canada). In 1994, she changed fields to work on consciousness, and is presently in the Physics Department at the University of Auckland.

Cognitive Theories of Consciousness

V de Gardelle and S Kouider, Département d'Etudes Cognitives, CNRS/EHESS/ENS-DEC, Paris, France

Glossary

Connectionism – Connectionism is a framework in cognitive science, according to which all of the processes achieved by the mind can be modeled by parallel and distributed processing among simple operational units. It is mostly based on the development of artificial neural networks, and it has been traditionally opposed to the position that mental processes are based on symbolic computations.

Functionalism – Functionalism is a doctrine in cognitive science, according to which a mental state is defined by its functional role, rather than by its intrinsic structure and its implementation. In other words, a functional model of the mind (or of an operation that is achieved by the mind) involves mental states that are causally related to sensory inputs and other mental states, and behavior.

Homunculus – Literally, 'little man,' in Latin, that is in the context of cognitive theories of consciousness, a conscious observer, which is at the top of the cognitive system.
A Homunculus is a hypothetical construct that operates (i.e., with no further explanation) the very operation that is supposed to be explained. Thus, a theory that relies on a homunculus at some point is incomplete in the same extend.

Introduction

Consciousness is probably the most privileged topic in psychology. The study of consciousness is considered to be at the origin of the separation between psychology and philosophy during the nineteenth century, as psychologists were motivated by the will to tackle this issue in a scientific way. As such, developing a scientific theory of consciousness has been the Holy Grail of psychology since its earliest days. Today, consciousness continues to be a central topic of interest, extending its interest across almost all disciplines of cognitive science.

Studying consciousness, however, has not always been an acceptable question in psychology. Indeed, the issue of consciousness was totally rejected by the dominant behaviorist school during the first half of the twentieth century. The failure of earlier psychologists, who were stuck in unsolvable debates related to introspection, led the behaviorists to reject consciousness as a plausible scientific issue. Instead, behaviorists sought to bring psychology into the scientific domain by restricting it to objective data and reproducible methods. Even the so-called 'cognitive revolution' that transformed psychology during the second half of the twentieth century did not favor a renewal of interest in consciousness. Indeed, the purpose of the cognitive approach was to reintroduce the notion of internal representations or mental states, not the notion of consciousness per se. Furthermore, an important new assumption was that internal representations were largely unavailable to consciousness. Nonetheless, in this new information-processing perspective, cognitive scientists developed several key elements that largely influenced the forthcoming cognitive theories of consciousness. These precursors included new models of attention and working memory, and also new concepts, such as the distinction between modules and central processes, or that between automatic and controlled processes.

In the present article, we will first overview the precursors that allowed the development of cognitive theories of consciousness. Then we will present a selection of influential contemporary accounts of consciousness. These theories will be

grouped according to three main themes: theories that consider consciousness to result from specific architectural elements within the cognitive system; theories claiming that some features of consciousness are in fact illusory; and theories that focus on the relation between consciousness and learning. We will conclude this article by emphasizing the common challenges that current cognitive models of consciousness have to face: the pressure from the philosophically defined hard problem, on one side, and the pressure from neurobiological evidences, on the other side.

Precursors for a Cognitive Perspective on Consciousness

In this section, we highlight the most influential precursors for a theory of consciousness. Most of these elements emerged during the cognitive revolution. Although they were originally sketched out in light of an information-processing perspective, those elements are now largely linked to the dissociation between conscious and unconscious processing.

Attention and the Central Executive

Various influential models developed in the 1960s referred to a central processor, a central executive system, or a supervisory system. Processing within the central system can be considered as analogous to conscious processing, even if the word consciousness was still largely banished in the scientific community. This system is at the top of the hierarchy in the cognitive architecture: it is involved in higher-order computations (decision, monitoring, planning, etc.) and leads to selection and control over lower-level subsystems. As in many contemporary accounts of consciousness, the central system was considered the most integrative element of the cognitive system, granting flexibility and control over behavior.

Another key element was the simple but powerful metaphor of attention as a filtering mechanism that was put forward by Broadbent. In a nutshell, peripheral processors in this theory provide sensory information to the central system dealing with control and decisions. Because multiple sensory channels are continuously acting in parallel, a huge quantity of information becomes available to the rest of the system. However, the central system is very limited in terms of computational resources. Hence, a selection mechanism is needed to prevent overload. As such, attention operates by selecting the most relevant information and by filtering out that which is irrelevant. Then, the most relevant information, which is under the focus of attention, becomes the target of the central system and can thus benefit from deeper and more enriched processing. Once again, although consciousness was not the main concern, one consequence of attentional selection was that it allowed the target information to become conscious. In this perspective, attention and consciousness are two tightly related notions.

The notion of short-term memory put forward by George Miller and later extended to the notion of working memory is also an important precursor. For example, in their model of working memory, Baddeley and Hitch relied on a central executive system, which has top-down control over the distinct specific subsystems, namely the phonological loop and the visuospatial sketchpad. Here, the content of working memory may be roughly equated with the content of consciousness, an aspect that will also be important for future cognitive theories of consciousness.

Norman and Shallice, in turn, proposed a model of action selection implicating a supervisory attentional system. This central system receives sensory evidence and determines the appropriate behavior by selecting instruction schemes for action mechanisms. In addition, the supervisory attentional system can be modulated by the goals of the organism, and it is primarily involved when a new or critical situation appears. Here too, the central part of the model shares some properties that are associated with consciousness, namely flexibility, reactivity regarding unexpected situations, decision, and control over behavior.

In sum, these influential early models depicted the global architecture of the cognitive system by emphasizing the following components: sensory inputs in the periphery that are processed in parallel in multiple channels, attention that performs selection upon these sources of information, a working memory component that keeps tracks of the selected information, and finally a central

system that acts as a supervisor. But one major limitation of this view is that it falls into the homunculus trap, when it comes to the question of consciousness. Indeed, if this central supervisor is governing the whole cognitive system, one may ask who is in turn governing the central supervisor! That is, if we were to rephrase this question by focusing on consciousness, it would be problematic to rely on a hypothetical little man in our head (i.e., a homunculus) that has consciousness, which is the same property we are supposed to explain. This approach unavoidably leads to an infinite regression. Because consciousness was not the main issue for these early models, this crucial issue was left out or even denied during the development of early cognitive models with a central supervisor. As we will see below, current theories of consciousness will overcome this limitation by proposing various cognitive architectures, sometimes including a central system, that take into account the homunculus issue.

Specialized Modules versus Flexible Integration

Closely related to this distinction between central and peripheral processes is the very influential theory of modularity developed by Jerry Fodor. In this framework, modules are fast and efficient devices that process inputs in an automatic and mandatory fashion. They are tuned to a particular kind of computation on a particular kind of information. In other words they are functionally specialized and they constitute the small computational bricks of cognitive architecture. While modules operate in the periphery in this architecture, they have been classically opposed to central processes that can be slow but flexible, and can integrate inputs from different modalities. Closely related to this architectural dissociation is Posner and Snyder's dissociation between, on the one side, automatic processes that are mandatory and fast and, on the other side, controlled processes that are assumed to be strategic and voluntary.

Here again, although this was not explicitly acknowledged in these various works, the fast and automatic processes operated by modules were assumed to reflect unconscious processing, while the control processes involving the central system were assumed to be conscious. These dichotomies between central and peripheral, controlled and automatic, flexible and hard-and-fast processes have provided the ground for the distinction between conscious and unconscious processing, which has been central in the development of current cognitive theories of consciousness. It is of note that the difficulty of studying consciousness did not arise only when researchers decided to face it. Although a few serious attempts have been made to propose functional description of the central system, such as in the Adaptive Control of Though (ACT) theory by John Anderson, this notion was itself often unspecified and often considered as a mysterious but needed component. For instance, Fodor strongly defended the idea that although the program of cognitive science was to understand how modules work, we would surely be in an impasse when trying to address the nature of central processes.

Architectural Accounts of Consciousness

We present in this section three of the most influential cognitive theories of consciousness. For each of them, consciousness is grounded in an information processing system. Baars' global workspace theory uses the metaphor of global broadcasting to describe conscious processing, Jackendoff and Prinz' intermediate level theory emphasizes the need for consciousness to be focused on intermediate representations, and Tononi's information integration theory proposes to relate consciousness with complexity in the cognitive system. The three accounts all share the same will to link consciousness with a particular representational aspect of the cognitive system. These theories differ, though, in many respects and thus provide a diversified sample of what cognitive accounts of consciousness can be.

The Global Workspace Theory of Consciousness

Grounded on the distinction between conscious and unconscious processes, Bernard Baars' global workspace theory is one of the most influential

cognitive theories of consciousness. This theory relies on the metaphor of a theater. In this theater, unconscious specialized processors (equivalent to modules) are assumed to be the actors and the audience. While the audience represents the set of passive processors, actors represents active processors playing on the stage of the theater (i.e., the workspace). These actors are engaged in a competition for being seen by the audience: by broadcasting their information they actually compete for more broadcasting. Active processors with the highest coherent activity can form local coalitions that strengthen them in this competition process. The strongest coalition in this competition dominates the workspace, in a winner-take-all fashion, and corresponds to the content of consciousness. The workspace is equated by Baars to working-memory, in which only the most active content becomes conscious. Additionally, the dominant coalition benefits from global broadcasting, which allows it to recruit new processors from the audience in order to form a global coalition. Here, consciousness allows for the integration of computational resources in a large-scale coordination and for the exchange of information among processors that would otherwise remain separated. In this theory, each processor can operate in the conscious mode if it benefits from global broadcasting through the workspace, or it can operate in the unconscious mode when disconnected from the workspace.

An important feature of the global workspace theory is the presence of contexts as stable coalitions shaping access to the workspace. Contexts are constituted of unconscious processors reflecting, in a hierarchical manner, our expectations, our beliefs, our goals, and ultimately our self. In particular, attention is implemented as a goal context in this theory. It is described as a mechanism that controls access to the workspace, acting as a filter and biasing the competition process toward a particular set of actors. At any given moment, the dominant coalition is under the spotlight of attention, and its informational content becomes the content of conscious experience.

A crucial aspect of Baars' theory is that it avoids the problem of the homunculus by reducing it to an audience of multiple unconscious processors. Here, there is no need for a hypothetical single conscious observer in the system, and thus there is no issue of infinite regression with a homunculus inside another homunculus. Instead, consciousness is considered to reflect the global broadcasting of information to an audience of unconscious processors. As the audience is unconscious, unsupervised, and receptive rather than attending to the information, it does not constitute an internal homunculus.

The Intermediate Level Theory of Consciousness

The intermediate level theory originally proposed by Ray Jackendoff and further defended and specified by Jesse Prinz proposes that within the hierarchy of representations that are used to describe the cognitive system, conscious experience occurs only for specific levels of representation.

The theory is rooted in Jackendoff's analysis of different cognitive systems such as vision, language, and music and the subsequent observation that consciousness does not arise anywhere within these systems. According to Jackendoff, consciousness is not associated with low-level, nor with high-level representations, but rather with those implying intermediate levels of processing. For instance, in the domain of object recognition, it is assumed that the visual system comprises a low level with local computations of visual features, an intermediate level reflecting binding and object recognition, and a higher level computing viewpoint invariance and representing abstract categories. According to Jackendoff and Prinz, conscious experience is not comprised of a disunified picture of visual features, nor is it represented by view-invariant categories. Rather it is composed of bound and specific instances of objects that are assumed to be computed at the intermediate level of representation. In an analogous manner, speech perception can be decomposed into three levels: an acoustic representation of speech sounds at the lower level, a high level involving abstract lexical and syntactic categories, and in between a word recognition level relying on phonological representations. This theory explains why the conscious experience associated with speech perception mostly involves phonological representations, rather than other levels of representations. In Jackendoff and Prinz' theory,

the privileged role of the intermediate level of processing is based on the need for real-time computational efficiency. Indeed, this level of representation is assumed to be the most relevant one regarding ecological and functional needs.

Another important aspect of this theory concerns the central role of attention during conscious experience. Here, attention is defined as a selection process that acts as a gate to working memory mechanisms. It performs the function of selecting the relevant information that is amplified afterward and then becomes conscious. Indeed, Prinz acknowledges that activation of an intermediate-level representation on its own cannot be a sufficient condition for consciousness, given that those representations can be activated during subliminal perception. However, this theory makes the crucial postulate that the amplification of intermediate-level representations by attention is a necessary and sufficient condition for consciousness. In sum, for each domain of processing, the content of consciousness at a particular moment is supported by a representational structure of intermediate level for that domain, which is selected to enter short-term memory, and enriched by attentional processing.

The Information Integration Theory of Consciousness

'The information integration theory of consciousness' has been proposed by Gulio Tononi to explain how consciousness arises from dynamic complex systems. It originates from Tononi's work with Gerald Edelman and their observation that conscious states share two fundamental properties: they are both differentiated and integrated. Conscious states are highly differentiated in the sense that the occurrence of a particular conscious state results from its selection among a huge repertoire of possible conscious states. As such, a conscious state carries an important amount of information, as it reflects a large reduction in uncertainty. At the same time, conscious states are integrated as a unified experience. For instance, one does not have the experience of the color of a particular shape independently from the experience of the shape itself. A given state in a system is considered to be integrated if it results from the interactions of diverse subsets within this system. To account for integration, Edelman and Tononi relied on the notion of neuronal reentry within a thalamocortical dynamic core.

The information integration theory, formulated more recently by Tononi, is more concerned with how any physical system, brain or machine, with both integrated and differentiated information can lead to conscious experience. In this theory, consciousness is a property of a system that can integrate differentiated information: the more one system exhibits integrated and differentiated states, the more it is conscious. Accordingly, Tononi proposed to measure information integration by means of a function labeled Φ, whose value allows one to assess the degree of consciousness within the system. This function Φ takes high values for systems with high complexity, such as small-world architectures where connectivity patterns between units are heterogeneous. Conversely, it has low values for simple and feedforward systems. Importantly, Tononi gives an operational method for the computation of Φ in a given system, based on decomposition of the system into its subsets. As such, he also puts forward the notion of a complex in a system: a complex is mathematically defined as a subset of the system that is not part of a subset of higher Φ value. Importantly, according to the information integration theory, the content of consciousness at a given moment corresponds to the information processed in the complex, which exhibits the highest Φ value, called the main complex of the system. As the system processes information dynamically, interactions between the different parts of the system are continuously changing. Thus, the main complex changes accordingly, and so does the content of consciousness.

One important aspect of this approach is that it considers consciousness to be a quantitative and graduate variable. Furthermore, as consciousness is only determined by the Φ measure, it is only a matter of system complexity in any system. Consequently, animals or mechanical systems exhibiting similar properties can be considered as having a certain degree of consciousness. Still, although the value of Φ can be computed in theoretical situations, with fully specified systems, one obvious difficulty is the measure of Φ in natural systems. The decomposition of the mind into relevant

subunits is still a matter of research, and the assessment of information processed by these subunits has been to date an untargeted issue.

Illusory Features Accounts of Consciousness

Several approaches have claimed that some features associated with conscious experience are in fact illusory. Here, we present the most popular views on this matter. A first perspective is represented by Daniel Dennett's multiple drafts model of consciousness, where the appearance of a unified stream of consciousness reflects an illusion produced during introspection. A second view is held by Daniel Wegner whose theory of apparent mental causation claims that free will and the fact that we consciously determine our actions is illusory. A third account, the sensory-motor theory of consciousness by Kevin O'Regan and Alva Noe, also takes phenomenal experience as a retrospective illusion. However, this theory also associates consciousness with a learning process, and thus so it will be addressed in the next section on 'Learning process accounts of consciousness.'

The Multiple Drafts Model of Consciousness

The quest for a conscious subsystem in the brain has been overtly criticized by Dennett who explicitly related it to the homunculus assumption. Instead, he proposed a multiple drafts model of consciousness in which information does not need to be represented in front of a conscious observer within our heads. In this model, the stream of consciousness is neither unified nor is it produced by a single narrative system. Instead, what makes the stream of consciousness apparently unified is a retrospective reconstruction involving multiple drafts describing the story.

In the multiple drafts model, the cognitive system continuously processes information in parallel in different threads, either in perceptual, conceptual, or motor domains. In fact, threads look like Fodorian modules or specialized processors of the global workspace, and their computations in progress are logged in a temporary draft. As such, multiple drafts are edited in parallel and continuously revised within the system. In addition, these drafts have different fates: some will be read by the rest of the system and will affect subsequent behaviors, while others will simply fade out. In Dennett's model, cerebral celebrity makes a particular draft conscious (or 'fame in the brain'), that is, the extent to which it affects other processes in the system, and eventually subsequent behaviors and responses. In particular, by introspecting ourselves and thus directing our attention to one particular thread, we let the content of this thread affect our behaviors and thus become conscious. Introspection can also have the consequence of modifying the content of the draft itself. For instance, if a thread is probed too late, the associated draft will not be available anymore, or it will be totally reconstructed on purpose. Conversely, if the thread is probed to early, its process is interrupted, and the draft that becomes conscious will not reflect further edition.

The multiple drafts model of consciousness is an early and influential cognitive theory of consciousness, developed with the will to eradicate problematic homunculus assumptions. Indeed, the theory emphasizes that there is not a single observer that would receive all the information and provide a single and unified narrative stream of consciousness. Rather there are multiple on-going processes, from which some drafts have sufficient impact to influence behaviors and lead to consciousness one after the other. According to Dennett, the illusion of a single narrative stream stems from the fact that the story is continuously revised in order to be more plausible. Though less specified than more recent accounts that are similar in principle (e.g., Baars' global workspace theory), this theory remains an interesting instance of a strongly reductive view, which offers to replace the central homunculus system by parallel and distributed processing in a network of threads or daemons. In this approach there is nothing more to consciousness than the causal impact that one particular content has on subsequent processing and behavioral reports. However, one possible criticism linked to this feature is that the explanation provided by Dennett is a theory of report rather than of conscious experience. This latter argument reflects more generally the critical problem of assessing consciousness without relying

on some kind of report (we return to this point in the conclusion of this article).

The Theory of Apparent Mental Causation

In everyday life as well as in scientific accounts, consciousness is usually associated with the determination and control of appropriate behaviors. In his apparent mental causation theory, Daniel Wegner takes a different view in which consciousness and will are actually determined by unconscious causes, and have no real causal role in return. This view is also called epiphenomenalism, as it considers that conscious experience is an epiphenomenon that accompanies unconscious processes, but has no functional role.

In Wegner's theory, our conscious thoughts do not necessarily cause our behaviors. Rather, both conscious thoughts and behaviors are caused by unconscious mechanisms. These underlying unconscious causes of thoughts and the unconscious causes of behaviors are different, though they can influence each other. Because of the mutual influence between these two types of unconscious causes, their effects (i.e., conscious thoughts and conscious behaviors) are correlated as well. Because conscious thoughts happen just before conscious behaviors, they are taken to be the causes of initiated actions. Here, the attribution of a causal role to conscious thoughts is an illusion based on what is apparent to consciousness, not on what really happens. In addition, the theory specifies the condition under which this illusion occurs: thoughts have to appear just before an action (priority), they have to be consistent with the action (consistency), and they have to be the only possible explanation of the action (exclusivity). When these conditions are satisfied, conscious thoughts contain a predictive model of the forthcoming action, and when the action is realized in agreement with the predictions, we grant authorship for it and we experience ourselves as causal agents.

The theory of apparent mental causation does not aim at explaining how consciousness arises in a cognitive system. Rather, it explains how our conscious experience of will is an illusion that stems from our ignorance of actual unconscious causal chains. This approach has the advantage of trying to eliminate a false but still well-established *a priori* about the experience of conscious will. Nonetheless, one might wonder why then would we experience this illusion? Wegner proposes that it may help the subject to maintain his goals through consciousness or to build a better representation of the world, in which his own contributions are marked as such. This theory, however, suffers from an important difficulty, as it is expressed in terms that remind us of the homunculus problem, as pointed out by Dennett. Indeed, there is still in Wegner's account one self: someone who is conscious, someone who attributes causality to conscious thoughts, someone who is experiencing the illusion of conscious will, and who has in fact the properties of a homunculus.

Learning Process Accounts of Consciousness

Here, we present three theories that emphasize the influence of learning on consciousness. In their sensory-motor theory of consciousness, Kevin O'Regan and Alva Noë put forward the notion of learnt sensory-motor contingencies. In both Axel Cleeremans' radical plasticity thesis and Hakwan Lau's higher-order Bayesian decision theory, consciousness results from the ability of the cognitive system to learn about its own internal states.

The Sensory Motor Theory of Consciousness

Most models of vision are based on internal detailed representations that are active when a particular stimulus is present in the visual world. The sensory-motor theory of consciousness proposed by Kevin O'Regan and Alva Noë takes an alternative view in which there is, according to them, no need for detailed representations in the brain, and in which conscious experience is produced by the mastery of sensory-motor contingencies.

In normal situations, the observer knows that he only has to direct his eyes or his attention toward it in order to obtain detailed information. In other words, the world is an external memory, and the information it carries is usually sufficient for action. As a consequence, rather than relying

on internal representations that would be at the origin of conscious experience, this theory considers consciousness to be an active and exploratory process in between the observer and the external environment. In support of this theory, several studies, including some experiments by O'Regan and Noë, have shown that normal observers can suffer from 'change blindness,' a situation predicted by the idea that our memory lies in the outside world. In this situation, observers have the illusion that they are conscious of the whole visual scene while, actually, they fail to notice important modifications in the scene. Importantly, these changes are noticed when participants direct their eyes or their attention to the critical location. These findings show that observers have an illusory and overconfident estimation of their visual capacities. The sensory-motor theory of consciousness also proposed to explain some features of conscious experience on the basis of the characteristics of the sensory-motor contingencies, i.e., the principles that link exploration acts to sensorial consequences. In vision, for instance, a saccade to the left will shift the visual input on the retina accordingly, but even if the position of the object in front of you has changed on your retina, you would still feel that this object has not moved: this principle is embodied in your sensory-motor contingencies. Besides, the different sensorial modalities are different means for exploring the environment, and among these modalities, the differences in the sensory-motor contingencies (e.g., optical laws differ from acoustical laws) are the basis for the differences in the structure of conscious experience. Importantly, these contingencies apply at different levels of abstraction: some relate to the physical apparatus of the stimuli in a given modality, while others relate to features or categorical attributes. When we look at a particular object from a changing viewpoint, the visual image changes but the category of the object remains constant.

The most original idea expressed by the sensory-motor theory is that external stimuli do not have to be represented in detail in the brain. Importantly, however, O'Regan and Noë are NOT against any form of representation, or any storage of information in the brain. Following their own terms, they grant at least that the knowledge of the laws of sensory-motor contingencies have to be represented. What is rejected is the notion of continuous detailed representations of the outside world, and the fact that having these representations could be sufficient to create consciousness, without making use of it, in the sense of exploring it through sensory-motor contingencies. Regarding this question, one interesting argument put against this theory was the issue of dreams or mental imagery. Since those phenomena provide compelling intuitive support for the existence of such detailed internal representations, how does the sensory-motor theory deal with that? The answer provided by O'Regan and Noë is that there are still some differences between normal visual experience and dreaming or visual imagery situations, which correspond to the fact that in the latter cases the subject cannot make use of all the sensory-motor contingencies. Additionally, they deny that dream-experiences are stable in the details, as they miss the stability of the world as a memory.

The Radical Plasticity Thesis of Consciousness

While many cognitive models use symbolic and discrete representations, connectionist models rely on sub-symbolic and distributed units in artificial neural networks. In these models, representations are patterns of activation over processing units. Following this perspective, Axel Cleeremans proposed a conceptual framework termed the Radical Plasticity Thesis that put a strong emphasis on the link between conscious awareness and learning. This theory is based on three main principles.

The first main idea states that learning is a mandatory consequence of information processing, leading the cognitive system to develop representations of higher quality. Here, the quality of a representation is assessed by the stability and strength of activation in the dynamic network, and distinctiveness, which is equivalent to differentiation in the theory put forward by Tononi (see 'The information integration theory of consciousness' above). The second important idea is that consciousness reflects the quality of representations within the cognitive system. In this theory, the more representations achieve high

quality (i.e., high strength, stability, and distinctiveness), the more they participate in conscious experience. Hence, in this theory, conscious experience is a graded and continuous variable. Given those two principles, learning is associated with higher quality representations, which are in turn more likely to be conscious. Finally, the third principle highlights the implication of metarepresentations for self-consciousness. Cleeremans proposes that high-quality representations are efficient detectors of a particular content, and that they can be the target of metarepresentations. These metarepresentations capture the associations between first-order representations, which are developed through learning and past experience. The theory also proposes that a metarepresentation helps the first-order representations on which it is focused to achieve higher quality. In other words, one system can support consciousness insofar as it is able to learn about its environment and create internal representations, and also be able to learn about its own representations and increase their quality. Here, the more the system knows about its own rules, the more it is assumed to be conscious.

Cleeremans further distinguishes between different aspects of conscious experience and describes how these aspects correlate with the increase, through learning, in the quality of representations. The formation through learning of internal representations is depicted in three stages. The first one relates to implicit cognition: a poor-quality representation can influence behavior, but it is not strong enough to let the subject know about these influences or to have much control over them. Through exposure and learning, the representation achieves higher quality and becomes explicit. In this second step, the availability to control and the potential impact on the cognitive system also increase dramatically and reach a maximum. When the representation is sufficiently learned, it becomes automatic. According to Cleeremans, this third stage is associated with high-quality representations readily available to conscious awareness, though the subject has less control over their influences as they operate in a mandatory way. Thus, in this final idea, Cleeremans takes a view that can be contrasted with the classical assumption that automatic processes are unconscious.

The Higher-Order Bayesian Decision Theory of Consciousness

Signal detection theory and Bayesian frameworks have recently undergone a great renewal of interest among cognitive scientists. These conceptual tools bring useful insights in the description of behavioral performance, such as discrimination, detection, and decision. In a nutshell, signal detection theory proposes that discrimination between target and noise relies, on the one hand, on the objective distance (discriminability) between their two signal distributions on a psychophysical continuum and, on the other hand, on the particular setting of a decision threshold (criterion or bias) on that continuum. Bayesian decision theory, in turn, proposes a way to optimize the setting of the decision threshold, through prior learning over time of the probability distributions of the noise and target signals.

In many empirical studies on consciousness, participants' awareness of a given stimulus is equated with their performance on discrimination tasks (i.e., discriminability). Conversely, chance level performance on a discrimination task is often assumed to imply that the participant is completely unaware of the feature targeted by the task. Hakwan Lau's higher-order Bayesian decision theory of consciousness uses empirical dissociations between performance and awareness to support the idea that consciousness may not always be associated with an increase in discriminability. Rather, the hypothesis defended here is that it is related to the setting and the maintaining of the criterion threshold used for the perceptual decision. This theory associates some features of the higher-order thought theory with the Bayesian decision framework. More precisely, it proposes that while the lower-order system implements discriminability, the higher-order system, in turn, implements the decision threshold. In this view, the lower-order system performs a certain number of discriminations upon external signals, and the higher-order system learns about the distribution of states of the lower-order system, so as to interpret the signal, and to be able to set the threshold in an optimal manner. While Lau's theory is to date clearly not developed as far as other proposals, this work provides a new idea to the current theoretical landscape. It addresses a theoretical issue related to signal detection theory, which

is a methodological tool of increasing importance in the field of consciousness.

Conclusion

In this article we have presented an overview of the most representative cognitive accounts of consciousness. Most of these theories radically differ in their conception of what consciousness is. While some consider that it reflects the activation of attended intermediate level representations (Jackendoff, Prinz), or the involvement of complexes in a system (Tononi), others would equate consciousness with global broadcasting (e.g., Baars, Dennett), and still others would associate it with learning upon ones' own representations (e.g., Cleeremans, Lau) or upon sensory-motor contingencies (O'Regan and Noë). As such, it is obvious that consciousness is not yet a well-defined notion. In addition, cognitive accounts are now facing two epistemological constraints that impose important pressure on their development. The first one has been put forward by philosophers and corresponds to the need to focus on the 'hard problem' rather than the 'easy problem' of consciousness. The second one is related to the increasing amount of empirical evidence resulting from the study of the brain. We conclude this article by focusing on these two constraints.

The philosopher David Chalmers termed a dichotomy between the easy problem and the hard problem of consciousness in order to delineate the two major features related to consciousness. On the one hand, consciousness offers a processing advantage, as it allows for the information in working memory to be processed in a long-run by multiple devices. On the other hand, consciousness carries the qualitative property of subjective experience. Chalmers states that the first issue is in fact an interesting though only a computational problem; hence it is easy to study scientifically. The second one, however, is much more mysterious. How can the subjective quality of experience arise from squishy organic matter is a question that seems to go far beyond our possible understanding. The distinction between the hard problem and the easy problem can be mapped onto the dichotomy between access consciousness and phenomenal consciousness proposed by the philosopher Ned Block. Access consciousness relates to the global use of conscious information, and the possibility through consciousness to trigger complex and integrated processes such as reasoning, control of actions, decisions, and verbal reports. In contrast, phenomenal consciousness refers to the mere subjective experience, the 'what it is like' question expressed also by Thomas Nagel. Both Chalmers and Block defend phenomenal consciousness and the hard problem, claiming that conscious experience of a stimulus is not reducible to its information processing and its causal influences in the system.

It turns out that most cognitive models are expressed in information processing terms, and as such they are bound to take a reductive approach when trying to explain phenomenal consciousness. Intrinsically, they favor functionalist perspectives whereby information processing is all there is to conscious experience. On a more general perspective, science deals with measurements and measurements are by definition targeted to a piece of information that is measured. Hence, in most cognitive accounts, the hard problem is either reduced to the easy one or even completely denied. For instance, some will stand that it might be in fact necessary to revise our definition of what consciousness is, in order to eradicate any reference to some mysterious 'phenomenal' properties of the mind (e.g., Dennett). Indeed, one should not overlook the possibility that phenomenal and access consciousness are two notions that have to be dissociated only conceptually. They are not easy to dissociate experimentally since any measure of phenomenal consciousness can hardly be dissociated from the involvement of access consciousness. Indeed, measuring phenomenal consciousness in an experimental perspective must be based on some form of report, hence on access consciousness.

The other limit of purely functional accounts of consciousness comes from the brain. Indeed, as they remain distant from the biological implementation, purely functional or philosophical perspectives on consciousness are now likely to miss this crucial dimension, and source of evidence. Consequently, they will lack the same amount of explanatory power. Indeed, consciousness has become for many scientists a biological problem whose answers will be found by studying the brain. In

fact, some authors go a step further in arguing that the operational definition for the scientific study of consciousness should be expressed in neural terms. In that perspective, the psychological tools that we use may all be discarded in favor of a more physiologically grounded approach. Even if we do not want to go that far, it is a matter of fact that today basic observations of brain processes might help, by providing new concepts that would help directing research, and new critical test that would help discarding unfitting theories. Memory is a good example. Functional accounts of memory had to go back to the drawing board when the neurology of memory began to be understood in a more precise way. On the other hand, functional and neurological accounts have sometimes worked together productively, as for instance Baars' global workspace theory, which has been extended at the neurobiological level by Stanislas Dehaene and colleagues. Of course, without theoretical knowledge of the functions that are to be explained, a purely biological theory of cognition would be impossible to construct. In other terms, the union between psychological and neurobiological perspectives makes both approaches stronger, and future models of consciousness will be bound to include brain evidences and hence to be transformed into neurocognitive rather than purely cognitive accounts of consciousness.

See also: History of Philosophical Theories of Consciousness; Neurobiological Theories of Consciousness.

Suggested Readings

Anderson JR (1983) The Architecture of Cognition. Cambridge, MA: Harvard University Press.
Baars BJ (1988) A Cognitive Theory of Consciousness. New York: Cambridge University Press.
Baddeley AD and Hitch GJ (1974) Working memory. In: Bower GA (ed.) The Psychology of Learning and Motivation, pp. 47–89. Academic Press.
Block N (1995) On a confusion about a function of consciousness. Behavioural and Brain Sciences 18: 227–247.
Block N, Flanagan O, and Güzeldere G (eds.) (1997) The Nature of Consciousness: Philosophical Debates. Cambridge, MA: MIT Press.
Broadbent DE (1958) Perception and Communication. Pergamon Press.
Chalmers D (1995) Facing up to the problem of consciousness. Journal of Consciousness Studies 2(3): 200–219.
Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.
Cleeremans A (2005) Computational correlates of consciousness. Progression in Brain Research 150: 81–98.
Cleeremans A (2008) Consciousness: The radical plasticity thesis. Progression in Brain Research 168: 19–33.
Dehaene S and Naccache L (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79(1–2): 1–37.
Dennett DC (1991) Consciousness Explained. Little, Brown & Company.
Edelman GM and Tononi G (2000) A Universe of Consciousness: How Matter Becomes Imagination. New York: Basic Books.
Fodor J (1983) The Modularity of Mind: An Essay on Faculty Psychology. MIT Press.
Jackendoff R (1987) Consciousness and the Computational Mind. Cambridge: MIT Press.
Lamme VA (2006) Towards a true neural stance on consciousness. Trends in Cognitive Sciences 10: 494–501.
Lau HC (2008) A higher-order Bayesian decision theory of perceptual consciousness. Progression in Brain Research 168: 35–48.
Mack A and Rock I (1998) Inattentional Blindness. Cambridge, MA: MIT Press.
Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63: 81–97.
Nagel T (1974) What is like to be a bat? Philosophical Review 83: 434–450.
Norman DA and Shallice T (1986) Attention to action: Willed and automatic control of behavior. In: Davidson R, Schwartz G, and Shapiro D (eds.) Consciousness and Self Regu lation: Advances in Research and Theory, 2nd edn., vol. 4, pp. 1–18. New York: Plenum.
Online Papers on Consciousness. Compiled by David Chalmers.http://consc.net/online/
Posner MI (1994) Attention: The mechanism of consciousness. Proceedings of the National Academy of Sciences of the USA 91(16): 7398–7402.
Posner MI and Snyder CRR (1975) Attention and cognitive control. In: Solso RL (ed.) Information.
Prinz J (2005) A neurofunctional theory of consciousness. In: Brook A and Akins K (eds.) Cognition and the Brain: Philosophy and Neuroscience Movement, pp. 381–396. Cambridge: Cambridge University Press.
Rumelhart DE and McClelland JL (1986) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. Cambridge, MA: MIT Press.
O'Regan JK and Noë A (2001) A sensorimotor account of vision and visual consciousness. Behavioural and Brain Sciences 24(5): 883–917.
Schneider W and Shiffrin RW (1977) Controlled and automatic human information processing: Vol. I. Detection, search, and attention. Psychological Review 84: 1–66.
Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5: 42.

Tononi G and Edelman GM (1998) Neuroscience – Consciousness and complexity. Science 282(5395): 1846–1851.
Velmans M and Schneider S (eds.) (2007) The Blackwell Companion to Consciousness, pp. 225–235. Oxford, UK: Blackwell Publishing.
Wegner DM (2002) The Illusion of Conscious Will. MIT Press.
Wegner DM (2003) The mind's best trick: How we experience conscious will. Trends in Cognitive Sciences 7: 65–69.

Biographical Sketch

Vincent de Gardelle is a doctoral student supervised by Sid Kouider at the Ecole Normale Supérieure (Paris, France). His research addresses the question of biased conscious perception, by examining the perceptual illusions that might result from a subject's expectations under poor sensory evidence.

Sid Kouider is a cognitive neuroscientist working at the Ecole Normale Supérieure (Paris, France) on the neurobiological and psychological foundations of consciousness. His work focuses on contrasting conscious and unconscious processes both at the psychological and at the neural level, using various behavioral and brain imaging methods. Recently, he extended this line of research to study the neural correlates of consciousness in prelinguistic babies.

Coma, Persistent Vegetative States, and Diminished Consciousness

A Demertzi, S Laureys and M Boly, University of Liège, Liège, Belgium

Glossary

Apnea testing – A test needed to confirm brain death by checking whether the patient has a breathing reflex when disconnected from the positive pressure ventilator.
Brain–computer interfaces (BCIs) – Real-time muscular-independent systems that permit the translation of the electrical activity of the brain into commands, to control devices.
Deep brain stimulation (DBS) – An invasive surgical treatment involving the implantation of a medical device (brain pacemaker), which sends electrical impulses to specific parts of the brain.
Default mode network – A set of brain areas, encompassing the posterior cingulate cortex/precuneus, the medial prefrontal cortex, and bilateral temporoparietal junctions, which seem to be activated in the absence of any external stimulation, and show decreased activity during cognitive processing.
Event-related potentials (ERPs) – Averaged EEG signals that detect time-locked responses to sensory, motor, or cognitive activities. Short-latency or exogenous ERPs, ranging from 0 to 100 ms after the presentation of a stimulus, correspond to the activation of the ascending pathways to the primary cortex. Cognitive or endogenous ERPs are obtained after 100 ms of the presentation of a stimulus, and reflect both subcortical and cortical structures, including associative areas.
Functional connectivity – The temporal correlation of a neurophysiological index (i.e., cerebral metabolic rates of glucose, regional cerebral blood flow) measured in different remote brain areas.
Neuron-specific enolase – The neuronal form of the glycolytic enzyme enolase, which is found almost exclusively in neurons and cells of neuroendocrine origin and is used as a marker of ischemic brain damage.

Introduction

The management of coma and related disorders of consciousness (DOC) is a major clinical challenge. Patients in a vegetative state and minimally conscious state continue to pose problems in terms of their diagnosis, prognosis, and treatment. Bedside assessment remains the gold standard. Neuroimaging and electrophysiological measures can now identify signs of awareness inaccessible to clinical examination, which permit a better understanding of the mechanisms of human consciousness and improve our care of DOC patients.

Defining Consciousness

Consciousness is a first-person experience, which consists of two major components, wakefulness and awareness. Wakefulness refers to the level of consciousness and it is supported by the function of the subcortical arousal systems in the brainstem, the midbrain, and the thalamus. Clinically, it is indicated by opening of the eyes. Awareness refers to the contents of consciousness and it is thought to be supported by the functional integrity of the cerebral cortex and its subcortical connections. Awareness can be further reduced to awareness of the environment and of self. Clinically, awareness of the environment is assessed by evaluating command following and observing nonreflex motor behavior, such as eye tracking and oriented

responses to pain. Awareness of self, clinically a more ill-defined concept, can be assessed by the patients' response to autoreferential stimuli, such as the patients' own face in the mirror. An illustrative example of the relationship between the two components of consciousness is the transition from full wakefulness to deep sleep: the less aroused we get, the less aware we become of our surroundings and ourselves (see Figure 1).

A Short History of Disorders of Consciousness

About 50 years ago, before the era of neurocritical care, things were relatively simple. After a severe brain damage, comatose patients either died or, more rarely, recovered with more or less cognitive deficits. The invention of the positive pressure mechanical ventilator by Bjorn Ibsen in the 1950s, and the widespread use of intensive care in the 1960s, in the industrialized world, changed the picture. They stated that severely brain damaged patients could now have their heartbeat and systemic circulation sustained by artificial respiratory support. Such profound unconscious states had never been encountered before as, until that time, all these patients had died instantly from apnea.

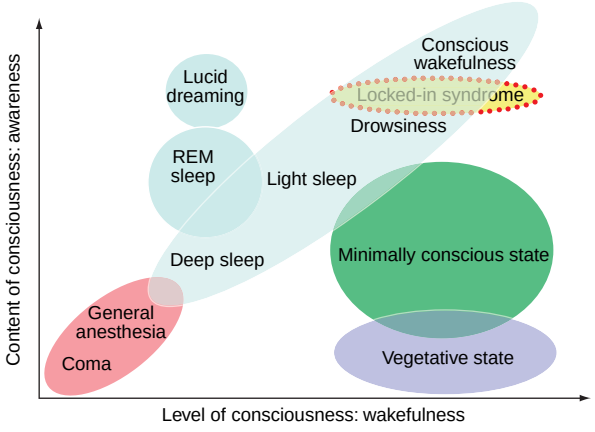


Figure 1 Simplified illustration of the two major components of consciousness and the way they correlate within the different physiological, pharmacological and pathological modulations of consciousness. Reproduced from Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. Trends in Cognitive Sciences 9: 556–559.

As a consequence, medicine was forced to redefine death, using a neurological definition, that of brain death.

In the 1960s, Fred Plum and Jerome Posner described for the first time the locked-in syndrome (LIS), to refer to fully conscious coma survivors who are unable to communicate due to physical paralysis. In 1972, Bryan Jennet and Fred Plum published the clinical criteria of another artifact of modern intensive care, the vegetative state (VS), a state of 'wakefulness without awareness.' In 2002, the Aspen Neurobehavioral Conference Workgroup realized that clinical reality was yet more complicated. Some patients showed signs of voluntary behavior, and therefore they were no longer vegetative, but still remained unable to functionally communicate. Based on these observations, they published the diagnostic criteria of a new clinical entity, the minimally conscious state (MCS).

Defining the Clinical Entities of Consciousness

Brain Death

Brain death means human death determined by neurological criteria. The current definition of death is the permanent cessation of the critical functions of the organism as a whole, such as, neuroendocrine and homeostatic regulation, circulation, respiration, and consciousness. Most countries, including the United States, require death of the whole brain including the brainstem. Some other countries, like the United Kingdom and India, rely on the death of the brainstem only, arguing that the brainstem is at once the throughstation for nearly all hemispheric input and output, the center generating wakefulness (an essential condition for conscious awareness), and the center of respiration. Classically, brain death is caused by a massive brain lesion, such as trauma, intracranial hemorrhage, or anoxia. Using the brainstem formulation of death, however, unusual but existing cases of catastrophic brainstem lesions, usually of hemorrhagic origin, sparing the thalami and cerebral cortex, can be declared brain dead in the absence of clinical brainstem function, despite intact intracranial circulation. Hence, a patient with a primary brainstem lesion who did not

develop raised intracranial pressure might theoretically be declared dead by the UK doctrine, but not by the US doctrine.

In 1995, the American Academy of Neurology published the criteria for brain death, which have been used to model many institutional policies. The criteria are (1) demonstration of coma; (2) evidence for the cause of coma; (3) absence of confounding factors, including hypothermia, drugs, electrolyte, and endorcrine disturbances; (4) absence of brainstem reflexes; (5) absent motor responses; (6) positive apnea testing (see 'Glossary'); (7) a repeat evaluation in 6 h is advised, but the time period is considered arbitrary; and (8) confirmatory laboratory tests are only required when specific components of the clinical testing cannot be reliably evaluated. At present, no recovery from brain death has been reported.

Coma

Patients that sustain severe brain damage may spend some time in coma, which lasts for a couple of days or weeks. Patients in coma cannot be awakened even when intensively stimulated and, hence, are not aware of the environment and of themselves (see Figure 1). Coma is distinguished from syncope or concussion in terms of its duration, which is at least 1 h. Coma can result from bihemispheric diffuse cortical or white matter damage or brainstem lesions bilaterally, affecting the subcortical reticular arousing systems. Many factors such as etiology, the patient's general medical condition, age, clinical signs, and complementary examinations influence the management and prognosis of coma. Traumatic etiology is known to have a better outcome than nontraumatic anoxic cases. In terms of clinical signs, after 3 days of observation, a bad outcome is heralded by the absence of pupillary or corneal reflexes, stereotyped or absent motor response to noxious stimulation, bilateral absent cortical responses of somatosensory-evoked potentials (SEPs) (see 'Glossary'), and (for anoxic coma) biochemical markers, such as high levels of serum neuron-specific enolase (see 'Glossary').

Vegetative State

In the VS there is dissociation between wakefulness, which is preserved, and awareness, which is absent (see Figure 1). These patients regain sleep–wake cycles. However, their motor, auditory, and visual functions are restricted to mere reflexes and show no adapted emotional responses. The VS is usually caused by diffuse lesions on the gray and white matter. According to the 1994 Multi-Society Task Force on persistent vegetative state (PVS), the criteria for the diagnosis of VS are the following: (1) no evidence of awareness of self or environment and an inability to interact with others; (2) no evidence of sustained, reproducible, purposeful, or voluntary behavioral responses to visual, auditory, tactile, or noxious stimuli; (3) no evidence of language comprehension or expression; (4) intermittent wakefulness manifested by the presence of sleep–wake cycles; (5) sufficiently preserved hypothalamic and brainstem autonomic functions to permit survival with medical and nursing care; (6) bowel and bladder incontinence; and (7) variably preserved cranial nerve and spinal reflexes.

The VS may be a transition to further recovery, or may be permanent. 'Permanent' VS refers to patients whose chances for recovery are close to zero. This is the case for VS that lasts more than 1 year after traumatic, or 3 months after nontraumatic (anoxic) injury. The VS is characterized as 'persistent,' when a patient is in this state for more than 1 month. As both terms are abbreviated as 'PVS,' it has been suggested to avoid these terms and, instead, mention the etiology and the time spent in VS. At present, there are no validated prognostic markers for individual patients except that the chances for recovery depend on patient's age, etiology, and time spent in the VS.

Minimally Conscious State

The MCS has been defined in 2002 by the Aspen Workgroup as a DOC in order to describe noncommunicating patients that show inconsistent, but discernible signs of behavioral activity that is more than reflexive in at least one of the following behavioral signs: (1) purposeful behavior, including movements or affective behavior that occurs in contingent relation to relevant environment stimuli and is not due to reflexive activity, such as: pursuit eye movement or sustained fixation occurring in direct response to moving or salient stimuli, smiling or crying in response to verbal or visual

emotional but not neutral stimuli, reaching for objects, demonstrating a relationship between object location and direction of reach, touching or holding objects in a manner that accommodates the size and shape of the object, and vocalizations or gestures occurring in direct response to the linguistic content of questions, (2) following simple commands; (3) gestural or verbal yes/no response, regardless of accuracy; and (4) intelligible verbalization.

Like the VS, the MCS may be chronic and sometimes permanent. Emergence from the MCS is defined by the ability to exhibit functional interactive communication or functional use of objects. Given that the criteria for the MCS have only recently been introduced, there are few clinical studies of patients in this condition. Similar to the VS, traumatic etiology has a better prognosis than nontraumatic anoxic brain injuries. Preliminary data show that the overall outcome in the MCS is more favorable than in the VS.

The Locked-In Syndrome

The LIS describes patients who are awake and conscious, but have no means of producing speech, limb, or facial movements, resembling patients in a VS. LIS most commonly results from lesions to the brainstem. According to the 1995 American Congress of Rehabilitation Medicine criteria, LIS patients demonstrate: (1) sustained eye-opening (bilateral ptosis should be ruled out as a complicating factor), (2) quadriplegia or quadriparesis, (3) aphonia or hypophonia, (4) a primary mode of communication that uses vertical or lateral eye movement or blinking of the upper eyelid to signal yes/no responses, and (5) preserved cognitive abilities. Since there is only motor output problem, LIS is not a DOC, but it is included here as it can be misdiagnosed as one. Based on motor capacities, LIS can be divided into three categories: (1) classic LIS, which is characterized by quadriplegia and anarthria with eye-coded communication; (2) incomplete LIS, which is characterized by remnants of voluntary responsiveness other than eye movement; and (3) total LIS, which is characterized by complete immobility including all eye movements, combined with preserved consciousness.

Once an LIS patient becomes medically stable, and given appropriate medical care, life expectancy now is for several decades. Even if the chances of good motor recovery are very limited, existing eye-controlled, computer-based communication technology (i.e., BCI, see 'Glossary') currently allows these patient to control their environment. Neuropsychological testing batteries adapted and validated for eye-response communication, have shown preserved intellectual capacities in LIS patients, whose lesions are restricted to brainstem pathology. Recent surveys show that chronic LIS patients self-report a meaningful quality of life and the demand for euthanasia, albeit existing, is infrequent.

Evaluation of the Disorders of Consciousness

Good medical management starts with good diagnosis. However, as awareness is a first-person perspective, its objective assessment is difficult. For that reason, at the bedside, clinicians need to infer it via the evaluation of motor activity and command following. Diagnosing DOC correctly is extremely challenging. This is mainly because these patients are usually deprived of the capacity to make normal physical movements and may show limited attentional capacities. Aphasia, apraxia, and cortical deafness or blindness are other possible confounders in the assessment of DOC. This, in combination with the difficulty to define uncertain behavioral signs as voluntary or reflexive, can partially explain the high rate of incorrect diagnosis of DOC, which has been estimated to be around 40% of the cases. Besides these difficulties, one should also consider that some of the diagnostic criteria for VS and MCS do not share international consensus, such as, visual fixation, eye tracking, blinking to visual threat, and oriented motor responses to noxious stimuli.

Behavioral Evaluation

In 1974, Teasdale and Jennett's Glasgow coma scale (GCS) was published in '*The Lancet*.' This standardized bedside tool to quantify consciousness

became a medical classic, thanks mainly to its short and simple administration. The GCS measures eye, verbal, and motor responsiveness. There may be some concern as to what extent eye-opening is sufficient evidence for assessing brainstem function. Additionally, the verbal responses are impossible to be measured in cases of intubation and tracheotomy. Most importantly, the GCS is not sensitive enough to detect transition from the VS toward the MCS.

To differentiate VS patients from MCS patients, the most appropriate scale is the coma recovery scale-revised (CRS-R). The CRS-R has a similar structure to the GCS, containing, in addition to motor, eye, and verbal subscales, also auditory, arousal, and communication subscales. Despite its longer administration (i.e., c. 20 min) as compared to the GCS and the full outline of unresponsiveness (FOUR), it is the most sensitive in differentiating VS patients from MCS patients. This is because it assesses every behavior according to the diagnostic criteria of the VS and the MCS, such as, the presence of visual pursuit and visual fixation. Importantly, the way we assess these behavioral signs need to be standardized and uniform, permitting between-centers comparisons. For example, for the assessment of visual pursuit, some scales use an object or finger (FOUR), some use a mirror, a person, an object, and a picture (Western Neuro-Sensory Stimulation Profile), some use an object and a person (Wessex Head Injury Matrix; Sensory Modalities Assessment and Rehabilitation Technique), and some a moving person (Coma/Near Coma Scale). We have shown that the use of a mirror is more sensitive in detecting eye tracking and, hence, identify MCS patients. These findings stress that self-referential stimuli have attention-grabbing properties and are important in the assessment of DOC.

Despite their pros and cons, each scale contributes differently in establishing the diagnosis and prognosis of DOC. The administration and interpretation of findings should be decided and discussed in terms of the person who uses the scale, the place where it is administered (e.g., intensive care vs. chronic rehabilitation settings), and the reasons for administration (e.g., clinical routine vs. research purposes).

In Search for Objective Markers of Consciousness

Electrophysiology

The EEG allows recording of the spontaneous electrical brain activity, permitting the identification of the level of vigilance and the detection of functional cerebral anomalies, such as seizures or encephalopathy. In brain death, the EEG shows absent electrocortical activity with a sensitivity and specificity of around 90%. In coma, a burst suppression in the EEG heralds a bad outcome. In the VS, the EEG often shows a diffuse slowing and it is only sporadically isoelectric. Similarly, in MCS there is a general slowing on the EEG. In LIS, the EEG does not reliably distinguish these patients from VS patients. However, a close-to-normal EEG should have the physician consider the possibility of LIS.

The use of ERPs (see 'Glossary') is useful to predict the outcome in DOC. Bilateral absence of cortical potentials (i.e., N20) or SEPs heralds a bad outcome in coma. The presence of 'mismatch negativity' (MMN), a late cognitive ERP component that is elicited in auditory 'oddball' paradigms, is predictive of recovery of consciousness. In VS, SEPs may show preserved primary somatosensory cortical potentials (SEPs), and brainstem auditory-evoked potentials (BAEPs) often show preserved brainstem potentials. Endogenous-evoked potentials, measuring the brain's response to complex auditory stimuli, such as the patient's own name (as compared to other names) permits to record a P300 response, which delayed in DOC patients when compared to controls. However, a P300 is not a reliable marker of consciousness as it can also be detected during deep sleep and anesthesia.

Resting cerebral metabolism

Cortical metabolism in coma survivors is reduced on an average to 50%–70% of the normal values. A global depression of cerebral metabolism is not unique to coma. When anesthetic drugs are titrated to the point of unresponsiveness, the resulting reduction in brain metabolism is similar to that observed in pathological coma. Another example of transient metabolic depression can be observed during slow-wave sleep. In this daily physiological condition, the cortical cerebral

metabolism can drop to nearly 40% of the normal values – while in REM-sleep, the metabolism returns to normal waking values (see Figure 2).

In brain death the so-called 'empty-skull sign' is observed, denoting functional decapitation. VS patients show substantially reduced, but not absent, overall cortical metabolism, up to 40%–50% of the normal values. In some VS patients who subsequently recovered, global metabolic rates for glucose metabolism did not show substantial changes. Hence, the relationship between the global levels of brain function and the presence or absence of awareness is not absolute. It rather seems that some areas in the brain are more important than others for its emergence. Statistical analyses of metabolic positron emission tomography (PET) data have identified a dysfunction in a wide frontoparietal network encompassing the polymodal associative cortices: bilateral lateral frontal regions, parieto-temporal and posterior parietal areas, mesiofrontal, posterior cingulate, and precuneal cortices (see Figure 3). However, awareness seems not to be exclusively related to the activity in this 'global workspace' cortical network, but, as importantly, to the functional connectivity within this system and with the thalami. Long-range, frontoparietal, and thalamocortical 'functional disconnections,' with nonspecific intralaminar thalamic nuclei, have been identified in the VS. Moreover, recovery is paralleled by a functional restoration of this frontoparietal network and part of its thalamocortical connections.

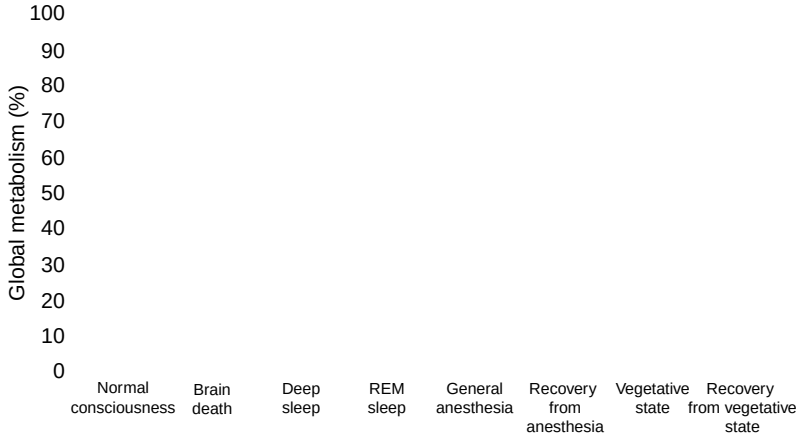


Figure 2 Global cerebral metabolism in healthy, pharmacological and disorders of consciousness. Adapted from Laureys S, Owen AM, and Schiff ND (2004) Brain function in coma, vegetative state, and related disorders. Lancet Neurology 3: 537–546.

Figure 3 The frontoparietal ''awareness network'' (orange) is systematically the most impaired region in the vegetative state. The blue arrows represent the functional disconnections within this ''awareness network'' and with the thalami. The green area represents the relatively spared activity in the brainstem and hypothalamus. Adapted from Laureys, et al. (1999), NeuroImage.

Cortical activation to passive external stimulation

In brain death, external stimulation does not lead to any neural activation. In coma and VS patients, noxious stimulation was shown to activate only low-level primary cortices. Hierarchically higher-order areas of the pain matrix, encompassing the anterior cingulate cortex, failed to activate. Importantly, the activated cortex was shown to be isolated and functionally disconnected from the frontoparietal network, considered critical for conscious perception.

Similarly, auditory stimulation in VS was found to activate primary auditory cortices, but not higher-order, multimodal areas, from which they were disconnected (see Figure 4). In MCS, the activation was more widespread and there was an integrate functional connectivity between primary auditory cortices and the posterior temporal/temporoparietal and prefrontal associative areas.

Emotionally complex auditory stimuli, such as stories told by a familiar voice, lead to more widespread brain activation as compared to meaningless noise. Such context-dependent, higher-order auditory processing in MCS, often not assessable at the patient's bedside, indicate that content does matter when talking to these patients.

However, given the absence of a thorough understanding of the neural correlates of consciousness, functional neuroimaging results must be used with caution as proof or disproof of awareness in severely brain-damaged patients. Recently, Adrian Owen from Cambridge University in collaboration with our laboratory proposed a more powerful approach to identify 'volition without action' in noncommunicative brain-damaged patients. Rather than using passive external stimulation paradigms, patients were being scanned while asked to perform a mental imagery task. In one exceptional VS patient, task-specific activation was observed, unequivocally demonstrating consciousness in the absence of behavioral signs of consciousness. Interestingly, the patient subsequently recovered. Other studies also showed that VS patients with atypical brain activation patterns, after functional neuroimaging, showed clinical signs of recovery of consciousness – albeit sometimes many months later.

Treatment

To date, there are no 'standards of care' for therapeutic management in DOC. Many studies have been conducted under suboptimal or uncontrolled settings, and for that reason, no evidence-based recommendations can be made. MCS patients, however, were shown to benefit more than VS

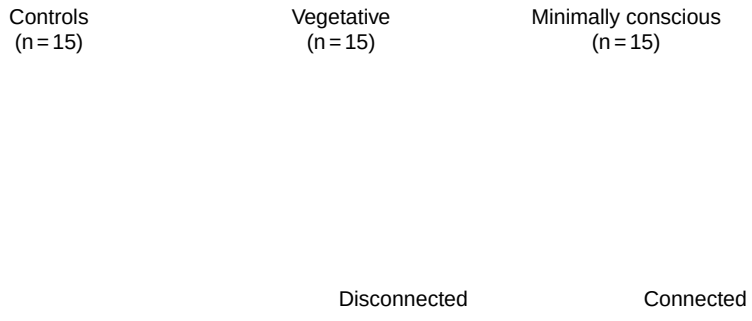


Figure 4 External stimuli still induce robust activation in primary sensory areas in vegetative patients. In the minimally conscious state, the activation is more widespread extending to multimodal associative areas. Functional connectivity studies (see 'Glossary') show that the activity of the primary cortex is isolated and disconnected from the rest of the brain, like the parahippocampal gyrus (red areas in the left inset). In the minimally conscious state, we observe a more integrated processing with preserved functional connectivity between low-level sensory areas and frontalparietal regions, which are thought to be involved in the emergence of conscious perception (blue areas in the right inset). Adapted from Boly, et al. (2004), Archives of Neurology.

after invasive treatment with DBS (see 'Glossary'). More particularly, bilateral thalamic stimulation, implanted over 6 years after acute trauma, has just been shown to cognitively improve an MCS patient, resulting in stimulation-related recovery of functional object use and intelligible verbalization. In the VS, despite some sparse evidence that DBS may benefit these patients, its effectiveness to this population is limited, mainly due to uncontrolled experimental settings. In any case, the technique awaits confirmation from studies on larger cohorts of patients, but illustrates that DBS in well-chosen patients, selected on the basis of functional neuroimaging results, can offer a real therapeutic option, at least in chronic MCS patients.

Pharmaceutical interventions with amantadine, mainly a dopaminergic agent, was shown to increase metabolic activity in a chronic MCS patient. Similarly, zolpidem, a nonbenzodiazepine sedative drug, may improve arousal and cognition in some brain-injured patients. However, placebo controlled randomized trials are needed before we making assertive conclusions about the effectiveness of the drug in DOC patients.

Conclusion

Currently, it is an exciting time for the study of DOC. The gray zone transitions between them, in the clinical spectrum following coma, are beginning to be better defined by adding powerful imaging methodology to bedside behavioral assessment. However, it should be stressed that these exciting developments are not yet a reality. The first obstacle to be overcome relates to the engendered ethical problems. An ethical framework that emphasizes balancing clear protections for patients with DOC along with access to research and medical progress is preferred. Moreover, most of the discussed areas of advances in coma science regard single case studies. Only large scale multicentric clinical trials will enable these research tools to find their way to a better evidence-based care for coma survivors.

Acknowledgments

Athena Demertzi is funded by the DISCOS Marie Curie research Training Network. Steven Laureys is senior research associate at the Belgian Fonds National de la Recherche Scientifique (FNRS). Melanie Boly is research fellow at FNRS. This research was funded by the European Commission, Mind Science Foundation, James McDonnell Foundation, French Speaking Community Concerted Research Action, and Fondation Médicale Reine Elisabeth.

See also: Ethical Implications: Pain, Coma, and Related Disorders; General Anesthesia.

Suggested Readings

American Congress of Rehabilitation Medicine (1995) Recommendations for use of uniform nomenclature pertinent to patients with severe alterations of consciousness. Archives of Physical Medicine and Rehabilitation 76: 205–209.

Boly M, Phillips C, Tshibanda L, et al. (2008) Instrinsic brain activity in altered states of consciousness: how conscious is the default mode of brain function? Annals of the New York Academy of Sciences 1129: 119–129.

Boly M, Faymoville ME, Schnakers C, et al. (2008) Preception of pain in the minimally conscious state with PET activation: an observational study. Lancet Neurology 7(11): 1013–1020.

Boveroux P, Bonhomme V, Boly M, et al. (2008) Brain function in physiologically, pharmacologically, and pathologically altered states of consciousness. International Anesthesiology Clinics 46(3): 131–146.

Demertzi A, Vanhaudenhuyse A, Bruno MA, et al. (2008) Is there anybody in there? Detecting awareness in disorders of consciousness. Expert Review of Neurotherapeutics 8(11): 1719–1730.

Di H, Boly M, Weng X, et al. (2008) Neuroimaging activation studies in the vegetative state: predictors of recovery? Clinical Medicine 8(5): 502–507.

Fins JJ, Illes J, Bernat JL, et al. (2008) Neuroimaging and disorders of consciousness: envisioning an ethical research agenda. American Journal of Bioethics 8(9): 3–12.

Giacino JT, Ashwal S, Childs N, et al. (2002) The minimally conscious state: Definition and diagnostic criteria. Neurology 58: 349–353.

Kubler A and Kotchoubey B (2007) Brain–computer interfaces in the continuum of consciousness. Current Opinion in Neurology 20: 643–649.

Laureys S (2005) The neural correlate of (un)awareness: Lessons from the vegetative state. Trends in Cognitive Sciences 9: 556–559.

Laureys S and Boly M (2008) The changing spectrum of coma. Nature Clinical Practice Neurology 4(10): 544–546.

Laureys S, Owen AM, and Schiff ND (2004) Brain function in coma, vegetative state, and related disorders. Lancet Neurology 3: 537–546.

Laureys S, Pellas F, Van Eeckhout P, et al. (2005) The locked-in syndrome: What is it like to be conscious but

paralyzed and voiceless. Progress in Brain Research 150: 495–511.
Laureys S, Perrin F, and Bredart S (2007) Self-consciousness in non-communicative patients. Consciousness and Cognition 16: 722–741.
Majerus S, Gill-Thwaites H, Andrews K, and Laureys S (2005) Behavioral evaluation of consciousness in severe brain damage. Progress in Brain Research 150: 397–413.
Owen AM, et al. (2006) Detecting awareness in the vegetative state. Science 313: 1402.
Posner J, Saper C, Schiff N, and Plum F (2007) Plum and Posner's Diagnosis of Stupor and Coma. Oxford University Press.
Schiff ND, Giacino JT, Kalmar K, et al. (2007) Behavioral improvements with thalamic stimulation after severe traumatic brain injury. Nature 448: 600–603.
Schnakers C, Majerus S, Giacino J, et al. (2008) A French validation study of the Coma Recovery Scale-Revised (CRS-R). Brain Injury 22(10): 786–792.
Schnakers C, Ledoux D, Majerus S, et al. (2008) Diagnostic and prognostic use of bispectral index in coma, vegetative state and related disorders. Brain Injury 22(12): 926–931.
Schnakers C, Perrin F, Schabus M, et al. (2008) Voluntary brain processing in disorders of consciousness. Neurology 71(20): 1614–1620.
The Multi-Society Task Force on PVS (1994) Medical aspects of the persistent vegetative state (1). The New England Journal of Medicine 330: 1499–1508.
The Quality Standards Subcommittee of the American Academy of Neurology (1995) Practice parameters for determining brain death in adults (summary statement). Neurology 45: 1012–1014.
Vanhaudenhuyse A, Giacino J, Schnakers C, et al. (2008) Blink to visual threat does not herald consciousness in the vegetative state. Neurology 71(17): 1374–1375.
Voss HU, Uluc AM, and Dyke JP (2006) Possible axonal regrowth in late recovery from the minimally conscious state. The Journal of Clinical Investigation 116: 2005–2011.

Relevant Websites

http://www.comascience.org.

Biographical Sketch

Athena Demertzi, MSc, PhD student, graduated from the Faculty of Psychology at the Aristotle University of Thessaloniki, Greece in 2005. Soon after, she pursued her research master's in cognitive neuroscience, neuropsychology, and psychopathology, at Maastricht University, The Netherlands, where she specialized in the field of neuropsychology. During her master's, she conducted her research internship at the Blixembosch Rehabilitation Centre, Eindhoven, The Netherlands, where she studied self-awareness deficits in everyday life following brain injury. She graduated in August 2007, and next joined the Coma Science Group as an early stage researcher appointed by the Marie Curie Research Training Network 'DISCOS' – Disorders and Coherence of the Embodied Self. Under the supervision of Steven Laureys, she investigates the neural basis of the elementary personal identity in patients with altered states of consciousness, such as vegetative and minimally conscious patients.

Steven Laureys, MD, PhD, is a senior research associate at the Belgian National Fund of Scientific Research (FNRS) and Clinical Professor at the Department of Neurology, Sart Tilman Liège University Hospital. He graduated as a medical doctor from the Vrije Universiteit Brussel, Belgium. While specializing in neurology he entered his research career and obtained his MSc in pharmaceutical medicine working on pain and stroke, using *in vivo* microdialysis and diffusion magnetic resonance imaging (MRI) in the rat (1997). Drawn by functional neuroimaging, he moved to the Cyclotron Research Center at the University of Liège, Belgium, where he obtained his PhD (2000) and his 'thèse d'agrégation de l'enseignement supérieur' (2007), studying residual brain function in coma, vegetative, minimally conscious, and locked-in states. He is board-certified in neurology (1998), and in palliative and end-of-life medicine (2004). A recipient of the William James Prize (2004) from the Association for the Scientific Study of Consciousness (ASSC) and the Cognitive Neuroscience Society (CNS) young investigator award (2007), he recently published *The Boundaries of Consciousness* (Elsevier 2005) and The Neurology of Consciousness (Academic Press 2009). He nowadays leads the Coma Science Group at the Cyclotron Research Centre at the University of Liège, Belgium.

Melanie Boly, MD, PhD student, is currently a research fellow at the Belgian National Funds for Scientific Research (FNRS) and Neurologist in training at the University Hospital CHU Sart Tilman. Under Steven Laureys' supervision, she performed several studies comparing auditory and noxious stimuli cerebral processing in minimally conscious and vegetative state patients. In collaboration with the team of Adrian Owen in Cambridge, she also elaborated a method to assess the presence of voluntary brain activity, and thus of consciousness, in noncommunicative, brain-injured patients. This method has already proven to be of potential interest in the early detection of signs of awareness in patients previously diagnosed as being in a vegetative state. Her interests include the study of recovery of neurological disability and of neuronal plasticity by means of multimodal functional neuroimaging (EEG-fMRI, PET, and MEG), and behavioral assessment in severely brain-damaged patients with altered states of consciousness.

Concepts and Definitions of Consciousness

D M Rosenthal, City University of New York, New York, NY, USA

Introduction

The term 'consciousness' is used in several ways: to describe a person or other creature as being awake and sentient, to describe a person or other creature as being 'aware of' something, and to refer to a property of mental states, such as perceiving, feeling, and thinking, that distinguishes those states from unconscious mental states. Distinguishing these different concepts of consciousness is crucial in evaluating the major theories of what it is for a state to be conscious. Among those are first-order theories, on which a mental state is conscious if being in that state results in one's being conscious of something; global-workspace theories, on which a state is conscious if it's widely available for mental processing; inner-sense theories, on which a state is conscious if one senses or perceives that state by way of a special inner faculty; and higher-order-thought theories, on which a state is conscious if one is aware of that state by having a thought about it. We will consider the advantages and shortcomings of these theories and variants of them.

Concepts of Consciousness (I)

The ubiquity of consciousness in human life and mental functioning makes it easy to overlook that the term 'consciousness' is used for three distinct phenomena. Though related in various ways, these phenomena are different, and distinguishing them is important both conceptually and theoretically.

The term 'conscious' is used most frequently to refer to the condition of people and other creatures when they are awake and responsive to sensory stimulation. A creature lacks consciousness in this first sense when it is asleep, anaesthetized, in a coma, and so forth. The main concern with this kind of consciousness is to explain in biological terms the difference between creatures' conscious and unconscious conditions. Important progress has been made on that front, for example, by Giulio Tononi and colleagues and by Steven Laureys. Because consciousness of this sort is a property of creatures, it is convenient to refer to it as creature consciousness.

A second important phenomenon we call consciousness is a creature's being conscious, or aware, of something. There are two ways creatures are conscious of things. A person or other animal is conscious of an object by seeing, hearing, or touching it, or sensing it in some other way. But one is also conscious of something, even without sensing that thing, if one has a thought about it as being present to one, that is, a thought that represents that thing as being in one's immediate environment. Because we describe this phenomenon by reference to a grammatical object, we may call it transitive consciousness. Explaining transitive consciousness consists in explaining what it is for a thought to be about something and what it is for a perception or sensation to be of something.

A third phenomenon is more controversial in nature, and is the subject of much recent scientific and philosophical literature. We are conscious of various things in virtue of our having perceptions of them or thoughts about them. But those perceptions and thoughts can themselves be conscious or not conscious. Subliminal perception is an example of nonconscious perceiving, and it is widely accepted that many thoughts occur nonconsciously as well, that is, outside our stream of consciousness. Since this phenomenon is a property of mental states, rather than of creatures that are in those states, it is convenient to call it state consciousness.

Mental states, such as thoughts, perceptions, and feelings, were until the latter part of the nineteenth century seldom described as being conscious or not conscious. Theorists before that time tended to regard mental states as invariably conscious; so it was idle to mark a distinction between mental states that are conscious and

those that are not. Thus Descartes held that "we cannot have any thought of which we are not aware at the very moment when it is in us" (Fourth Replies), echoing Aristotle's claim in that "if we perceive, we perceive that we perceive, and if we think, that we think" (Nichomachean Ethics 1170a32).

Brentano, whose University of Vienna lectures Freud attended for a time, maintained as late as 1874 that all mental states are conscious. Still, he broke ranks with previous tradition in his Psychology from an Empirical Standpoint by denying that there is any contradiction in the notion of a mental state that is not conscious, thereby opening the door to the possibility that mental states might after all sometimes not be conscious.

As long as consciousness was widely thought to be essential to mentality, little attention was given to explaining why that is so, or even to explaining what it is for states to be conscious. Brentano's breakthrough, very likely noted by Freud, was to focus attention on those questions. And Brentano himself offered an explanation both of what it is for states to be conscious and of why, as he held, all mental states are conscious.

Theoretical discussions of consciousness often fail to be clear which of these three phenomena are at issue. This is sometimes innocuous, but running these phenomena together also sometimes causes theoretical difficulty. Thus conflating creature consciousness with the consciousness of mental states may lead one to hold that the mental states a creature is in when that creature is conscious are themselves all conscious states. But, since mental states occur without being conscious, we have no reason to think that all the mental states a conscious creature is in are conscious states. Perhaps, indeed, the mental states of some creatures, such as lizards and frogs, are never conscious, even when those creatures are conscious; other creatures might only sometimes be conscious without any of their mental states being conscious. A creature's being conscious does not by itself show that its mental states are conscious.

Concepts of Consciousness (II)

Mental states have two characteristic types of mental property. One is intentional content, which represents things in a way that can be expressed by a full sentence. States with intentional content also have mental attitude that one holds toward that content, such as mental affirmation, doubt, wonder, and so forth. In contrast with those intentional properties, there are various mental qualities, which are characteristic of bodily and perceptual sensations. Each mental quality has a particular location in a quality space that is characteristic of the relevant sensory modality, effect, a quality space of mental colors, sounds, and the like; this account has been developed by Clark and by Rosenthal. Some states, such as perceptions and emotions, have both intentional and qualitative properties; the mental properties of other states, such as thoughts and sensations, are of only one of the two types.

When a state with qualitative character is conscious, there is, as Thomas Nagel has put it, something it's like for one to be in that state. By contrast, we do not typically say that there is something it's like for one consciously to think some particular thing, or to doubt it, though some have contested that. The consciousness of purely intentional states is in any case intuitively distinct from that of states that have some qualitative character.

Pressing in part on that intuitive difference, Block has distinguished two ways in which states can be conscious. A state is access conscious if its content is "poised to be used as a premise in reasoning,. . . [and] for [the] rational control of action and. . . speech". By contrast, a state exhibits phenomenal consciousness if there is something it's like to be in that state. In part because qualitative consciousness seemingly differs from the consciousness of nonqualitative states, Block's distinction has been influential both in the philosophical and in the scientific literature.

Block regards these two types of state consciousness as conceptually independent; access and phenomenal consciousness reflect two distinct concepts of state consciousness. Block has more recently argued in addition that the two occur independently and have distinct neural realizations. If so, distinct theoretical treatments are required for the two.

The notion of access consciousness plays a central role in so-called global-workspace theories, developed by Baars, Dehaene and Naccache, and Tononi, on which a state is conscious if it has the

potential for having a global effect on memory, behavior, and other psychological functioning. As Dennett vividly puts it, ''[c]onsciousness is cerebral celebrity.'' Such global effects are, moreover, thought by some to be the function that consciousness has, in virtue of which it is useful for an organism's mental states to be conscious. The concept of access consciousness in effect purports to isolate a kind of consciousness by reference to its mental function.

The potential for global effects on mental functioning and behavior does sometimes accompany the consciousness of mental states, but that is arguably not always so. Conscious peripheral perceptions have little if any global effect, and many conscious passing thoughts and desires also have none.

Conversely, much thinking occurs without being conscious, as with the nonconscious thoughts that are steps in much problem solving. Nonetheless, these nonconscious thoughts sometimes have a significant effect on mental functioning. So it is unclear that a state's potential to have global effects coincides with its being conscious. And if it does not, such potential would not then be a distinctive function that conscious states serve in contrast to mental states that are not conscious.

Questions can also be raised about Block's notion of phenomenal consciousness. Block explains phenomenal by saying that there is always something it's like to be in a phenomenally conscious state. But he also argues that phenomenal consciousness occurs in connection with subliminal vision, extinction, and other clinical conditions in which the relevant states are not in any intuitive way conscious. So it is tempting to construe Block's phenomenal consciousness as simply a matter of a state's having mental qualities, independent of whether that state is conscious.

Conscious qualitative character is intuitively such a distinctive mental phenomenon that it has been thought by some not to be susceptible of any informative explanation. Thus Levine has argued that even if brain function subserves qualitative states, there is an explanatory gap that may make it impossible to explain why particular brain events result in the particular qualitative states they do, or indeed in any at all. Chalmers argues similarly, maintaining that this is the Hard Problem of consciousness.

It may be, however, that whatever explanatory difficulty now confronts us is not ineluctable, but is rather due simply to our current state of knowledge about qualitative character, and its relation to brain function. Levine urges that our understanding of the neurological basis of qualitative consciousness can never be firm and complete in the way our current understanding is of the chemical nature of water. But it may be that as our understanding of the neutral basis of qualitative consciousness approaches the completeness and theoretical sophistication of current chemistry, the intuitive contrast in explanatory adequacy of the two cases will disappear.

Another factor that seems to block any informative explanation of qualitative consciousness is the view of some theorists that we can know about qualitative properties only by the way we are conscious of them. This view reflects the traditional idea, inspired by Descartes, that consciousness gives us infallible or in any case incorrigible access to our own mental states, and indeed that this access exhaustively reveals their mental nature.

The view that we can know about mental qualities only by way of consciousness underlies the familiar view, advanced by Locke in *An Essay Concerning Human Understanding*, that it is conceivable that the mental quality two individuals have on seeing the same object differ in undetectable ways. And it is sometimes held to be conceivable that an individual physically and functionally identical to us might undetectably lack mental qualities altogether. Such occurrences would be undetectable only if one's consciousness of one's own mental qualities were the only way to gain knowledge about them, which would block any explanation of mental qualities in terms other than consciousness. In particular, it would prevent explaining mental qualities in terms of their neural basis.

But it is arguable that mental qualities are individuated by their location in a quality space that corresponds to the quality space of the perceptible properties accessed by the relevant sensory modality. Thus mental red, for example, is individuated by its relation to other mental color qualities, corresponding to the relations physical red has with perceptible physical colors. If so, mental qualities are not individuated after all by one's individual access to those qualities. The conceptual

ties between families of mental qualities and perceptible physical properties would then make undetectable inversion and absence of mental qualities conceptually incoherent. And there would then be little reason to see an explanatory gap as inevitable.

Both intentional and qualitative states often occur consciously. But some theorists hold that whereas intentional states also occur without being conscious, that is not so for qualitative states. And that view leads some to use the term 'consciousness' to refer simply to conscious qualitative character.

But even if all qualitative states were conscious, the property of being conscious would only be one aspect of their mental nature. As G. E. Moore noted, conscious qualitative states differ among themselves in respect of mental quality, though they have in common the property of being conscious. Consciousness is accordingly a distinct property from any mental quality. Focusing on what it's like for one to be qualitative states yokes together these two aspects of their mental nature, making it seem that they cannot occur independently. But each does occur apart from the other, since nonqualitative, intentional states are sometimes conscious, and qualitative states sometimes occur without being conscious.

Concepts and Theories

Conflating distinct concepts of consciousness can result in confused theories. Block has urged that this sometimes happens when theorists fail to distinguish access from phenomenal consciousness. Failing to distinguish creature, transitive, and state consciousness can also have important consequences for theories of consciousness.

As already noted, failing to distinguish creature consciousness from state consciousness may encourage the view that mental states never occur without being conscious. And that may tempt one to identify being conscious with being mental, and so to hold that there is nothing more to a state's being conscious than its simply being mental. And since many mental states are states in virtue of which one is conscious of things, identifying consciousness with mentality will invite the view that a state's being conscious consists simply in its being a state in virtue of which one is conscious of something. This has come to be known as a first-order theory of consciousness, best exemplified by Dretske.

Holding all mental states to be conscious encourages a first-order theory of consciousness. Nonetheless, traditional thinkers from Aristotle to Descartes, Locke, and Brentano did not endorse that view. As noted in the section 'Concepts of consciousness (I),' it was rare until Brentano's time to describe mental states as conscious at all. Even though Descartes and Locke were plainly writing about the property we describe as a state's being conscious, they did not say that our mental states are all conscious, but rather that we are conscious of all our mental states.

The difference is significant. On a first-order theory, a state's being conscious is its being in a state of transitive consciousness, a state such that one's being in that state constitutes being conscious of something. What we describe as a state's being conscious was traditionally described in terms of one's being conscious of that state. Because it appeals to transitive consciousness, we can refer to the view that a state's being conscious consists in one's being conscious of that state as the Transitivity Principle (TP). And because being conscious of a state involves some higher-order awareness, theories that adopt TP are known as higher-order theories.

The contrast between higher-order and first-order approaches marks a major theoretical divide in explaining consciousness. On a higher-order theory, a state is conscious simply if one is transitively conscious of it; on a first-order view, a state is conscious instead if it is itself a state of transitive consciousness.

Each approach faces difficulties that the other avoids. Because first-order theories classify as conscious any state in virtue of which one is conscious of something, such theories may be unable to account for the occurrence of nonconscious, subliminal perception and thinking that intuitively fails to be conscious.

There is extensive evidence that perceiving does sometimes fail to be conscious. As Anthony J. Marcel, Bruno G. Breitmeyer and Haluk Ögmen, and Zoltan Dienes and Josef Perner, have shown,

subjects in masked-priming experiments are presented with a visual stimulus followed at a specific interval by another. Without the second stimulus, subjects would see the first stimulus consciously; but when the second does occur it masks the first, leading subjects to see consciously only the second. Nonetheless, there is evidence that subjects do after all see the first stimulus, since it primes them for enhanced performance in various tasks, including largely correct guesses about that stimulus. Subjects see the first priming stimulus, but not consciously.

In blindsight, the study of which was pioneered by Lawrence Weiskrantz, subjects with damage to an area of primary visual cortex report not seeing stimuli presented in the area of their visual field corresponding to cortical damage, but again their guesses about these stimuli are well above chance. Subjects evidently see the stimuli, though the seeing is not conscious.

The view that all mental states are actually conscious may be problematic in another way. Explaining what it is for a state to be conscious plainly must appeal to mental properties of that state. But if a state's being mental coincides with its being conscious, any explanation of consciousness in terms of mentality risks being circular.

First-order theorists would reply that, since consciousness does coincide with mentality, we explain what it is for a state to be conscious by explaining what it is for that state to be mental. Such theorists would also point to difficulties that higher-order theories seem to encounter. Most pressing, they urge, is the possibility of inaccurate higher-order awareness. The way we are aware of things is not always accurate; so if a state's being conscious consists in one's being aware of that state, perhaps that higher-order awareness can itself fail to be accurate. But it is unintuitive to suppose that consciousness could be inaccurate; with consciousness, many maintain, appearance and reality coincide.

Another challenge for higher-order theories is to explain why any such higher-order awareness occurs at all. Perhaps that awareness serves some function, so that having that awareness confers some adaptive advantage. But it is unclear what advantage such higher-order awareness might confer. First-order theories avoid this challenge, since they hold that a state's being conscious consists in its being in a state of being conscious of something. And it is plain that being conscious of things is crucial for a creature's successful functioning. The remaining discussion will examine in more detail the issues that divide these two approaches.

First-Order Theories

An apparent advantage of first-order theories is that subjectively we seldom seem to have the sort of higher-order awareness that higher-order theories posit. John Searle has recently appealed to this in denying that we ever observe our mental states, or that we even could. When we see something, the seeing and the thing seen are distinct, but Searle insists that this distinction does not apply to our awareness of our own mental states.

Observation is a frequent model for how we are aware of our own mental states; as Locke famously put it, ''[c]onsciousness is the perception of what passes in a Man's own mind.'' But observation is not the only way we might be aware of our mental states, and intuitively it is the least likely. More important, since higher-order theories countenance mental states that are not conscious, whatever higher-order awareness they posit need not itself consist in states that are conscious. And if those higher-order states are not conscious, it will seem subjectively that we have no such higher-order awareness. The higher-order awareness such theories appeal to is a theoretical posit, not something to be found in the phenomenological appearances.

A first-order theorist might insist that no higher-order mental state could result in one's being conscious of the first-order state it is about unless that higher-order state is itself conscious. That would decisively undermine the higher-order approach, since it would result in a vicious regress of higher-order awarenesses; each higher-order awareness, to be itself conscious, would require a higher-order awareness of it.

But that argument presupposes the first-order view that a state's making one transitively conscious of something coincides with its being a conscious state. Subliminally perceiving things results in one's being conscious of those things;

otherwise such perceiving would not affect one's mental functioning. Indeed, subliminal perception often results in qualitative discriminations of just the sorts we make by consciously perceiving things, as shown by Breitmeyer and Ögmen. States need not themselves be conscious to result in our being aware of things.

Consciousness is simply a matter of the phenomenological appearances. So it may seem that a higher-order awareness that is not itself part of those phenomenological appearances cannot explain consciousness. But in appraising any explanation, we must distinguish between what is to be explained and the considerations in virtue of which the explaining proceeds. Any satisfactory theory of consciousness must do justice to the phenomenological appearances. But it does so by explaining those appearances; the considerations that do the explaining need not themselves be limited to those phenomenological data, any more than we explain ordinary macroscopic phenomena solely by appeal to such macroscopic phenomena. Indeed, explaining the phenomenological appearances solely by appeal to those appearances would be circular and uninformative.

Dretske has advanced an elegant argument in support of a first-order approach. It often happens that we see two scenes that differ in some slight way, though without being conscious that they differ. Perhaps the scenes are alike except that one has ten trees, one of which is missing in the other. Nonetheless, one may consciously see the entire scenes, and so consciously see the tenth tree in the scene in which it occurs. So one has a conscious visual experience of the tenth tree. But despite that, one is not conscious of the experience of the tenth tree, since one is unaware of the two scenes differing. Dretske concludes that conscious experiences occur of which one is not conscious.

Scenes that differ in some unnoticed way are common in everyday experience. Still, Dretske's argument seems not to be decisive against TP and higher-order theories. One can be conscious of something in one respect and not in another. So one might in Dretske's example well be conscious of the experience of the tenth tree only as a part of the overall experience of the scene, though one is not conscious of the experience of the tenth tree as the way in which the two overall experiences differ. Since one could be conscious of the experience of the tenth tree, though not in the way Dretske argues against, Dretske's example does not establish that an experience can be conscious despite one's not being conscious of it.

Visual presentations that differ in some salient way that is not consciously noticed are the focus of experimental work on change blindness, in which salient changes occur that subjects do not consciously see, as shown by James Grimes and by Daniel Simons and Ronald Rensink. But if we are in some way blind to such unnoticed changes, perhaps we do not, as Dretske maintains, always consciously see the things in virtue of which two scenes differ in unnoticed ways. Dretske has recently addressed one experimental paradigm, developed by Grimes, in which the unnoticed changes occur during saccades, arguing that since visual input during saccades does not reach cortical areas, subjects are not blind to things that change, but only to the differences that result from those changes.

There is, however, a crucial way in which subjects are indeed blind to the changed objects. In one case of change blindness, a large parrot switches back and forth between being red and green. Dretske acknowledges that what subjects see corresponds to the actual stimulus; when the parrot is red subjects see red and when it is green they see green. But even when the parrot's color changes, there is often no change in what it's like for subjects; Grimes's subjects often continue seeming to see red when the parrot is green. These cases exhibit a divergence between the seeing and how we are conscious of it, which points toward TP and higher-order theories.

First-order theories, by arguing that a state is conscious if it is a state of being conscious of something, seem to leave no room for subliminal perceiving and nonconscious thinking. Dretske has also addressed this issue. Refining his first-order view, Dretske adds as a condition for perceiving's being conscious that the individual can cite the perceived fact as a justifying reason for doing something. This rules out subliminal perceiving, in which subjects deny perceiving anything and so cannot cite what they perceive in any way, much less as a reason for action.

Subjects' inability to give a justifying reason, however, may not show that they have no such

reason, but only that they have no conscious reason. We often do things for reasons that do not figure in our stream of consciousness; this is evident from cases in which such reasons later come to be conscious. So Dretske cannot accommodate subliminal perceiving without explaining how conscious reasons for doing things differ from reasons that are not conscious (see the section 'Language and function').

There is another empirical challenge for first-order theories. Libet and Haggard have shown that the cortical event that corresponds to subjects' deciding to make a basic movement occurs significantly before they are aware of that deciding. The most straightforward interpretation of these findings is that acts of deciding occur prior to those decidings coming to be conscious. If so, the mental state and its being conscious are distinct occurrences, contrary to first-order theories.

Dennett has sought to undermine higher-order theories by arguing that the hierarchy of mental states such theories posit is psychologically unrealistic. Being in a mental state results in things' seeming a certain way to one. But there is no difference, Dennett urges, between how things seem to one and how they seem to seem. So there cannot be higher-order states in virtue of which it seems to us that we are in particular first-order states. This conclusion points toward the first-order approach.

But Dennett's view again has difficulty with subliminal perceiving, since that in effect consists of something's seeming to one though it does not seem to one that it does. Conscious perceiving, by contrast, is perceiving in which it does seem to one that something seems to be some particular way. Dennett urges that these subliminal cases are not genuine perceiving at all, but mere ''events of content-fixation.'' But since having content is a mark of the distinctively mental, it may be more reasonable to accept the subliminal cases as being genuine perceiving, and thereby a second level of awareness in conscious perceiving.

Global-workspace theories, on which a state is conscious if it has the potential to affect a broad range of mental functioning and behavior, are in effect a type of first-order theory, since they appeal to no higher-order states. Robert Van Gulick has sought to combine global-workspace theory with aspects of a higher-order theory, arguing that a state's having global connections results in one's being conscious of oneself. But this is at best a qualified type of higher-order theory, since being conscious of oneself need not by itself involve being conscious of any particular mental state.

One could combine global-workspace theory with a higher-order approach by stipulating that the global connections a conscious state has must include a higher-order awareness of that state. Still, a theory must specify what it is for a mental state to be conscious, and it is not obvious whether, on such a hybrid theory, a state's being would be a matter of the higher-order awareness or of the global connections. If the global ties were seen as responsible for consciousness, that would still be a first-order explanation of consciousness.

Higher-Order Theories (I)

The alternative to the first-order approach is a higher-order theory that conforms to TP, on which a state's being conscious consists in its being a state of which one is conscious. As noted in the section 'Concepts and theories,' this approach dominated traditional discussion of consciousness from Aristotle through Descartes and Locke to Brentano. The higher-order theory most often advanced has been the inner-sense theory, developed by Armstrong and by Lycan, on which we are aware of our conscious states by sensing or perceiving those states.

This way of implementing TP has a number of advantages. For one, we are conscious of things most often by sensing and perceiving them. We are also conscious of something if we have a thought about it as being present; but sensing and perceiving are what come first to mind in connection with being conscious of things. Indeed, it is very likely by analogy with sensing that we regard having a thought about something as being conscious of that thing only if the thought represents it as present to one.

A second reason inner sense is inviting has to do with qualitative consciousness. If qualitative states are conscious in virtue of our perceiving them, that may help explain qualitative consciousness, since perceiving is itself qualitative. In particular, the mental qualities we are conscious of differ in

myriad fine-grained ways, seemingly outstripping the ability of concepts to capture those differences. So perhaps the differences in virtue of which we are conscious of our qualitative states can be captured only by higher-order perceptual awareness, which itself involves mental qualities.

Another advantage of inner sense has to do with a condition that higher-order awareness must satisfy. Not all higher-order awareness of one's own mental states results in those states' being conscious. One may be aware of one's own mental states by theorizing about oneself or by taking the word of somebody who knows one very well. But these kinds of higher-order awareness do not by themselves result in the relevant states' being conscious. The higher-order awareness must, it seems, be immediate in some way; as Descartes put it, "I use [the term 'thought'] to include everything that is within us in such a way that we are immediately aware of it" (Fourth Replies). Inner sense captures this constraint, since perceiving something seems subjectively to result in one's being immediately conscious of the objects perceived.

Finally, it is tempting to explain why any higher-order awareness occurs in the first place by appeal to the usefulness of our monitoring our first-order mental states. Since perceiving monitors our environmental and body conditions, it is arguable that the higher-order awareness in virtue of which such monitoring occurs is very likely perceptual in nature.

Despite these advantages, a number of difficulties face any inner-sense theory. The qualitative character of sensing and perceiving underlies several of those advantages, but is also the source of the principal difficulty. Though higher-order awareness results in our being of first-order mental qualities, we have no reason to think that our higher-order awareness itself has any qualitative character. We never subjectively encounter higher-order mental qualities, in addition to those of our first-order qualitative states. Perhaps that is only because our higher-order awareness is seldom itself conscious, so that we are not conscious of our higher-order mental qualities. But sometimes we are introspectively aware of our conscious states, conscious of them, that is, in a way that is reflective and attentive. When we are, we are conscious also of our higher-order awareness of first-order conscious states; but even then we are never conscious of higher-order mental qualities.

Even though our higher-order awareness evidently lacks qualitative character, it might resemble perceiving in other significant ways. Thus Lycan has recently argued that we attend to our conscious states much as we attend to the things we perceive. And he urges that the voluntary control we have over which perceptual states in our sensory fields we are conscious of more closely resembles the voluntary control we have in perceiving than in thinking about things.

It is unclear that we have much voluntary control over our awareness of our conscious states. But that aside, we arguably have as much control over our thought processes as over our perceiving. And thinking about things allows us to focus attention on them no less than perceiving them. It is questionable whether any nonqualitative aspects of perceiving will sustain a compelling analogy with the higher-order awareness we have of our conscious states.

Though the appeal to higher-order qualitative character is inviting in explaining first-order conscious qualities, it is also very likely circular; higher-order mental qualities would need explaining no less than their first-order counterparts. And, though higher-order mental qualities would capture all the first-order differences among mental qualities that we are conscious of, a purely conceptual form of higher-order awareness may be able to do that as well (see the section 'Higher-order theories (II)').

We monitor environmental and bodily conditions perceptually, but that might not be necessary for one's own mental states. It would be enough for thoughts about those states would monitor those states if the states are causally implicated in leading to the thoughts. Indeed, feedback training can enable subjects to have reliable thoughts about their blood pressure and heart rate, based on visceral input that may involve no perceptual modality, and shown by findings in the 1970s by Brener and Jones and by Cinciripini, Epstein, and Martin. Subjects seem spontaneously to have reasonably reliable thoughts about what their heart rate or blood pressure is.

More important, consciousness does not always play any monitoring role whatever. As Nisbett and Wilson showed in their well-known work in the

1970s, we are sometimes conscious of ourselves as being in various mental states that we are not actually in. Such confabulatory consciousness, which also occurs in various dissociative disorders, evidently serves to make our mental lives seem more sensible or otherwise acceptable to ourselves or to others. In these cases, which may not be all that rare, our being conscious of ourselves as being in particular mental states does not serve to monitor our actual mental functioning, thereby undermining the monitoring analogy with perception.

Higher-Order Theories (II)

Aristotle held a mixed higher-order theory, on which our higher-order awareness is perceptual for conscious perceiving, but consists of higher-order thinking for conscious thought. According to inner-sense theory, higher-order awareness is perceptual for all our conscious states, both qualitative and intentional. But perhaps purely intentional higher-order states will work at least as well for conscious intentional and qualitative states alike.

In what is arguably its most straightforward form, developed by Rosenthal, this theory posits distinct, occurrent higher-order thoughts (HOTs), in virtue of which we are aware of our conscious states. For one to be aware of one's first-order states, these HOTs must have the content that one is, oneself, in the state in question. As with other thoughts, HOTs could occur in creatures without language; nonlinguistic creatures often express thoughts and other purely intentional states, thereby providing evidence for their occurrence.

HOTs have many of the advantages of inner sense. As noted in the section 'Higher-order theories (I),' HOTs can subserve such monitoring as actually occurs if the monitored state is causally implicated in the occurrence of a HOT. More importantly, HOTs can accommodate the apparent immediacy with which we are aware of our conscious states.

It is worth noting that the traditional claim that our consciousness of our mental states is immediate rests solely on subjective appearances. And that warrants holding only that such consciousness seems to be unmediated, not that nothing actually mediates between the mental states and the corresponding higher-order awareness. Indeed, that is the situation with perceiving; it seems subjectively that nothing mediates between perceiving and what we perceive even though there is much that does mediate.

HOTs can yield a higher-order awareness that is no less spontaneous and subjectively unmediated. It may be that some inference, observation, or other mental processes lead to a HOT that one is in a particular state. But if one is not conscious of those processes, one's awareness of the first-order state will be subjectively direct and immediate. Moreover, HOTs would seldom themselves be conscious; they would be conscious only if there was a yet HOT about them. And when HOTs are not conscious, the higher-order awareness that results would appear spontaneous and unmediated.

As noted in the section 'Higher-order theories (I),' mental qualities differ in ways that outstrip our concepts for specific qualities. We can consciously distinguish vastly more mental color qualities, for example, than we have concepts for the specific qualities. But concepts for specific qualities are not needed here. When asked to describe specific colors, we typically do so comparatively, saying that a particular color is darker than another or closer to one color for which we have a name than to another. And we do the same in describing specific mental qualities. We use comparative concepts to fine-tune our ability to describe, and hence conceptualize, particular mental qualities. This fits with the suggestion in the section 'Concepts of consciousness (II)' that we individuate mental qualities by their location in a quality space homomorphic to that of the perceptible properties accessed by the relevant modality. We make extensive use of comparative concepts to locate specific mental qualities and the corresponding perceptible properties in their respective quality spaces, thereby individuating them in conceptual terms. This idea is developed by Rosenthal.

There is compelling evidence that we do individuate mental qualities comparatively. As Raffman has noted, our ability to determine whether two simultaneously presented qualities are the same or how they differ is far more accurate than our ability to identify, recognize, or remember the very same qualities when they occur successively. The best explanation is that we are conscious

of very fine qualitative similarities and differences comparatively; having comparisons available greatly enhances our discriminative abilities.

Higher-order perceiving cannot explain the lighted-up qualitative character of conscious mental qualities, since the higher-order mental qualities would themselves need to be explained. But since HOTs are purely intentional states and so have no qualitative character, it may seem intuitively implausible that they could result in there being something it's like for one to be in conscious qualitative states.

There is reason, however, to think that HOTs can actually do this. We sometimes become conscious of differences among mental qualities only when we have words to hang those differences on, as with the different mental qualities that result from tasting various wines. Such mental taste qualities may, at an early stage, be consciously indistinguishable. But we sometimes come, upon learning suitable wine-tasting terms, to be conscious of the qualities as distinct. Learning new words reflects the learning of the concepts those words express, concepts that result in our being able to have more fine-grained thoughts about our mental qualities. Since purely intentional states about mental qualities can by themselves result in what it's like for one to be more fine-grained, HOTs can presumably result in there being something it's like for one in the first place.

A theory that posits distinct, occurrent, HOTs is not the only type of theory that posits purely conceptual states to explain our higher-order awareness. Brentano advanced a theory on which that higher-order awareness is due to intentional content that is intrinsic to each conscious state. This approach has been more recently defended by Gennaro and Kriegel.

Intrinsicalism about higher-order awareness has many of the inviting features of first-order theories, for example, in squaring with the phenomenological sense that higher-order awareness seldom occurs. Intrinsicalism thereby seeks to combine the advantages of both first-order and higher-order approaches.

Intrinsicalism also promises to handle a problem some have raised for higher-order theories. A distinct higher-order perception or thought could misrepresent one's mental life, either by making one conscious of a state in a way that distorts its nature or by making one conscious of a state that one simply is not in. And Levine has argued that there is no principled answer to what it would be like for one in such a case. Would having a sensation of red along with a higher-order awareness of that sensation as green be subjectively like seeing red? Or would it be like seeing green? Intrinsicalism appears to help, since an intrinsic higher-order awareness might be unable to misrepresent the state of which it is a part.

But higher-order theories face no difficulty about such cases. What it's like for one on these theories is determined by the way the higher-order awareness represents the first-order state. Consciousness is a matter of mental appearance, that is, of how our mental lives appear to us, and on those theories that mental appearance is due solely to the higher-order awareness. And intrinsicalism could not help in any case. There is no reason why a state's higher-order content could not misrepresent that state.

Intrinsicalism also does not fit comfortably with Libet's and Haggard's results, on which we are conscious of states only slightly after those states themselves occur. Perhaps higher-order content arises slightly after the rest of the state, but the intrinsicalist must explain why that higher-order content counts as intrinsic.

Intrinsicalism may conform to the phenomenological sense that we seldom are conscious of any higher-order awareness. But as noted in the section 'First-order theories,' phenomenology determines only the psychological reality to be explained, not what theories should posit to do that explaining. Kriegel has argued that we are generally conscious of our higher-order awareness, but only peripherally. That is unlikely with many perceptions and thoughts that are themselves only peripherally conscious. But such peripheral consciousness of higher-order awareness could in any case occur equally with distinct HOTs.

There is, finally, a difficulty about the mental attitude of conscious intentional states. Wondering or doubting whether one is in a mental state does not result in one's being conscious of that state. The higher-order awareness must be one of mental affirmation. No intentional state, moreover, has more than one mental attitude; no state is a case of

both wondering and doubting or both doubting and mentally affirming. So when wondering or doubting is conscious, since the higher-order awareness must be a case of mental affirmation, it must be distinct from the wondering or doubting itself.

Carruthers has argued that we lack the cognitive and cortical resources to sustain occurrent HOTs for all our conscious states, and that occurrent HOTs would also confer no adaptive advantage that could explain the evolution of creatures with such HOTs. So he has developed a theory on which a state is conscious not when a HOT actually occurs, but when it is simply disposed to occur.

The dispositionalist HOT theory is intuitively inviting. We are seldom conscious of our higher-order awareness, but focusing on the state often results in being conscious of that awareness. So perhaps higher-order awareness does not actually occur whenever a state is conscious, it is disposed to occur on attending to it. But the better explanation is that we simply are not conscious of the higher-order awareness that accompanies ordinary conscious states. And again, the phenomenological appearances should not in any case guide what posits a theory makes.

Since we do not now know what cortical resources subserve specific thoughts, we have no reason to think we lack the cortical resources needed for occurrent HOTs. And we may need fewer resources than it seems subjectively. Conscious visual perception seems equally acute throughout our visual field. But as Dennett has stressed, that subjective sense is confabulatory; so we doubtless need far fewer HOTs for conscious parafoveal vision than for conscious central vision.

The dispositionalist theory, moreover, faces a difficulty in implementing TP, which motivates higher-order theories generally. Being disposed to have a thought about something does not make one conscious of that thing; so being disposed to have a HOT does not make one conscious of the mental state that HOT would be about.

Language and Function

As noted in the section 'First-order theories,' Dretske seeks to accommodate subliminal perceiving within a first-order framework by holding that perceiving is conscious only if one can cite what is perceived as a reason for doing something. But we can cite reasons only when they are conscious. So that suggestion requires being able to distinguish conscious from nonconscious reasons, which seems in turn to point toward a higher-order theory.

A standard test for a mental state's being conscious is that the individual can report being in that state. If we have good reason to think somebody thinks, perceives, or feels something but the person sincerely denies being in that state, we conclude that the thinking, perceiving, or feeling is not conscious. This test guides us in both in commonsense contexts and experimental psychology.

Higher-order theories can explain the reliability of this test. Every sincere speech act expresses an intentional state that has the same content as the speech act and a mental attitude that corresponds to the speech act's illocutionary force. So a sincere report that one is in a mental state expresses an assertoric thought that one is in that state, and the ability to report being in some mental state is the same as the ability to express a HOT that one is in that state. We can best explain why a state's being conscious coincides with one's ability to report that state by supposing that a state is conscious only when the HOT such a report would express is present. The reportability test for consciousness actually supports the HOT hypothesis.

Reporting a mental state reveals consciousness only if the report is subjectively noninferential; a report based on conscious inference or observation is compatible with the state's not being conscious. This fits with the requirement noted in the section 'Higher-order theories (II)' that HOTs themselves not rely on any inference of which one is conscious. The reportability test applies only to creatures with language and, indeed, the ability to talk about their mental states. But we can use the test to determine what is responsible for consciousness in that case, and then apply that to nonlinguistic creatures.

In the case of humans, there is an even closer tie between consciousness and speech. Whenever we verbally express a first-order thought, that thought is invariably conscious. By contrast, thoughts expressed only by nonverbal behavior often fail to be conscious. This may have led Descartes to

insist that nonlinguistic creatures have no conscious thoughts, since if they did they would express them verbally (letters to More and to Newcastle).

But the HOT theory can explain the tie in humans between consciousness and speech without appeal to Descartes's remarkably implausible view. Humans seldom note the difference between reporting a thought and verbally expressing it; one may not even recall a moment later whether one said that something is so or that one thinks it is. Whenever one says something, one might as easily have said that one thinks that thing.

Because we are disposed to report a thought whenever we verbally express it, we are also disposed to have the HOT that such a report would express. So, as Rosenthal has shown, verbally expressing a thought results in that thought's being conscious. Expressing a thought nonverbally does not dispose us to report having that thought; so nonverbally expressed thoughts often fail to be conscious.

On first-order theories, a state's being conscious consists in that state's making one conscious of something, which is pivotal for an organism's functioning. So, as noted in the section 'Concepts and theories,' first-order theories readily explain why mental states are conscious. Higher-order theories face a challenge on this score, since it is unclear what the function of higher-order awareness might be.

A higher-order theorist might reply that monitoring one's mental states serves an important function. Thus Armstrong urges that a state's being conscious enhances problem solving and planning. But monitoring may not significantly enhance problem solving and planning, since those processes rely largely on causal connections among first-order thoughts and desires, which in turn reflect the intentional content of those states. Indeed, those processes can be more successful when they are not conscious, as shown in recent work by Dijksterhuis and colleagues and by Leib Litman and Robert Reber.

In creatures with suitable linguistic ability, reportability coincides with a state's being conscious. But such reportability confers no function that the state lacks when it occurs nonconsciously. Rather than report a state, an individual can convey being in that state by expressing it verbally; reporting adds no relevant information. And the consciousness of verbally expressed thoughts in the case of humans hinges on our being strongly disposed to a thought whenever we express it verbally, which itself seems to serve no particular function.

If a state's being conscious adds little significant function to what the state has occurring nonconsciously, adaptive value cannot explain why many states come be conscious. But a nonadaptive explanation is possible. Creatures sometimes come to be aware that a perceptual error has occurred. That realization involves a creature's having the thought that it was in an erroneous state, say, of the sort that occurs when, for example, a red object is in front of it. And that in time will dispose the creature to have such thoughts whenever it is in such states, and thereby to be conscious of those states.

Purely intentional states require a different account. Consider creatures that can report their own thoughts, but only by inferring from their behavior what it is likely that they are thinking. Since their reports always rely on observation and inference, the thoughts they report this way are not conscious.

But in time the difference between reporting their thoughts and expressing them verbally will come to be unimportant to them; whenever they are disposed to say something, they will also be disposed to say that they think that thing. Reporting their thoughts will become automatically interchangeable with verbally expressing them. As Rosenthal has shown, since being disposed to report a thought is being disposed to express a higher-order awareness of it, those thoughts will then often be conscious.

This account applies only to creatures with suitable linguistic abilities, but it may be that nonlinguistic creatures are conscious of their mental states only in respect of their qualitative character. It is likely that we can explain why some mental states are conscious independent of any added function that the consciousness of those states might confer.

See also: Cognitive Theories of Consciousness; Folk Theories of Consciousness; Functions of Consciousness; History of Philosophical Theories of Consciousness; Intentionality and Consciousness; Language and Consciousness.

Suggested Readings

Armstrong DM (1968) A Materialist Theory of the Mind. New York: Humanities Press; second revised edition, London: Routledge & Kegan Paul, 1993.

Baars BJ (1988) A Cognitive Theory of Consciousness. Cambridge: Cambridge University Press.

Block N (1995) On a confusion about a function of consciousness. The Behavioral and Brain Sciences 18(2): 227–247.

Breitmeyer BG and Haluk Ö (2006) Visual Masking: Time Slices through Conscious and Unconscious Vision. 2nd edn. New York: Oxford University Press.

Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press.

Clark A (1993) Sensory Qualities. Oxford: Clarendon Press.

Dehaene S and Lionel N (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workshop framework. Cognition 79(1–2): 1–37.

Dennett DC (1991) Consciousness Explained. Boston: Little, Brown and Company.

Dienes Z and Josef P (2004) Assumptions of a Subjective Measure of Consciousness: Three Mappings. In: Gennaro RJ (ed.) Higher-Order Theories of Consciousness, pp. 173–199. Amsterdam and Philadelphia: John Benjamins Publishers.

Dretske F (1995) Naturalizing the Mind. Cambridge, MA: MIT Press/Bradford Books.

Dretske F (2006) Perception without awareness. In: Gendler TS and Hawthorne J (eds.) Perceptual Experience, pp. 147–180. Oxford: Clarendon Press.

Grimes J (1996) On the failure to detect changes in scenes across saccades. In: Akins K (ed.) Perception, pp. 89–110. New York: Oxford University Press.

Haggard P (1999) Perceived timing of self-initiated actions. In: Aschersleben G, Bachmann T, and Müsseler J (eds.) Cognitive Contributions to the Perception of Spatial and Temporal Events, pp. 215–231. Amsterdam: Elsevier.

Hirstein W (2005) Brain Fiction: Self-deception and the Riddle of Confabulation. Cambridge, MA: The MIT Press.

Kriegel U and Kenneth W (2006) Self-Representational Approaches to Consciousness. Cambridge, MA: The MIT Press/A Bradford Book.

Levine J (2001) Purple Haze: The Puzzle of Consciousness. New York: Oxford University Press.

Libet B (2004) Mind Time: The Temporal Factor in Consciousness. Cambridge, MA: Harvard University Press.

Litman L and Arthur SR (2005) Implicit cognition and thought. In: Holyoak KJ and Morrison RG (eds.) The Cambridge Handbook of Thinking and Reasoning, pp. 431–453. New York: Cambridge University Press.

Lycan WG (1996) Consciousness and Experience. Cambridge, MA: MIT Press/Bradford Books.

Raffman D (1995) On the persistence of phenomenology. In: Metzinger T (ed.) Conscious Experience, pp. 293–308. Exeter, UK: Imprint Academic.

Rosenthal DM (2005) Consciousness and Mind. Oxford: Clarendon Press.

Rosenthal DM (2008) Consciousness and its function. Neuropsychologia 46(3): 829–840.

Simons DJ and Ronald AR (2005) Change blindness: Past, present, and future. Trends in Cognitive Sciences 9(1): 16–20.

Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5(42): (November 2).

Weiskrantz L (1997) Consciousness Lost and Found: A Neuropsychological Exploration. Oxford: Oxford University Press.

Biographical Sketch

David M. Rosenthal is professor of philosophy and coordinator of the interdisciplinary concentration in cognitive science at the Graduate Center, City University of New York. He has written extensively in philosophy of mind, especially about consciousness, and has lectured widely on consciousness and related topics. He has developed and defended a higher-order-thought theory of the consciousness of mental states, and a theory of mental qualities and qualitative consciousness based on a homomorphism of mental qualities with the perceptible properties of physical objects. He is currently the president of the Association for the Scientific Study of Consciousness, and has been a McDonnell Visiting Lecturer and a McDonnell-Pew Fellow at the University of Oxford and a Resident Fellow at the Center for Interdisciplinary Research (ZiF) at the University of Bielefeld. He received his PhD from Princeton University and his AB in history from the University of Chicago.

Conscious and Unconscious Control of Spatial Action

B Hommel, Leiden University, Leiden, the Netherlands

Glossary

Action effect – Sensory consequences of actions (reafference), assumed to be integrated with the actions they accompany and to serve as their 'mental' retrieval cues.
Dual-pathway models of sensorimotor processing – Models assuming that sensory codes are translated into motor activity along two processing pathways, one subserving direct online sensorimotor transformation and another allowing for perceptual elaboration and off-line action planning.
Executive ignorance – The phenomenon that voluntary agents have conscious access to their action goals only but no insight into how these are translated into action.
Feedback – Feedback control considers action-produced, reafferent stimulation to fine-tune or stop the action, that is, preliminary results of an ongoing process can influence earlier stages of the process in a continuous loop.
Feedforward – Feedforward control consists of the complete prespecification of an action, which then runs off in a ballistic, context-insensitive fashion.
Motor program – Originally a concept to refer to lists of muscle instructions, later relaxed to allow for more abstract representations of only the invariant properties of actions.
Somatic marker – A representation of the affective consequences of an action (action effects), which can be used as a shortcut to the action in selection processes.

Introduction

Any overt action carried out with the muscles moves the body in space. This is particularly obvious if we actively navigate through our environment or for manual actions, such as grasping a cup of coffee or playing piano. But even actions that serve entirely nonspatial purposes, such as speaking to someone or singing, are controlled at lower motoric levels by targeting particular spatial end configurations of one's jaws, lips, and tongue. Not to forget the actions we carry out to generate perceptual information, such as scanning a visual scene with one's eyes, turning one's head and ears to sound sources, and systematically exploring the texture of surfaces with one's hand. Hence, we move our body in space all day long, and thereby change the configuration of our body parts and their relationship with the environment. At the same time, however, there is very little we know about all these changes and the ways we achieve them. Consider, for instance, you would be asked how you tie your shoes or how you ride a bike (assuming that you master these skills). Presumably, there is very little of interest you could say, simply because you know very little about the details. Most of your description will not differ much from what any other observer could see just as well (that you pick the right shoe lace with the index finger and thumb of your right hand, and so forth), which implies that you have no privileged knowledge about how you move your body in space. How is this possible, that you can move almost any limb of yours in almost any physically possible way, and yet have no idea how? This phenomenon of executive ignorance, as it is sometimes called, has been studied for more than 150 years by now and, fortunately, some progress has been made.

Action Control

Human action is characterized by its voluntary nature, that is, with very few exceptions we do not respond to stimuli in an invariant, reflexive manner but carry out movements to reach a particular goal. One may ask whether this is possibly

an illusion and actions are only interpreted as being goal-directed after the fact. However, the final goal actions aim at often shapes the movement elements needed to reach it, so that actions reflect their goal right from the start. For instance, grasping an object entails a number of systematic movement features, with one being the aperture of the hand adjusting to the size of the object long before the contact is made. Another example is the way people pick up an object they want to move – such as when putting a book back into a shelf – which is commonly chosen to guarantee a comfortable position of the hand at the end of the action (the end state comfort principle). These and other observations have lead to the idea that voluntary actions are mediated and driven by motor programs, which are conceived of as stored representations of actions that can be retrieved and used for controlling an action at will, even in the absence of feedback. Numerous findings have supported this idea of a central engram of action. For instance, patients with a complete loss of kinesthetic feedback are still able to carry out goal-directed actions with the afflicted limb and experimentally deafferented monkeys can still grasp, walk, and jump. Other interesting observations stem from analyses of action errors. Errors sometimes anticipate later action elements and thus are actually correct action components at the wrong time – famous examples are the expressions delivered from William Archibald Spooner, such as 'the queer old dean' (where he actually wanted to refer to 'the dear old queen'). Less entertaining examples are everyday lapses in the order of sequential manual actions, such as when making tea or coffee, or typing errors, such as correctly doubling the wrong 'leeter.' Further evidence is provided by experimental studies of actions varying in complexity, such as the number of action steps or their accuracy demands, that show that the time needed to execute an action increases with complexity. Hence, there is strong evidence that actions are commonly guided by internal representations, including the actions' goal.

Actions and Effects

Combining the phenomenon of 'executive ignorance' (i.e., the observation that people do not have much insight into the 'how' of their voluntary actions) with the assumption that voluntary actions are guided by something like motor programs poses an important question: If we have no conscious access to the contents of motor programs, how are we able to select them? One possible, philosophically very interesting answer will be considered in the next section: consciousness may not have anything to do with the selection of actions but, rather, may merely serve to make sense of or justify these actions after the fact. But there is another possibility. As ideomotor theorists in the nineteenth century such as Hermann Lotze and William James have suggested, conscious experience may well be associated with (and perhaps even causally effective in) the selection of the action goal, which then in one way or another takes care of the further motor programming itself. The developing infant and the novice facing a novel motor task may start by carrying out all sorts of involuntary movements, a process sometimes called 'motor babbling.' The motor patterns generating these movements may be set up entirely by chance or follow genetically specified reflexes, but they should systematically produce particular sensory feedback. An automatic integration mechanism may associate the motor patterns and the perceived action effects in a bidirectional fashion. If so, the motor pattern could later be intentionally retrieved by endogenously activating an aimed-at action effect – thinking of the goal triggers the movements necessary to reach it without any conscious insight into their inner workings.

Numerous studies have in fact revealed that people pick up novel effects of their movements spontaneously and continuously, thereby steadily increasing their action repertoire and the number of goals they can realize. Spontaneous acquisition has been observed in adults, children, and infants, and experiments have demonstrated that agents consider the novel action effects in the selection of voluntary actions. Brain imaging studies have shown that episodic action-effect associations are stored in the hippocampus and that reactivating them primes motor representations in the supplementary motor area, which again is known to play a crucial role in voluntary action planning.

Many studies have focused on the perhaps most obvious sensory effects, such as visual and auditory

consequences. However, recent theorizing has emphasized that the sensory codes of affective consequences may be of particular importance for action planning. According to Antonio Damasio and colleagues, we integrate motor patterns not only with codes of the reafferent feedback they produce but also with codes of the ways they make us feel. Carrying out an action can have more or less positive or negative consequences, which lead to corresponding affective states. Increasing evidence suggests that representations of these states (so-called somatic markers) are associated with the actions they accompany, so that reactivating an action tends to reevoke the associated somatic marker. If so, somatic markers can be used to select among possible actions and guide one to the action alternative that 'feels best' – a selection mode that one may call intuition or deciding 'by gut.' Interestingly, realistic studies of decision-making in complex everyday situations have revealed that good deciders are not at all following the rules of logic (which do allow good decisions on simple problems), but decide in ways they are often unable to explain themselves. Moreover, simulation studies have shown that logic-based decision-making is likely to be way too slow to work under realistic circumstances, in which decisions often have to be fast. In these cases, somatic markers may provide a kind of shortcut to appropriate decisions.

Will and Consciousness

According to ideomotor reasoning, consciously representing a goal (i.e., the intended action effect) is sufficient to retrieve and execute the necessary motor patterns. Conscious representations play thus a restricted but still crucial role in action control. Some researchers have questioned even this role however. If conscious representations would really be causal in bringing about intentional action, so they argue, the causal experience should temporally precede the action it causes. That this may not be the case is suggested by a classic experiment of Benjamin Libet. He asked subjects to carry out voluntary movements and measured the time at which they showed a readiness potential (an EEG component preceding voluntary actions) and when they began consciously intending the movement. Surprisingly, the readiness potential was measured much earlier than the conscious intention. This suggests that the 'brain' had made the decision to act long before this decision was consciously experienced, which again undermines the idea that it is the experience that drives the action. Some authors have argued that the conscious experience of intentions may be the product rather than the cause of the neural processes leading to action, which means that consciousness does not play any causal role in action control. Other researchers have argued that this holds true only for the initiation of actions while a conscious 'veto' may still be conceivable. In other words, conscious experience may be functional not in producing but in preventing actions.

On the one hand, the available evidence clearly shows that the relationship between conscious intentions and the actions they refer to is more complicated than the common sense model of action (perception → conscious decision-making → action, see the section 'Sensorimotor processing') suggests. Clearly, voluntary actions can be prepared and carried out without tight conscious monitoring, which is also evident from many everyday observations: We walk without thinking of every single step and drive home with very little cognitive involvement. And yet, there is no evidence that voluntary actions are possible without a conscious representation of the goal. We cannot exclude that, in Libet's study, consciously constructing the overarching task goal and preplanning the possible actions once in the beginning of the experiment was sufficient to drive remaining activities more or less automatically, just as consciously intending to drive home is enough to have the remaining action run in 'autopilot' mode.

Motor Programs and Action Plans

The original concept of a motor program that controls the execution of a movement was derived from the domain of computer programming. The idea was that sequences of muscle movements could be stored and rerun in a purely feedforward fashion whenever needed. Historically, this approach was a counterreaction to the strong emphasis on stimulus-driven behavior associated with American behaviorism, which dominated the psychological scene in the first half of the twentieth century. Behaviorist

approaches tried to reconstruct actions as responses, that is, as logical and empirical consequences of stimuli impinging on the sense organs. These approaches were surprisingly successful and did a good job in disenchanting the concept of voluntary action quite a bit. But in the 1950s and 1960s it became increasingly clear that there are too many indications that actions can be generated in a purely endogenous fashion and in the absence of any sensory and evaluative feedback to further rely on behaviorist concepts of action control. In sight of the upcoming computer revolution and the increasing temptation to apply computer metaphors to human cognition, it seemed only logical to consider action being controlled by programs that have a structure similar to computer software. If such software could produce 'overt' behavior on a computer monitor, why should mental software not generate muscle activity?

However, even if people would be able to maintain muscle-specific information associated with a particular movement, the idea is unlikely to solve a number of problems. One main problem relates to 'storage.' Consider all the possible pointing actions you could perform, all the possible locations in space that your index finger may occupy, and all the possible configurations of the limbs involved this would imply. If a motor program would really be a literal record of muscle movements, each single combination of all the factors would require a program on its own. Worse, which muscles need to move in which way to reach a particular movement goal is dependent on context and starting conditions, such as the current positions of the body and the limbs involved. A separate program would need to be constructed for each possible context. Considering that pointing is just one simple movement out of the many simple and complex movements you can perform, this would imply enormous amounts of memory to store all those programs, not speaking of the search time needed to relocate a required program in that memory. Even though the human brain entertains smart routines to handle large amounts of data, the implied memory demands seem too excessive. A second problem relates to 'novelty.' Having learned to point to one location is enough to generate any other pointing movement – even if different muscles are involved, and other movements generalize just as well. It is difficult to see, however, how generalization from muscle-specific programs would work. Hence, very detailed representations of movements, as implied by the original motor program concept, would be counterproductive.

The major theoretical move to fix these and other problems was to assume that the motor program proper consists of invariant information only (a kind of motor schema or plan), whereas parameter slots were defined for configuring a program for a particular task and purpose. The idea was that action control consists of two phases, one in which the proper program is selected and another in which the (commonly spatial and/or timing-related) parameters of this program are specified to tailor it to the situation at hand. Pointing would be a good example: Selecting a generalized pointing program would specify which limbs are involved and how they move in relation to each other, whereas a location parameter would specify the spatial target of the movement and a timing parameter would specify how fast the action is carried out. This approach successfully overcomes both the storage problem, as very few programs are actually stored, and the novelty problem, as using different parameters would be a natural way to generalize.

Sensorimotor Processing

The main motivation for the original motor program approach were findings that voluntary actions can be carried out even in the absence of any sensory feedback. Accordingly, the intention was to conceive of a control structure that can run in a completely feedforward fashion without any need for sensory information. However, even though it is possible to carry out at least some actions without any feedback, it is clear that the accuracy of actions is often enhanced if feedback is available. Moreover, sensory contributions to action programming do not need to be restricted to feedback, that is, to information that is produced by the action, but the sensory information available at the onset or during the execution of an action may be just as important. Indeed, once programs are thought to accept parameters, it makes sense to consider that parameter values are delivered by the environment. This

raises the question of how sensory information affects and shapes action control.

Single-Pathway Models

One way or another, all organisms are capable of responding to stimuli from their environment, be it flowers stretching toward the sun or bacteria fleeing areas of high phenol concentration. In lower organisms the links between sensors and effectors are rather short and direct, so that the action of a given effector is easy to understand from tracking the transmission of signals picked up by the sensors. In the beginning of the systematic investigation of human sensorimotor processing, the same logic was applied: Motor activity was thus conceived of a more or less direct extension of sensory processing, with the task being to trace stimulus signals through the processing system until it makes contact with the muscles. A famous example of this line of reasoning is the empirical and theoretical work of René Descartes, who suggested that the pineal gland transforms the electric signals it receives from the eye into hydraulic energy driving the muscles. Later researchers like Franciscus Cornelis Donders further developed this approach and suggested that human cognition can be understood as a (rather extended) sequence of processing stages beginning with the sensory registration of a stimulus and resulting in the movements of the muscles. The interesting implication of Donders' approach was that it allows for the separate measurement of the time demands of each single processing stage by systematically manipulating the processes necessary for performing a given task.

Not all researchers working with this single-pathway conception of human information processing were interested in the role of consciousness, but those who were located conscious experience right in the middle between sensory coding and muscle movement. That is, it was not assumed that the whole chain of processes would be accessible for conscious experience, but rather, conscious experience and decision-making was thought to be associated with the higher ends of perception (such as the conscious experience of meaningful objects and events) and the decision which action this perceptual experience calls for (hence, the conscious experience of will). In other words, the role of consciousness, or of the processes associated with conscious experience, was to separate perception from action, so that actions would be no longer directly driven by stimuli but could be planned ahead and triggered in the absence of any external stimulation. The theoretical idea shared by single-pathway models is sketched in Figure 1: In lower organisms sensory coding more or less directly leads to motor output, whereas higher organisms (and humans in particular) are assumed to have acquired means to decouple stimulus processing from motor programming – with consciousness being either functional in, or at least associated with, this decoupling.

Multiple-Pathway Models

Even though it seems clear that humans are no longer purely reflexive organisms, which certainly implies some additional processing capabilities, one can ask whether this necessarily implies the loss of more direct linkages between sensory and motor processing. Indeed, increasing evidence suggests that the two pathways sketched in Figure 1 do not represent 'alternative' processing routes but, rather, a tandem of concurrent processing streams that distribute labor in a particularly efficient way. Numerous observations have suggested that (1) there must be more than one pathway from sensors to effectors and that (2) not all pathways are monitored by consciousness and accessible for conscious experience.

Early evidence for the presence of multiple pathways was provided by prism studies, in which human subjects wore goggles that systematically

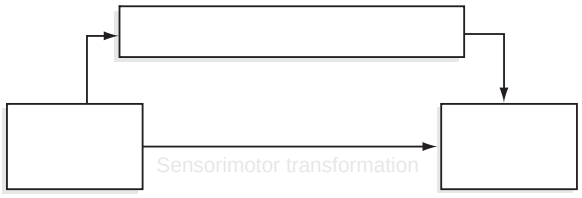


Figure 1 Dual-pathway models of sensorimotor processing distinguish between a low-level pathway translating sensory codes into motor activity and a high-level pathway subserving higher-order cognitive operations and action planning.

distorted their visual input, for example, by turning it upside down. From these studies, two observations are particularly impressive. First, the conscious experience of the visual world (or, perhaps better, the conscious visual experience of the world) turned to normal after some time, which means that the subjects have learned to 'correct' the visual input for the distortion. Second, however, the subjects regained their ability to carry out spatial actions (such as walking or riding a bike) much earlier, long before conscious experience was successfully readapted. This implies that actions have access to sensory information that is not available for, and certainly not controlled by phenomenal experience. Numerous dissociations between conscious perception and judgment on the one hand and (commonly) manual motor action on the other have been demonstrated since then – for example, people can correctly point at stimuli they consciously mislocate in space and correctly grasp objects they consciously misperceive in size. Likewise, even if people are unable to detect a displacement of an object they are in the process of grasping (by moving the object during an eye blink) their hand moves straight to the new location without any noticeable delay or hesitation.

Patient studies provided converging evidence. For instance, patients with lesions in their visual cortex are unable to consciously perceive stimuli falling into the visual field corresponding to the lesioned area; and yet they can correctly point to these stimuli (the so-called 'blindsight' phenomenon). Patients with bilateral damages in the posterior parietal cortex (the Balint–Holmes syndrome) are perfectly able to judge the orientation of a slit in a wooden board, but are unable to orient their hand so that it can pass through the slit. Reversely, patients suffering from visual form agnosia (due to lesions affecting visual cortex) are able to orient their hand accordingly but at the same time cannot judge the orientation of the slit. Hence, numerous findings converge on the conclusion that multiple pathways exist from sensory coding to motor processing and that not all pathways are consciously accessible.

Even though researchers agree that multiple pathways exist, they differ with respect to the question of how such pathways should be characterized. Early approaches assumed that one pathway is mainly responsible for the processing of identity-related information (the 'what' pathway), such as shape and color, while another pathway focuses on spatial information (the 'where' pathway). A later approach has suggested that one pathway mainly serves to inform conscious perception (the 'ventral' pathway, named after its assumed anatomical location) while another feeds motor processes with action-related information (the 'dorsal' pathway). Other approaches have focused more on the distinction between off-line processing (be it for perception or action planning) and online processing (be it for sensory coding or motor programming). Even though the debate is not entirely settled, most differences are rather semantic and restricted to details. A general agreement is that there must be two or more different processing pathways that are likely to represent a combination of what previously was thought to be alternatives: A direct link between sensory and motor processes that we share with lower organisms and a much more cognitively penetrated and in part consciously accessible link, including higher-level perception and the planning of goal-directed actions (see Figure 1).

Establishing multiple concurrent pathways for what looks to be the same purpose (bridging the gap between sensors and effectors) seems redundant and creating needless confusion and communication problems. However, close consideration reveals that the purposes the multiple streams apparently serve are rather different and to some degree mutually incompatible. First, there are different demands in terms of processing 'speed.' If we consider direct sensorimotor transformation processes (the lower branch in Figure 1) to continuously feed motor actions with online information about the current environmental state of affairs, these processes need to be fast and specific enough to guide a hand toward a visible goal. In contrast, speed is not so much of an issue when recognizing an object and planning an action via the higher-level route.

Second, the 'type of processes' operating in the two pathways is likely to differ. Sensorimotor transformation consists in the application of acquired transformation rules (which need readjustment when wearing goggles with reversal prisms, for instance), whereas recognition and action planning

require the extensive use of memory retrieval, prediction and probability estimates, and reasoning – all processes unnecessary for sensorimotor functioning.

Third, a related point, the 'amount of information' to be considered differs. Sensorimotor transformation works best if the incoming sensory information is as pure as possible, whereas recognition and action planning benefit from extensive use of anticipation, elaboration, and retrieved memory codes – the informationally richer the resulting representation, the better.

Fourth, sensorimotor and the more cognitive pathways need rather different 'spatial reference frames' to code information efficiently. Motor control needs to consider the body of the agent and the configuration of the effectors involved. Action goals need to be translated into body- and effector-related reference frames, so that the required movements can be properly programmed and executed. Many of such so-called pragmatic body maps have been identified in the primate brain and it is interesting to consider that the way most of these maps represent action space is not reflected at all in our conscious experience. In contrast, conscious experience commonly conceives of space as egocentric (considering the body as a whole) and allocentric (i.e., ndependent of the body). Even though the mechanism underlying this ability is not yet well understood, we are somehow able to construct the latter from the former by integrating many different egocentric 'snapshots' of experiences into a coherent allocentric representation, such as of the city we live in by navigating through it by foot, bike, subway, and car. The emerging allocentric representation is particularly important for planning purposes, but it is at the same time way too coarse to control the movements of our feet. Allocentric representations are also important for the recognition of objects irrespective of their orientation, another ability that requires the integration of egocentric snapshots into an ego-independent representational format.

Interacting Levels of Action Control

So far, we have seen three converging lines of theoretical developments. First, modern approaches to action control started with a strong emphasis on purely feedforward programming but became increasingly lenient with respect to the consideration of the available sensory information to adapt motor programs to current needs. Recent models assume that stored action-control structures are more like Swiss cheese: The general structure, a mere schema or action plan, is maintained after use but a number of slots are intentionally left open to be specified online whenever needed. Second, ideomotor approaches explain the executive ignorance of conscious experience with respect to the details of action selection and execution by referring to the automatic integration of motor patterns with representations of their sensory consequences (i.e., action effects). Accordingly, action-effect codes become retrieval cues for the associated motor patterns, so that the conscious representation of intended action effects is sufficient to launch the programs necessary to reach them. Third, modern approaches to sensorimotor processing assume that there are multiple streams from sensory processing to motor activity. Some of these streams link sensory codes more or less directly to motor execution, thus providing a fast and pure picture of the sensory state of affairs, whereas other streams are busy with the cognitive elaboration of the input and the setting up of action strategies.

Figure 2 shows how these assumptions fit together. The observation that conscious representations are restricted to the perceivable action consequences is consistent with the claim that conscious access is restricted to higher cognitive processing pathways, pathways that are specialized for the off-line elaboration of perceived events and anticipatory action plans. In some sense, the only difference between perceiving a current event and planning an action is the temporal reference: What we call perception is awareness of the sensory aspects of an event that happens to be right now, based on the activation of sensory codes describing this event, whereas what we call action planning is awareness of the sensory aspects of an event, again based on the activation of sensory codes describing it, that is still to occur – a sort of imagery. One might object that action planning has motor consequences while perception has not. However, apart from the fact that requiring information for perception almost

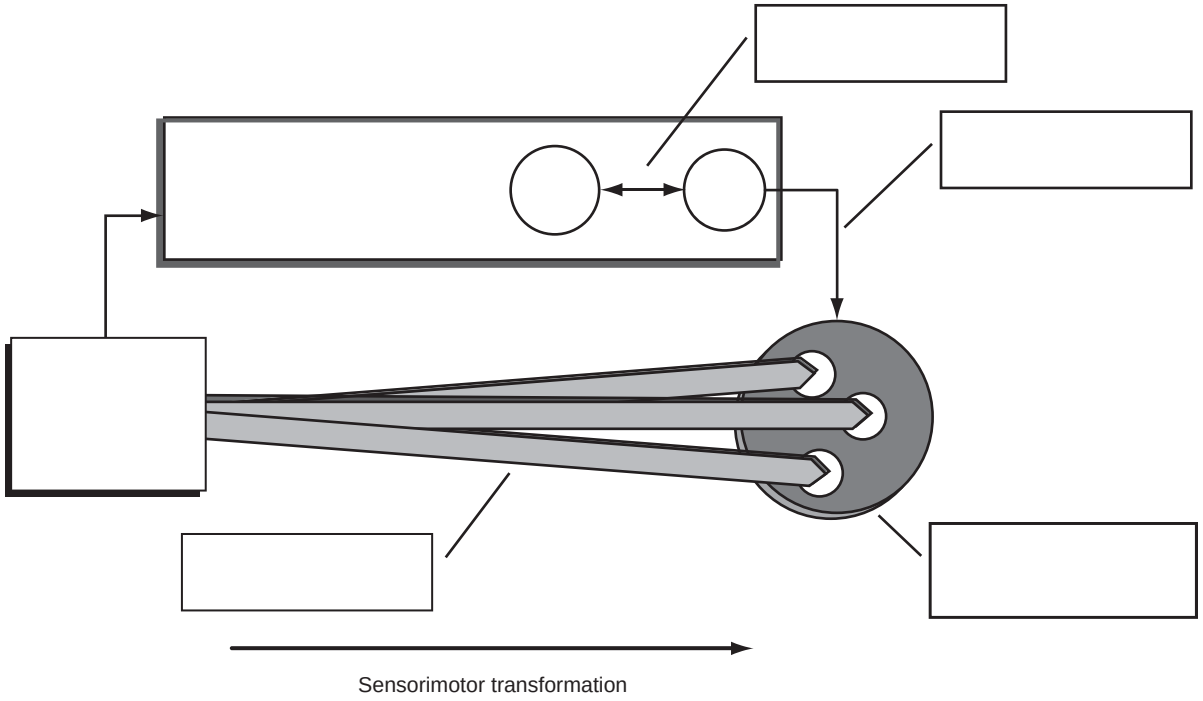


Figure 2 Action control in a dual-pathway system. Specifying a goal leads to the selection of codes representing the sensory consequences of actions suited to bring about the intended effects (action-effect codes or AE). Action-effect codes activate associated action plans (the off-line component of action control), which contain open parameter slots to be filled online by available sensory information (the online component of action control).

always involves motor activity, there is increasing evidence that perceiving an object primes the actions it affords and perceiving actions of other people activates compatible action tendencies. Moreover, if people judge and compare visual objects, their performance is affected by relations between the actions these objects imply, suggesting that action-related information is automatically activated in what looks like straightforward perceptual tasks. Hence, perception and action planning seem to bias motor processes in comparable ways, so that it makes sense that our consciousness treats them equally. Selecting a plan does not fully determine its motoric realization. Rather, it only makes sure that the intended sensory consequences are produced in one way or another (the equifinality principle). In which way they are is eventually determined by the present sensory situation, which fills in the open parameter slots. The whole filling-in operation and, presumably, the filled-in parameter values are not only commonly unconscious but presumably not accessible by conscious operations in principle.

Even though some authors have argued that the two main processing pathways are independent, this stands in contrast with a number of empirical findings, but it is also unlikely for logical reasons. If it is the case that low-level sensorimotor transformation procedures are entirely unaffected by higher-level perception, memory retrieval, decision-making, and action planning, how do they know which parameters to select and feed into which motor program? Consider a grasping movement. You are often facing numerous objects, so that without selecting one target object transformation procedures would not know which sensory information to use as input. Minimally, the perception-and-planning pathway would need to communicate to the sensorimotor transformation pathway which object or location in space is to be attended. Next, the sensorimotor pathway would need to know for which action the transformation is required. Grasping mainly calls for shape and location information while pointing is likely to rely more on contrast and/or color. This means that attention also needs to be directed to particular

stimulus dimensions, a job that relies on information about the planned action. Furthermore, analyses of sequential actions have revealed context effects and optimization strategies, which seem to imply rather massive interactions between pathways. For instance, the way one grasps a given object depends on what one is going to do with this object next – to move it somewhere else, to bring it to the mouth, or to throw it away. This means that rather concrete action parameters are biased by the next upcoming action elements, which means that parameter specification at lower levels must have some access to, or be affected by higher-level planning processes. Thus, even though the sensorimotor transformation pathway and the perception-and-planning pathway differ in many respects, and are better off doing so, good communication between them is essential for coherent goal-directed action.

Any distortion of this communication can lead to abnormal phenomena. As mentioned earlier, patients with optic ataxia (the Balint–Holmes syndrome) have no problem to consciously perceive objects, but at the same time they have difficulty to manipulate them manually, whereas patients with visual agnosia show the opposite pattern. In line with dual-pathway models, this seems to reflect selective impairments of sensorimotor transformation and perception-and-planning pathways, respectively. Damage of the frontal lobes, which are particularly important for action planning, has been reported to lead to so-called 'utilization behavior': Patients tend to carry out actions on objects that are 'sensomotorically correct' without making sense or following any goal. This seems to suggest that the sensorimotor transformation pathway is still functional but no longer under control and in the service of action planning. A possible explanation for delusion of control experiences reported by schizophrenics might be that action planning is still working but no longer represented in consciousness. Accordingly, the outcomes of actions are not consciously anticipated and therefore surprising. Daniel Wegner and others have suspected successful anticipation of action outcomes to be a major component of the attribution of agency, with anticipated outcomes indicating self-performed actions. If so, a loss of conscious anticipation is likely to prevent the attribution of self-agency.

Role and Function of Conscious Control

Now that we have roughly circumscribed the processes that are related to consciousness, it seems appropriate to ask what role consciousness may play in and for them. Why is it that we have conscious access to higher-order perceptual processes and action planning, but are widely ignorant with respect to the specifics of sensory and motor processing? Depending on one's philosophical stance regarding the concept of consciousness, different types of answers to this question are appropriate.

From an evolutionary standpoint, it is certainly difficult to say why conscious experience exists anyway and, if it exists, why it does not reflect all existing processes. However, it is also clear that the phylogenetic development of processing pathways on top of and in addition to the basic, relatively direct linkage between sensory organs and motor systems is associated with the emergence of conscious experience. In fact, the more a species shows evidence of similar multiple-pathway structures, the more signs they show for aspects of consciousness. This does not account for the *qualia* of conscious experience, but it points to an interesting correlation between the existence of consciousness and a particular alternative way to transform sensory information into motor activity, which allows for the decoupling of perception and action, for prospective planning, and for stimulus-independent, endogenously generated action.

From a functionalist point of view, it seems interesting to consider why this correlation exists. That is, what are the characteristics of processes that accompany conscious experience (irrespective of the issue whether this correlation expresses a particular causality or not)? Most interestingly, there is no evidence of any conscious access to 'processes'; rather, it is the 'states' these processes produce that are consciously represented: We perceive the sensory consequences of an action but not the operations necessary to make them accessible to our experience. Moreover, increasing evidence suggests that conscious experience is restricted to rather global, highly integrative states. This is particularly obvious for the generation of action goals. Even though we are commonly aware of the action goal we currently pursue (at least to

some degree), we commonly have no insight into why it is this goal we chose. This is not to say that we cannot think of justifications for the goals we have after the fact or point to the role of a subgoal in a broader action plan (like boiling water in the process of making tea), but why exactly we ended up having this but not other goals we commonly cannot say for sure. (Again, this is not any different from perception; e.g., while it is evident that we can switch between different views of multistable stimuli, we are unable to say why and how we switch.) Nevertheless, chosen goals often reflect multiple constraints, often more than we can enumerate, which suggests that the generation of a goal considers huge amounts of information collected from the senses, from memory, and from *ad hoc* reasoning. Again, our conscious experience does not have access to all that information and the processes integrating it, but is merely presented with the final outcome.

A given goal can lead to different types of action. Western legal systems are based on the idea of conscious deliberation and, indeed, at least the selection of novel and complicated actions to reach a particular goal is commonly accompanied by conscious experience. Again, the experience is not referring to processes but to states representing the effects the considered actions are expected to have. Whether or not the experience as such has any causal relevance – which given that experience relates to products rather than causes seems questionable, however – it seems clear that intermediate states of deliberation are consciously accessible and often associated with conscious experience. However, it is only the anticipated (perceptual and affective) consequences of the considered actions the experience refers to. Eventually, one alternative is being selected (often for justifiable but actually unknown reasons) and thereby made available for conscious representation. This is where conscious insight into action control ends; the next aspect we experience is only whether the action produced unexpected results (while expected results often go unnoticed).

The strong link between consciousness and action planning and the observation that conscious experience often refers to the consequences of actions have motivated a number of so-called 'simulation theories' of cognition. In a certain sense, these theories reverse the commonsense approach to action control. The commonsense approach suggests that it is consciousness that makes action planning possible by allowing one to mentally simulate and 'think through' alternative actions before deciding which action one prefers. Simulation theories acknowledge the value of such mind plays but assume that it is the ability to plan and play through an action that actually came first and conscious experience only emerged as a consequence. Making creative use of action-effect associations that originally evolved for much more basic functional reasons may have allowed extending one's thoughts beyond the current, perceptually available situation by reasoning in terms of what-if.

To summarize, conscious experience is most closely related to the goal of an action and its intended consequences, and the experience seems to focus on a rather general, highly integrated representation of these aspects. This observation fits with approaches that see consciousness as a kind of global work space, such as the theory of Bernard Baars. The idea is that much cognitive processing is going on in local modules that are highly specialized and fast, but their inner workings and intermediate results are not consciously accessible. Their products however, the information they generate, is sent to a global work space, where information can thus be related to each other, integrated, and acted upon. Even though this approach fails to explain the *qualia* aspect of conscious experience, it is consistent with the picture that emerges from the available evidence, the picture of a surprisingly ignorant and only globally informed voluntary agent.

See also: Intentionality and Consciousness; Perception, Action, and Consciousness; Perception: Subliminal and Implicit.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. New York: Cambridge University Press.

Blakemore SJ, Wolpert DM, and Frith CD (2002) Abnormalities in the awareness of action. Trends in Cognitive Sciences 6: 237–242.

Clark A (2007) What reaching teaches: Consciousness, control, and the inner zombie. British Journal of the Philosophy of Science 58: 563–594.

Damasio A (1999) The Feeling of What Happens: Body, Emotion and the Making of Consciousness. London: Vintage Books.
Hesslow G (2002) Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences 6: 242–247.
Hommel B (2007) Consciousness and control: Not identical twins. Journal of Consciousness Studies 14: 155–176.
Hommel B, Müsseler J, Aschersleben G, and Prinz W (2001) The theory of event coding (TEC): A framework for perception and action planning. Behavioral and Brain Sciences 24: 849–878.
James W (1890) The Principles of Psychology, vol. 2. Cambridge, MA: Harvard University Press.
Jeannerod M (2006) Motor Cognition: What Actions Tell to the Self. Oxford: Oxford University Press.
Miller GA, Galanter E, and Pribram KH (1960) Plans and the Structure of Behavior. New York: Holt, Rinehart, & Winston.
Milner AD and Goodale MA (1995) The Visual Brain in Action. Oxford: University Press.
Prinz W (1992) Why don't we perceive our brain states. European Journal of Cognitive Psychology 4: 1–20.
Wegner D (2002) The Illusion of Conscious Will. Cambridge: MIT Press.

Biographical Sketch

Prof. Dr. Bernhard Hommel holds the chair of "Gerneral Psychology" at Leiden University since 1999, after having worked as senior researcher at the Max-Planck Institute for Psychological Research (PhD at the University of Bielefeld in 1990; Habilitation at the Ludwig-Maximilians University of Munich). His research focuses on congnitive, computational, developmental, neural, and neurochemical mechanisms of human attention and action control, and the role of consciousness therein. He served as Associate Editor of *Acta Psychologica*, the *Journal of Experimental Psychology: Human Perception and Performance*, and the *Quarterly Journal of Psychology*, and currently is editor-in-chief of *Psychological Research*. He has (co-) edited two books on action control and the relationship between perception and action, and co-edited several special issues on attention and action control. He has (co-) authored more than 100 articles in international journals and numerous chapters in readers and psychological textbooks.

Consciousness and Memory in Amnesia

M Moscovitch, University of Toronto, Toronto, ON, Canada

Glossary

Anoetic, or nonknowing, consciousness – A type of consciousness associated with implicit memory; though conscious, the person lacks awareness of the memory that influences performance.

Anterograde amnesia – Loss of memory pertaining to stimuli, events, or experiences that occurred after the trauma or dysfunction that marked the onset of amnesia.

Autonoetic consciousness – The consciousness that underlies episodic memory, the recollection of a particular experience, an autobiographical event.

Conceptual implicit memory – Memory without awareness for the meaning of the stimulus, without regard to its perceptual qualities (e.g., for visual objects, what it signifies but not how it looks).

Declarative or explicit memory – Memory with conscious awareness.

Episodic memory – A type of declarative or explicit memory (memory with conscious awareness) which refers to memory for past events associated with a particular autobiographical experience, such as a memorable day on vacation, or an embarrassing event.

Familiarity – A process that enables 'knowing,' a sense of recognizing or experiencing a stimulus or event as old but with little or no information about the context in which it was encountered.

Hippocampus – The most prominent structure in the medial temporal lobe. It is implicated in the encoding, retention, and retrieval of episodic memories, particularly those that are recollected.

Medial temporal lobes – Structures on the inside aspect of the temporal that are concerned with memory.

Noetic awareness – The consciousness that underlies semantic memory.

Nondeclarative or implicit memory – Memory without conscious awareness.

Perceptual implicit memory – Memory without awareness for the perceptual, and possibly sensory, attributes of the stimulus, such as modality (e.g., auditory vs. visual) and within each modality, about perceptual and sensory qualities (e.g., for visual objects, the memory is concerned with aspects such as shape, color, viewpoint, size, and location).

Perirhinal cortex – A structure in the medial temporal lobe believed to support familiarity processes in recognition memory.

Procedural (implicit) memory – Memory without awareness for rules, motor sequences, conditioned responses, and subjective probabilities of stimulus–response or stimulus–stimulus associations or occurrence.

Recollection – A process that enables remembering, the reexperiencing or reliving a past event in the mind. Characterized by recovering and recreating the context in which a stimulus or event occurred, it is a hallmark of true episodic memory.

Relational binding – The encoding of associations among diverse elements into a memory trace. It is generally acknowledged to be a function of the hippocampus.

Retrograde amnesia – Loss of memory for information and experiences acquired prior to the trauma or dysfunction that marked the onset of amnesia.

Semantic memory – A type of declarative or explicit memory (memory with conscious awareness), which refers to general knowledge such as vocabulary and facts about the world, and even facts about oneself.

Introduction

Memory is a lasting, internal representation of a past event or experience (or some aspect of it) 'that is reflected in thought or behavior.' This definition, meant to encompass all memories in all organisms, does not distinguish among the different ways memory can be inferred to exist. According to this definition, one only needs to demonstrate that an experience has modified thought or behavior. Thus, for example, if repetition of a simple action or an action sequence, or repeated exposure to a stimulus, leads to better execution of the action or to better perception of the stimulus, we can infer that memory for the action or stimulus exists. Whether the person, or any other organism for that matter, is consciously aware of a memory for the action or for the percept is not taken into consideration by this definition. Because plasticity is a property of many structures of the nervous system, especially the brain, those structures can be modified by experience, so that memory has the potential of being represented almost anywhere in the nervous system, as much in the neural structures concerned with action and perception, as in those concerned with conscious awareness of a memory. What then distinguishes memory with conscious awareness from memories without it at both a behavioral and neural level? Are there different types of each kind of memory in the sense that each type represents different information, is governed by different principles, and is mediated by different neural structures? In this article, I will try to answer these questions.

Definitions

The terms 'declarative or explicit memory' refer to memory with conscious awareness, and 'nondeclarative or implicit memory,' to memory without conscious awareness. Explicit memory is characterized by an ability to declare or comment, verbally or by any arbitrary action, that one has a memory of a particular stimulus, concept, action, or event. Explicit memory can be tested simply by asking the individual directly whether he or she has a memory. By contrast, because the person or organism is unaware of implicit memories, their existence can only be inferred indirectly, by the individual or by an observer, if a change in behavior or performance is brought about by an experience. For example, suppose you asked a person with a profound memory loss whether she can ride a bicycle. She says that she imagines she can but has no memory of ever having done so. Her explicit memory of having ridden a bicycle is impaired. When placed on a bicycle, however, she performs expertly showing that her implicit memory of how to ride a bicycle is good. Put simply, the distinction between explicit and implicit memory is one between 'knowing that' and 'knowing how' and, as Cohen and Squire noted, it is a distinction that is honored by the brain.

Explicit and Implicit Memory in Amnesia

The difference between the two types of memory can be demonstrated in typical people in everyday life, and in the laboratory, but it is seen most dramatically in people with amnesia. Amnesia is a clinical disorder characterized by a severe impairment or total loss of the ability to acquire long-lasting explicit memories for personal experiences, public events, or general information, despite relatively preserved cognitive function in other domains, such as perception, action, and intelligence. Amnesia may also encompass some declarative memories acquired before the onset of the disorder, sometimes decades earlier. The cause of amnesia can be either primarily 'organic,' resulting from neurological conditions such as stroke, tumor, infection, anoxia, and degenerative diseases that affect brain structures implicated in memory; or it can be primarily 'functional or psychogenic,' resulting from some traumatic psychological experience. We will focus only on organic amnesia.

Neuroanatomy of Memory

Knowledge of the neuroanatomy of memory helps contribute to our understanding of the relation between memory and consciousness (see Figures 1–3). Though organic amnesia can have many etiologies,

it ultimately results from bilateral damage to the medial temporal lobes and related structures in the diencephalon. The hippocampal formation, in the medial temporal lobes, is the most prominent of the memory structures. It consists of the hippocampus proper with its various subfields and regions, plus the dentate gyrus and the subiculum. Communication between the hippocampus and neocortex occurs through a series of bidirectional relays. The hippocampus is connected directly to the entorhinal cortex which in turn is connected to the parahippocampal gyrus and perirhinal cortex which project bidirectionally, primarily to the temporal and parietal lobes of neocortex, respectively. There are also projections from the hippocampus and perirhinal cortex via the fornix and anterior cingulate to the mammillary bodies and the dorsomedial nucleus of the thalamus, respectively, both of which are in the diencephalon. The loop of medial temporal and diencephalic structures constitute the limbic system, but one should note that there are two related systems – the hippocampal system and its projections and related structures, and the perirhinal system and its projections and structures. Both also are connected directly and indirectly to the prefrontal cortex. The hippocampus is thus ideally situated to collate information both about the cognitive (neocortex) and emotional (limbic) state of the organism, and to bind that information into a memory trace that codes for all aspects of a consciously experienced event. By comparison, the perirhinal cortex is linked closely to perceptual and, possibly, conceptual systems, and may be an extension of them into the memory domain. As we shall see, recent studies implicate the two systems in different kinds of memory with awareness.

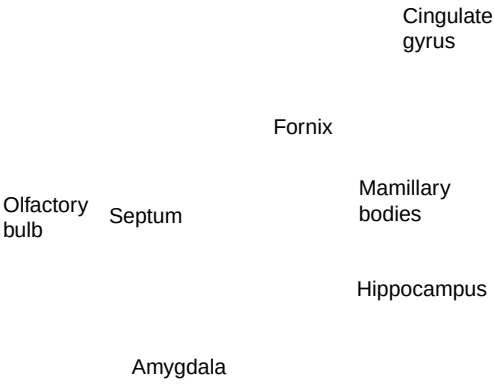


Figure 1 The hippocampus and related structures. Reproduced from Hamilton CW (1976) Basic Limbic Anatomy of the Rat. New York and London: Plenum.

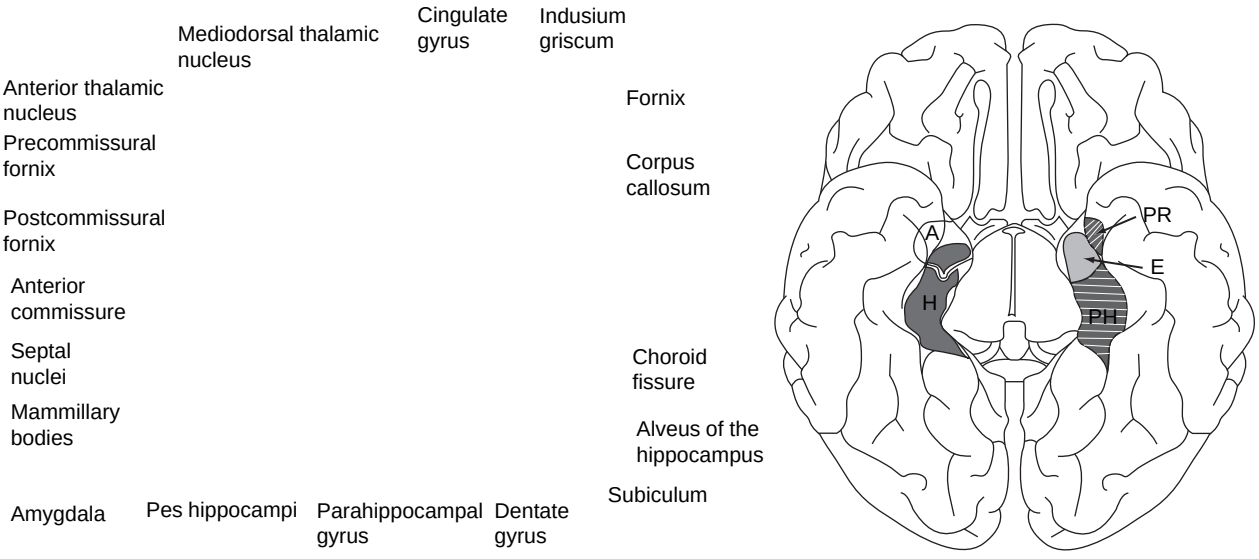


Figure 2 The medial temporal and related structures in diencephalon, and their projections (connecting pathways).

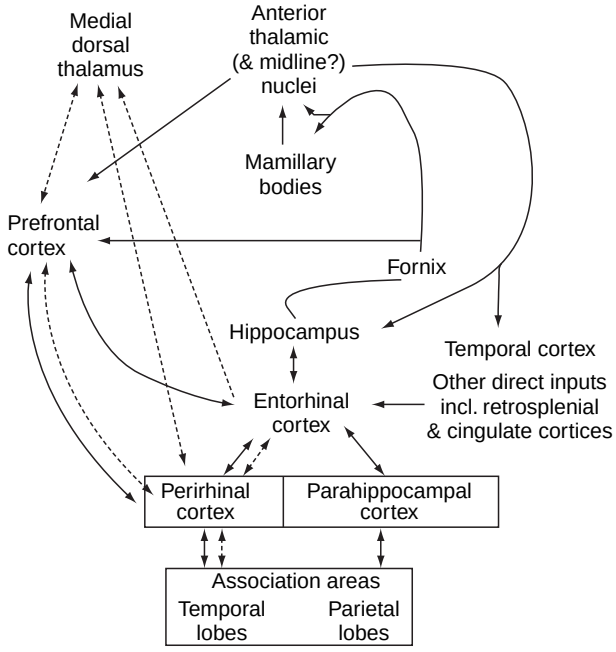


Figure 3 The hippocampal-diencepahalic systems, structures, and connecting pathways. There are two interrelated systems: The hippocampal-fornix-anterior thalamic system indicated by solid lines and the perirhinal-media/dorsal thalamic system indicated by dashed lines. Reproduced from Aggleton JP and Brown MW (1999) Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. Behavioral and Brain Sciences 22: 425–489.

The neural substrate for implicit memory is distributed more widely through the neocortex, basal ganglia, and cerebellum, depending on the type of implicit memory that is being considered.

Memory without Conscious Awareness in Amnesia

Almost since the time organic amnesia was first described in the neurological literature at the end of the nineteenth century, investigators have noted that implicit memories can be spared in the face of a severe and profound loss of explicit memory. Little research, however, was conducted on implicit memory until the late 1960s. It was noted then that people with severe amnesia caused by bilateral medial temporal lobe lesions, could learn and retain motor skills for months or years, though many had no memory of the learning episode in which they acquired the skill, minutes after the episode was over. The same was true of perceptual skills involved in reading mirror-reversed words, or in identifying degraded or fragmented pictures and words. Even words written in ordinary font or pictures to which they had been exposed could be identified more accurately and more quickly by amnesic people than new items, suggesting that they stored information peculiar to that item, even though at a conscious level they could not recall or recognize that they had studied it. Similar phenomena were demonstrated in normal people for material they could not consciously recall or recognize after a sufficiently long delay, suggesting that the dissociation between implicit and explicit memory was not peculiar to amnesia but was indicative of something fundamental about the organization of memory. Tulving associated implicit memory with 'anoetic, or nonknowing, consciousness': though conscious, the person lacks awareness of the memory that influences performance. In fact, the person is not even aware that memory is involved.

Types of Implicit Memory

Research on neurologically intact people and on people with amnesia has shown that implicit memory is not unitary but rather consists of a variety of different subtypes.

Although a detailed description of all them is not possible, we mention three types that have been identified and about which we know a great deal: 'perceptual implicit memory or priming, conceptual implicit memory, and procedural memory.' Perceptual implicit memory is concerned with the perceptual, and possibly sensory, attributes of the stimulus, such as modality (e.g., auditory vs. visual) and within each modality, about perceptual and sensory qualities such as shape, color, viewpoint, size, and location. Conceptual implicit memory, on the other hand, is concerned more with the meaning of the stimulus, without regard to its perceptual qualities – for example, does the stimulus represent a person, regardless of whether the representation is pictorial or verbal, visual, or auditory. In tests of both types of implicit memory, participants are first exposed to stimuli and at test they are asked to process a target without

reference to its prior occurrence (explicit memory). On tests of perceptual implicit memory, the participant may be asked to identify the target, by naming it, if it is a picture or word, or by completing a degraded version of it. Performance, as measured by speed or accuracy of the response to the test stimulus, will vary primarily according to how perceptually similar the test item is to the target, but not by conceptual similarity. Increases in accuracy or decreases in response latency to old studied items as compared to ones presented for the first time at test are indicative of implicit memory for the studied items. By comparison, on tests of conceptual implicit memory, the participant may be asked to classify the target semantically, or produce it in response to a semantic question. For example, having studied the word 'horse,' the participant may be asked to make a semantic decision to the word at test, or asked to produce it in response to the word 'animal.' Performance on conceptual tests, as measured by accuracy and latency, is influenced by semantic relatedness rather than perceptual similarity.

In studying implicit memory, great care must be taken to insure that performance on tests of memory that are ostensibly implicit are not contaminated by explicit components, such as might occur when asking participants to identify degraded stimuli they had studied earlier. One can identify the item by perceptual or semantic processing alone, or one can resort to an explicit memory of the studied item to aid performance. A number of procedures have been developed to reduce or eliminate contamination from explicit memory, among them verifying that the person is not aware that the response or test item refers to a studied stimulus, giving a comparable explicit test and seeing if performance is different (usually better) under these conditions, and dissociating implicit from explicit memory by requiring responses that could not be made accurately if explicit memory were implicated (e.g., do not respond with a word you had studied earlier). Because the amnesic person's explicit memory is so poor, equivalent performance between amnesic and intact people on an implicit test is taken as evidence that the test in question was not contaminated by explicit memory.

When the implicit test is not contaminated by explicit memory, the general finding across perceptual implicit tests in various modalities and across conceptual implicit tests, is that implicit memory is preserved in amnesic people, and often is indistinguishable from that of neurologically intact controls, though the amnesic person's performance on comparable tests of explicit memory is severely impaired. That is, amnesic people show a consistent advantage in processing old items compared to new ones on implicit tests, as do controls, despite amnesic people having little or no explicit memory of the items, and in severe cases, even of the study episode.

Tests of procedural implicit memory involve learning rules, motor sequences, conditioned responses, and subjective probabilities of stimulus–response or stimulus–stimulus associations or occurrence. As with perceptual and conceptual implicit memory, amnesic people perform normally, and show the benefits of experience and practice, without an explicit memory of even having performed any of the tasks, or of having encountered the stimuli and emitted the responses that constituted the tasks.

The Neuroanatomical and Functional Locus of Implicit Memory

As we noted, organic amnesia typically results from bilateral damage to the medial temporal lobes and parts of the diencephalon. Being damaged, these structures, so crucial for explicit memory, cannot support normal performance on tests of implicit memory. Instead, research on implicit memory suggests that performance is mediated by the very structures involved in stimulus analysis and perception (perceptual), in extraction of meaning (conceptual), and in response implementation and execution (procedural). With respect to perceptual implicit memory, these structures are the perceptual modules or structural representation systems in posterior neocortex that include higher-order visual areas in extrastriate and inferotemporal cortex, and higher-order auditory areas in the temporal lobes. Because of their inherent plasticity, these modules are modified by the act of processing some given material, leaving behind a record of that process. As a result, processing is faster and more accurate when the system is reexposed to that material as compared to new

material. The modules are domain specific, in that separate ones exist for objects, faces, for auditory and visual words, and possibly places, and so forth. Changes in one module are not transferred to another. This accounts for some of the specificity of perceptual implicit memory or priming – seeing an object will prime perception of that object but not of the word that denotes it, and vice versa. The modules are also presemantic in that they do not represent the meaning of the item but only its structure. Thus, performance on perceptual implicit tests is sensitive to changes in perceptual or structural aspects of the stimulus but not to changes in semantic aspects.

The opposite holds for performance on conceptual implicit tests which is mediated by conceptual systems in the lateral temporal lobe and inferior frontal cortex. As with perceptual modules, improvement in performance results from modifications in the conceptual systems themselves. Because conceptual implicit memory is sensitive to meaning and not to perceptual attributes, facilitation can be observed even when the stimuli change perceptually from study to test, as long as some common aspect of meaning is preserved.

The structures that have been identified as crucial for procedural learning are the cerebellum for classical conditioning, and the basal ganglia for the many of the other tests implicated in learning motor sequences, subjective probabilities and stimulus–response associations, with some indication of prefrontal involvement if strategic, sequential, or inhibitory components play a role. It is assumed that changes in these structures during planning and execution of procedures underlie the facilitation of performance.

What Function Does Implicit Memory Serve?

As there are many types of implicit memory, the function will depend on each type. What many of them have in common, however, is the formation, tuning, and priming of perceptual, semantic, and procedural structures with repetition, so that they can implement their operations efficiently and automatically when confronted with similar stimuli and situations. Thus, for example, implicit memory underlies some of the increased efficiency with repetition in reading words, identifying objects, acquiring stimulus–response associations, emitting response sequences, and learning probabilities of occurrence of some stimuli or sequences of them. In each of these cases, performance is driven by the stimulus or situation itself without regard to conscious awareness of where and when the stimulus or situation was encountered previously. Though such information is useful under some conditions (see below), under others it is often counterproductive to depend on it. In order to ride a bike or swing a racket properly (procedural memory), it is not necessary to remember consciously the episode in which we learned to do so, and it sometimes may be harmful to rely on such information to aid in executing the task. What is important is to modify the very structures that implement perception and action.

Types of Explicit Memory

Like implicit memory, explicit memory also is not unitary. Psychologists distinguish between two main types: 'semantic memory,' which refers to general knowledge such as vocabulary and facts about the world, and even facts about oneself, and 'episodic memory,' which refers to memory for past events associated with a particular autobiographical experience, such as a memorable day on vacation or an embarrassing episode. According to Tulving, each type of memory is also associated with, or dependent on, a distinct type of awareness or consciousness. Semantic memory is associated with 'noetic awareness,' which occurs when one thinks of something one knows, whereas episodic memory is associated with 'autonoetic consciousness,' which occurs when one remembers a particular experience, an autobiographical event. At the core of autonoetic awareness is the subjective, personal experience associated with the self, "which confers the phenomenal flavor that distinguishes remembering from other kinds of awareness" (Tulving, 1985: 3).

The distinction between autonoetic and noetic awareness is crucial to understanding different aspects of recognition memory. Basing himself on research conducted in the late 1970s and early 1980s, Tulving proposed a distinction between two aspects of recognition memory, 'remembering' and 'knowing.' Remembering involves reexperiencing or

reliving a past event in the mind, what he and others have termed 'mental time travel.' Characterized by recovering and recreating the context in which a stimulus or event occurred, it is a hallmark of true episodic memory. Knowing, on the other hand, is associated with a sense of recognizing or experiencing a stimulus or event as old but with little or no information about the context in which it was encountered. Though knowing refers to recognition of a memory associated with an episode, it has much in common with semantic memory. Because remembering and knowing are not pure processes, investigators refer to 'recollection' and 'familiarity,' respectively, as the processes that underlie them

Different Types of Explicit Memory, Their Neuroanatomical Substrates, and Dissociable Deficits in Amnesia

It has long been known that semantic memory and, by implication, noetic awareness is spared, for the most part, in amnesia, whereas episodic memory and autonoetic awareness is profoundly impaired, and forms the core of the amnesic deficit. Recent investigations, however, have linked recollection and familiarity to different regions of the medial temporal lobe. The hippocampus and its projections and associated diencephalic structures are implicated in recollection, and the perirhinal cortex and its projections and structures are implicated in familiarity. Thus, patients with damage restricted to the hippocampus, are impaired at recollection, but familiarity is relatively spared: They sometimes consciously can recognize that they had encountered an item at study, and distinguish it from an item encountered for the first time at test but they cannot, however, recover the particular context in which they encountered the item. In everyday life, this dissociation between recollection and familiarity manifests itself as an ability to 'know' that some event had occurred in the past, and sometimes to recognize items as familiar, but without the ability to reexperience the past event in any detail or recapture the context in which the familiar item was encountered. It is akin to our being introduced to someone at a party or at work, and later seeing that person in a different venue, recognizing the person as familiar, but not being able to place where the person was encountered previously.

Some investigators have claimed that this deficit in recollection applies not only to memories acquired since the onset of the amnesia ('anterograde amnesia') but also to memories acquired long before ('retrograde amnesia'), sometimes encompassing the patient's entire lifetime. Such patients 'know' that certain events happened to them (their wedding, graduation, etc.), but they cannot recollect these events; that is, they cannot travel mentally back in time and reexperience or relive them in any detail.

Patients also have been reported with damage that primarily affected the perirhinal cortex but spared the hippocampus. These patients have deficits in familiarity, though their recollection is intact. In other words, if they recognize a person or a stimulus as old, they do so because they can also recollect the context but they have little sense of familiarity without recollection.

To summarize, explicit memory associated with autonoetic awareness, namely explicit memory that allows one to recollect the past so as to have a conscious sense of reexperiencing it in great detail, depends on the hippocampus and related structures. It is particularly this conscious memory that is deficient in amnesia. Explicit memory associated with noetic awareness, either with respect to semantic memory or to familiarity, is dependent on the perirhinal cortex and its related structures, and it, too, can be absent if medial temporal damage extends to this region. Typically, both medial temporal structures are affected in organic amnesia and, consequently, so are both recollection and familiarity. Semantic memory depends on lateral and temporal neocortex, as well as inferior prefrontal cortex, and therefore, is spared in such patients.

Why is Memory with Conscious Awareness Associated with the Medial Temporal Lobes, and the Hippocampus in Particular?

The hippocampus is neuroanatomically situated so that it receives input from neocortical regions necessary for having conscious experiences of events. Through projections back to the regions that mediate this conscious experience, the hippocampus

encodes into a memory trace those co-occurring neural elements that mediate the conscious experience. The encoding of associations among diverse elements into a memory trace is called 'relational binding,' and is generally acknowledged to be a function of the hippocampus. The memory trace, therefore, consists of a bound ensemble of hippocampal–neocortical neurons. Because consciousness is an aspect of the experience, the neural elements that mediate consciousness are bound into the memory trace along with those that mediate the content of the experience. At retrieval, consciousness is recovered along with content, allowing for a rich conscious reexperiencing of the past. Damage to the hippocampus, therefore, not only prevents relational binding needed for the formation of new memory traces, but also impairs the recovery of relationally bound memories that depend on the hippocampus for their representation. As a result, conscious recollection of the past is not possible when the hippocampus is severely damaged, though memory of unbound elements of an experience, those single items that are recognized as familiar, should be preserved.

Another possible function of the hippocampus, derived from 'cognitive map theory,' is the encoding of allocentric representations of space. Allocentric representations are defined according to the relations that different spatial elements bear to each other as compared, say, to an egocentric (viewer centered) representation. Such representations are crucial components of any episodic memory, because all episodes occur in a particular place. The hippocampus is necessary to construct the scenes in which events take place and to recover that information at retrieval. Relational binding may still be the process that is involved in forming these allocentric representations and in binding the nonspatial content of the experience to them.

Recent Developments and New Directions

The Interaction between Implicit and Explicit Memory

New findings suggest a hitherto unexpected contribution of the medial temporal lobes, and hippocampus in particular, to perceptual and conceptual implicit memory. Such findings go against the hypothesis assumed throughout this paper that the domain in which the hippocampus operates is concerned only with memory with conscious awareness. Let me illustrate with a couple of examples.

A person is shown a picture of a detailed scene. After a delay, the scene is altered so that one of the elements is removed or its location is changed. Healthy controls are not consciously aware of these changes. Examination of their eye-movements suggest, however, that they have an implicit memory of the previous location of the altered item since they look at that spot more than at others in the picture. Interestingly, amnesic people do not show this bias to look at the changed location. These results are interpreted as indicating that the hippocampus is important for relational binding, regardless of whether one is consciously aware of the information that was bound in memory.

A similar conclusion has been reached by investigators who have examined the specificity of priming. They noted that it is not simply the repetition of the stimulus that is important, but also the response or decision associated with it. For example, consider a task in which a size decision is required when an item is first presented, and the same decision, or a different one (living-made) when the item is presented next. Priming is greater when the decision is the same than when it is different, even though it is the same item that is repeated. The same is true if only the direction of the size comparison changes from first presentation (bigger than a ___) to the second (smaller than a ___). In some cases, priming is eliminated entirely when the decision changes from first to second presentation. Importantly, this advantage for the item that is repeated along with the same decision is not seen in amnesic patients suggesting that the formation of associations between stimuli and decisions that underlie some forms of perceptual and conceptual priming may be dependent on the medial temporal lobes. Although one may consider this an instance of implicit memory being contaminated by explicit memory, that does not seem to be the case because the neurologically intact controls had no explicit memory of the association between the decision and the stimulus.

These, and related studies, open up new avenues of investigation on the interaction between

implicit and explicit memory. Past research was concerned with separating implicit from explicit memory, so as to enable investigation of one type of memory in isolation from the other. That strategy, an excellent one if one wants to understand the characteristics of each type, also created a situation that is artificial. The time seems ripe to remove these boundaries and examine the influences that one type of memory can have on the other, and how they are implemented.

What Function Does Recollection Serve?

Why do we need to reexperience the past consciously and in such great detail? One possible answer is that it allows us to have a sense of who we are and how we got to be that way. However, studies of amnesic patients, who have impaired recollection, have a good sense of who they are, suggesting that retaining a sense of self cannot be a major function of recollection. I am not aware of studies that examined on what basis amnesic people and neurologically intact individuals have these impressions of themselves. Perhaps amnesic people rely more on semantic memory than do individuals whose memory is intact.

More recently, a number of investigators have examined the implications of the notion that the hippocampus is necessary for mental time travel. If that is the case, it should make little difference conceptually which temporal direction one is traveling, into the past or into the future. A number of studies of amnesic patients have confirmed this prediction by demonstrating that they find it as difficult to imagine a detailed scenario of the future as they do to recollect an experience of the past. Both types of memories involve reconstructive processes dependent on the hippocampus. Perhaps one needs the ability to recover detailed past memories in order to imagine the future, or perhaps the relational binding and scene construction function of the hippocampus is needed to form new, imagined events or scenes, as it is to reconstruct past memories. New research is addressing this issue. Whatever the outcome, these findings suggest that the ability to form detailed, conscious memories of the past allows us to imagine new events, and in doing so, help us solve problems, plan for the future, and perhaps even, create works of art.

See also: Neurobiological Theories of Consciousness.

Suggested Readings

Aggleton JP and Brown MW (1999) Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. Behavioral and Brain Sciences 22: 425–489.

Cohen NJ and Squire LR (1980) Preserved learning and retention of pattern-analysing skill in amnesia: Dissociation of ''knowing how'' and ''knowing that''. Science 210: 207–209.

Eichenbaum H, Yonelinas AP, and Ranganath C (2007) The medial temporal lobe and recognition memory. Annual Review of Neuroscience 30: 123–152.

Gabrieli JDE (1998) Cognitive neuroscience of human memory. Annual Review of Psychology 49: 87–115.

Hamilton CW (1976) Basic Limbic Anatomy of the Rat. New York and London: Plenum.

Hassabis D, Kumaran D, Vann SD, and Maguire EA (2007) Patients with hippocampal amnesia can not imagine new experiences. Proceedings of the National Academy of Sciences USA 104: 1726–1731.

Hassabis D and Maguire EA (2007) Deconstructing memory with construction. Trends in Cognitive Science 11: 299–306.

Mayes A, Daniela Montaldi D, and Migo E (2007) Associative memory and the medial temporal lobes. Trends in Cognitive Sciences 11: 126–135.

Moscovitch M (1992) Memory and working with memory: A component process model based on modules and central systems. Journal of Cognitive Neuroscience 4: 257–267.

Moscovitch M (2000) Theories of memory and consciousness. In: Tulving E and Craik FIM (eds.) The Oxford Handbook of Memory, pp. 609–625. Oxford: Oxford University Press.

Moscovitch M (2008) The hippocampus as a ''stupid,'' domain-specific module: Implications for theories of recent and remote memory. Canadian Journal of Experimental Psychology 62: 62–79.

Moscovitch M, Vriezen E, and Goshen-Gottstein Y (1993) Implicit tests of memory in patients with focal lesions or degenerative brain disorders. In: Boller F and Grafman J (eds.) The Handbook of Neuropsychology, vol. 8, pp. 133–173. Amsterdam: Elsevier Science Publishers.

O'Keefe J and Nadel L (1978) The Hippocampus as a Cognitive Map. Oxford: Oxford University Press.

Roediger HL III, Rajaram S, and Geraci L (2007) Three forms of consciousness in retrieving memories. In: Zelazo PD, Moscovitch M, and Thompson E (eds.) The Cambridge Handbook of Consciousness, pp. 251–287. Cambridge, UK: Cambridge University Press.

Ryan JD, Alhoff RR, Whitlow S, and Cohen NJ (2000) Amnesia is a deficit in relational memory. Psychological Science 11: 454–461.

Schacter DL, Wig GS, and Stevens WD (2007) Reductions in cortical activity during priming. Current Opinion in Neurobiology 17: 171–176.

Slotnick SD and Schacter DL (2007) The cognitive neuroscience of memory and consciousness. In: Zelazo PD, Moscovitch M, and Thompson E (eds.) The Cambridge Handbook of Consciousness, pp. 251–287. Cambridge, UK: Cambridge University Press.
Suddendorf T and Corballis MC (2007) The evolution of foresight: What is mental time travel, and is it unique to humans? Behavioral and Brain Sciences 30: 299–351.
Tulving E (1985) Memory and consciousness. Canadian Psychologist 25: 1–12.
Tulving E (2002) Episodic memory from mind to brain. Annual Review of Psychology 53: 1–25.
Yonelinas AP (2002) The nature of recollection and familiarity: A review of 30 years of research. Journal of Memory and Language 46: 441–517.

Biographical Sketch

Morris Moscovitch received his BSc at McGill University in 1966 and his PhD at the University of Pennsylvania in 1972. He is a member of the psychology department at the University of Toronto since 1971, and of the Rotman Research Institute and Psychology Department of Baycrest Centre for Geriatric Care, he now holds the Glassman Chair in Neuropsychology and Aging. He conducts research on the cognitive neuroscience of memory, attention, and face-recognition. He is a fellow of the Royal Society of Canada; in 2007, he received the D.O. Hebb Award from CSBBCS (Canada) and the William James Award from the Association for Psychological Science.

Consciousness of Time and the Time of Consciousness

S Gallagher, University of Central Florida, Orlando, FL, USA; University of Hatfield, Hertfordshire, UK

Glossary

Act–content model – A description of consciousness that distinguishes between the differences in conscious acts (of perception, memory, imagination) and (representational or intentional) content.
A-series – McTaggart's term for the temporal change from future to present to past.
B-series – McTaggart's term for the permanent temporal order of events signified by the terms 'earlier than' or 'later than.'
Diachronic coconsciousness – Consciousness of successive contents which are nontransitive. A and B are experienced together, and B and C are experienced together, but A and C are not experienced together. Contents are experienced as present together, but not as simultaneous (cf. synchronous coconsciousness).
Flash-lag effect – A visual illusion in which a light flash and a moving object that coincide in the same location are perceived to be in different places.
Longitudinal intentionality (***Längsintentionalität***) – Prereflective awareness of one's just previous experience, which supports the 'transverse' intentional awareness of the content or object of that experience, and also generates the sense that the experience is one's own.
Phi phenomenon – A visual illusion of motion produced by a succession of still images first described by Max Wertheimer. The color *phi* phenomenon involves an illusion in which the 'moving' object seemingly changes color.
Protention – Husserl's term for the intentional aspect of consciousness responsible for providing an anticipatory sense of what is just about to be experienced.
Retention – Husserl's term for the intentional aspect of consciousness responsible for providing a sense of just-past experience
Specious present – A term coined by Robert Kelly (aka E. R. Clay) and used by William James and others to signify that our experience is not of a momentary duration but extends over time.
Synchronous coconsciousness – Consciousness of simultaneous contents as transitively simultaneous. If I am aware of A and B as simultaneous, and of B and C as simultaneous, then I am aware of A and C as simultaneous (cf. diachronic coconsciousness).
Transverse intentionality (***Querintentionalität***) – An awareness of the content or object of experience that is just past, which depends on the 'longitudinal' intentional awareness of one's just previous experience.

Introduction

In almost all cases when we are conscious of something we are at least implicitly aware of time passing. Our consciousness of time, however, is integrally related to the temporal structure of consciousness itself. As I consciously read this sentence my initial awareness of the first part of the sentence passes from present to past as the next part comes into view. This too becomes past as I move on to the next sentence. The sentences themselves, as text on paper, do not disappear into the past – they are still present on the page and can be made present in my consciousness again by rereading. But this rereading is not the original reading, the original consciousness of the

text, which is now firmly something that happened in the past. Each moment of my experience seems to be present for only a moment, and then to slip away into the past, even if the object that I am conscious of remains present and unchanging. Consciousness seems to involve a successive flow, and the often-used metaphor of a stream of consciousness seems very appropriate. Things become more complex if the object of which I am conscious is itself a temporal object, something that undergoes constant and obvious change, such as a melody. In that case we have two successions to explain – the successive flow of consciousness and the succession of the temporal object. In some way the flowing retreat of consciousness is able to maintain an orderly sense of the melody that runs off in time.

One question of interest to philosophers and psychologists especially in the nineteenth and early twentieth centuries concerned the relationship between the consciousness of succession and the succession of consciousness. Do we have an immediate experience or perception of succession, or is succession a reflective abstraction that derives from the passage of our experiences? This is a question about time perception. Psychological experiments were conducted to explore the maximum and minimum durations that we can sense, to identify the average unit of temporal experience, and to differentiate awareness of simultaneity and succession. These experiments resulted in no general agreement on these points but suggested that on all of these measures there are significant variations between individuals, and even within an individual depending upon the state of alertness or fatigue. Experiments also show that the experience of duration is different between different sense modalities within the same person.

Duration in these experiments is measured by a clock. Most psychological studies assume a linear concept of time and they often ask about the perception of moment-to-moment change in the environment, how perceptual consciousness properly estimates or conforms to that linearity, or how accurate, compared to a clock, our consciousness is. Time estimation studies, for example, show that one's sense of time passage does not always match up to clock time. This motivates the distinction between objective time, which is time objectively measured by the clock, versus psychological (or phenomenological) time, which is time as it is subjectively experienced. Good conversation can sometimes make time pass too quickly; time slows down when you are bored; time flies when you are having fun; and so forth. This is the way it seems subjectively. Psychological explanations of such phenomena are often made in terms of the quantity of the number of stimuli, mental events, perceptions, and memories that might be occurring during a conscious interval. How I experience a specific duration (objectively measured) is inversely proportional to the total amount of experience (numbers of sensations, perceptions, or amount of mental content, etc.) that takes place within the interval. The effects of attention and memory, as well as fatigue and emotional state, may qualify this. Furthermore, one should consider not just the quantity but also the possible semantic or emotional effects of content, that is, more meaningful or more salient content may affect the way we experience time. It is also well known that chemical changes in the brain caused by the ingestion of drugs, or the chemical imbalances associated with certain illnesses and psychopathologies, can alter the experience of time. Psychotomimetic drugs, such as LSD, mescaline, and psilocybin, produce physiological changes evidenced in a desynchronizing of EEG rhythm, increased muscle tension, and acceleration in the rate of metabolism. The phenomenological effect is that time seems to either speed up or slow down. Subjects may also suffer from disordered temporal experience, that is, abnormalities in time estimation and temporal orientation. Cannabis intoxication, for example, results in increased heart rate and abnormal EEG rhythms, disturbances in perceptual experience, and distortions in time estimation and time perspective, including flashbacks.

There is a long-standing interest in philosophy in the subjective experience of time, the discussion of which can be found in texts as far back as Aristotle and Augustine. Discussions of our consciousness of time focus on a number of perplexities. For example, although from one perspective I am always experiencing the present (it is always in some sense now) from a different perspective this now is constantly changing into something not-now as it moves into the past. Such considerations also

motivate a series of fascinating investigations into the perplexities of how we are able to experience time at all. Consider, for example, listening to your favorite song. At any one moment of time the only bit of the melody that exists is the note that is currently being played, and it is therefore the only bit of the melody that we can perceive in that moment. Notes that were previously played are no longer there to be heard; notes that will be played later in the melody do not yet exist. So literally it is possible to hear only one note at any one moment. At the next moment we hear the next note (or possibly a different part of the same note if its duration lasts several moments). If at any one moment we hear only one note or one bit of the melody, and are no longer hearing past notes, and not yet hearing future notes, then it does not appear to be possible to actually hear the melody itself. We hear one note, and then another, and then another – but we do not hear the succession of notes. It is difficult to see how we can have a direct perception of succession or of duration if this is the case, and yet we certainly seem to have such experiences. Is it possible that our experience of time is an illusion? Furthermore, such considerations demonstrate the importance of the question of time for understanding consciousness, for if our experience of a melody were exactly as described, our experience of anything would be of a similar kind – momentary experiences without continuity or coherent meaning.

The Purely Psychological Nature of Time

That our access to objective time depends in some way on our consciousness of it motivates some philosophers to ask whether time itself may be purely psychological and therefore not real. On this view objective time is at best a convention. In his 1908 article on 'The Unreality of Time,' McTaggart argues for the unreality of time in this way. With regard to the concept of time he introduced an important distinction between the A-Series and the B-Series. The A-series is signified by the terms 'past,' 'present,' and 'future'; the B-series by the terms 'earlier' and 'later.' The meaning of the term 'earlier' is not equivalent to the meaning of 'past,' nor is 'later' the same as 'future.' Rather, one event can be earlier than another, and both events can be in the past, or both in the future. Conceptually McTaggart defines the A-series as involving change, movement, or becoming. An event starts out in the remote future, is then in the near future, moves into the present, quickly becomes past, and continues to undergo temporal change in the sense that it becomes more and more removed from the present as it fades into the remote past. The B-series, in contrast, does not involve change. Once an event is earlier than another it remains that way. Events do not change their temporal position from earlier-than to later-than. The B-series thus expresses the permanent order of events. The Russian Revolution happens earlier than China's Cultural Revolution, and that order will never change. The B-series, then, signifies what we usually regard as the objective, chronological order of irreversible temporal succession.

McTaggart argues that the B-series is derivative from the A-series, but that the A-series cannot be found in reality because it involves a contradiction: any event is characterized by all of the incompatible determinations: past, present, and future. A future event, for example, will be present and then past. He considers the obvious response that no event has incompatible A-characteristics at the same time – it can only have them successively. This seems obvious from the required use of different verb tenses in statements about how a particular event can have such incompatible determinations. But, according to McTaggart, this response is viciously circular. Tense depends upon the A-series, and we end up escaping the initial contradiction of incompatible temporal determinations by an appeal to either a second-order contradiction (incompatible tenses) or an infinite regress from A-series to A-series. McTaggart solves this paradox by proposing that time is purely a matter of our experience. The distinctions of past, present, and future arise on the basis of consciousness and the cognitive functions of memory, perception, and expectation, respectively.

McTaggart's argument, however, assumes the principle of noncontradiction: a thing cannot be and not be at the same time. Clearly, however, this principle depends on the reality of time. If time is not real then the principle that McTaggart assumes in the demonstration of its unreality is

itself undermined. Others philosophers have tried to reformulate the principle of noncontradiction without presupposing time. Much of the critical discussion following McTaggart focuses on the status of tensed versus tenseless statements, or propositions and their truth conditions considered as objective facts, and in this regard is not directly relevant to consciousness.

The Specious Present

A number of philosophers accept the challenge of explaining how it is possible for us to be conscious of something like a melody, and one of the central concepts to be developed in this regard is the notion of the specious present. The term comes to us from Robert Kelly the anonymous author of *The Alternative: A Study in Psychology*, a work credited to E. R. Clay, which is the way William James cited it when he introduced the term into the mainstream philosophical discussion. The specious present doctrine consists of the claim that the present or now that we experience at every moment is not a knife-edge or punctate phenomenon, but includes a brief extended interval of time – a bit of the past and a bit of the future. The strict or real present is just the momentary piece of the now that is present; but this is always supplemented in consciousness by margins or penumbral horizons of the past or future. When I listen to a melody, for example, I hear not just the note that is currently being played, but I hear it in some way accompanied by some number of previous notes, and, perhaps, some number of notes that are to be played in the next seconds.

This direct experience of succession is not a matter of perceiving one note and supplementing it with the memory of a previous note (as philosophers like Thomas Reid and Franz Brentano had proposed). This can be made clear if we consider visual perception and the difference between perceiving the hour hand of the clock and the second hand of the clock. In perceiving the hour hand I get a sense of its movement only by comparing its current position to a memory of where it was a minute or two ago. In contrast, I can actually see the movement of the second hand and this does not seem to involve a comparative judgment based on memory. Where the second hand was a second ago seems to be intuitively present in my perception of its movement.

This concept of the specious present is meant to address an issue that is fundamental for understanding consciousness. A consciousness that did not have the kind of structure described by the specious present would seemingly be an experience of only one unconnected moment after another. In that case our experience of the world would be incoherent and inchoate. The idea that memory could bring coherency to this kind of experience, connecting together a set of discontinuous flashes of perceptual consciousness, is questionable simply because on such theories memory is itself a form of consciousness and would have no more intrinsic structure than perception. My memory of a melody can be coherent only if it is an awareness of more than one note at a time, that is, only if it is more than a series of knife-edge presents and it takes the form of the specious present.

The analysis presented by William James and his followers understands the specious present as a sensed or immediately experienced duration or succession. It explains that the content of consciousness has a temporal coherence, but it does not explain precisely how this is possible. Little attention is paid to the sensing or experiencing itself, the experienced content of which has the specious present structure. It does not explain whether or how the temporal structure of consciousness itself contributes to the temporal coherence of the experienced content. Indeed, a number of theorists understand the specious present to involve a momentary (nontemporal) act of awareness. But the idea of a momentary act of awareness brings along a number of perplexities that have motivated some theorists to reject the specious present doctrine.

For example, assume that my perception of a succession of events (VWXYZ) is laid out in the specious present form so that I am still aware of V and W as just past while I am currently aware of event X as present. If my perception is itself momentary, which means that for V, W, and X to be represented as occurring in succession they must be represented simultaneously, we must experience them all at once. On such an account of the specious present, to be aware of successive

objects consciousness needs to compare the earlier and later objects in a cognitive operation that makes the earlier and later simultaneous. Experienced content would, in some fashion, need to be at once both successive and simultaneous, both past and present. Since this is paradoxical, the critics argue that the specious present doctrine must be rejected.

Another problem with the specious present doctrine can be made clear by looking at C. D. Broad's account (see Figure 1). According to Broad, in any one moment of conscious experience (A) we are aware of a specific duration of contents, V – X. At the next moment of consciousness (B) we are aware of W – Y. One objection to this is to point out that we seemingly experience W – X twice in succession: once in momentary consciousness A and once in momentary consciousness B. If VWXY represents a melody, then we seemingly hear the notes of that melody twice.

To escape this problem, however, we simply and realistically have to think that perception (or any act of consciousness) itself has duration and is not momentary. If we do this, however, we run into a different problem, as pointed out by J. D. Mabbott in his critique of Broad. Mabbott assumes that the enduring conscious act A–B corresponds to the overlap of the two originally defined specious presents. In other words, the specious present of A–B is W–X. But this leads to the absurdity that the specious present varies inversely with the duration of the act of consciousness. Assume that the specious present of the originally defined momentary act of consciousness A is 6 s long. That would make the specious present of A–B, the overlap W–X, 3 s long. Even though A–B is itself longer than the momentary act A by 3 s, its specious present seems to have shrunk. If the act of consciousness lasts 6 s (e.g., A–C) the specious present shrinks to momentariness at X. Thus, conscious acts of 6 s or longer would have no specious present. The longer the duration of the perception, the shorter the specious present. And this is like saying, the longer I look at something, the less I see it. Does this mean that we should reject the specious present doctrine?

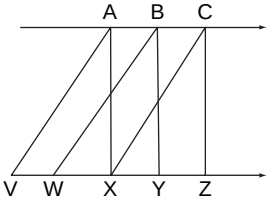


Figure 1 Broad's diagram of the specious present.

A different interpretation of Broad's diagram would avoid the absurdity. If we assume that the duration of the specious present remains constant no matter how long the duration of the act is, then we could conceive of the specious present as having the structure of a searchlight moving along the upper line in Figure 1, illuminating a constant duration along the lower line. This would avoid any logical absurdity. Still, there are empirical problems with this solution. Empirical studies indicate that the specious present does not remain constant, but varies even within a single individual. The searchlight widens or narrows depending on certain conditions. This seems consistent with the psychological issues described above – boredom, enjoyment, fatigue, attention, etc.

Empirical research also suggests that the specious present varies across different sense modalities (sight, touch, and hearing) even within the same individual. For example, intervals of auditory stimuli are experienced as lasting longer than objectively equal intervals of visual stimuli; visual experiences may vacate the specious present faster than auditory experiences. This might seem theoretically disconcerting. If, as in many instances, we experience through more than one sense modality simultaneously why does not our experience seem seriously incongruent? For example, when I watch a ballet, if my auditory specious present is not identical with my visual specious present, then the music could appear to be out of sync with the dancer's movements. The dancer would always be a little behind or ahead of where I think she ought to be. Obviously, since this does not appear to be the case in our actual experience, assuming we attend good ballet, then either the intersensory differences are resolved in some fashion or the specious present doctrine is wrong.

To preserve the specious present doctrine one can appeal to subpersonal processes that effect a temporal binding across different sense modalities. I see lightening before I hear the thunder. The physical differences in the relative arrival time of stimuli at the eye and ear can be accounted for by

the fact that light travels through air faster than sound: 300 000 000 versus 330 m s^{-1}. Although the transduction of sound waves at the ear takes less time than the chemical transduction of light at the retina, this solves the problem only for intermodal perception of simultaneous visual-auditory events 10 m away from the perceiver. Something else must account for all events that are closer or farther than 10 m and that are correctly synthesized across modalities. Neuroscientists (e.g., Varela, Pöppel) distinguish between:

- (S) neuronal system statesof approximately 30 to 100 ms, corresponding to a quantum of experience where differentiation of succession is not possible (note, however, that this may go as high as 250 ms in some cases, e.g., the difference between auditorily perceived speech and visually perceived lip movements has to be greater than 250 ms for the asynchrony to be perceived) and
- (W) a 0.5 to 3 s temporal window, which correlates with an experienced specious present, in which some neuronal mechanism effects a temporal integration of these S-states into a successive order.

The idea is that integration and ordering processes up to W are automatic and content-independent. In contrast, semantic binding comes into effect in the generation of W, the 3-s time scale. In other words, at or beyond the magnitude of W the coherency (or lack of coherency) of phenomenal experience may depend on content. For example, the often cited experiences of time seeming to slow down or speed up, when subjects are, respectively, bored or having fun, may be just such cases in which the specious present structure is affected by content.

The binding processes at the S scale (actually, the simple limitations of the system to discriminate succession at that scale) can explain why intersensory differences (nonsimultaneous processing of information across the various senses) do not show up in phenomenal experience. Intersensory differences are so small that they are integrated (or simply fail to count) at the S scale. This model may even help to explain some intrasensory perplexities. In the color phi phenomenon, for example, a subject is presented with two spots of differently colored light (e.g., blue and red), lasting 150 ms each, and flashed in 50 ms sequence. The time frame of the presentation is such that the perceiving subject experiences, not two separate dots in sequence, but a moving dot that changes color in midstream. If a spot of blue is flashed at point A and a spot of red is flashed 50 ms later at point C, the effect is that the subject is conscious of the red at a point B, between A and C, and at a phenomenal time that seems prior to the time the second color was actually flashed at C. The end point of the event seems to gain some representation at the midpoint of the experience; the subject sees red before there is red to be seen. One can make sense of this as follows. Assume that successive neural system states S1 and S2 have magnitudes of 60 ms each, and that S1 begins 35 ms prior to the beginning of the 50 ms interval between the flashing of a blue dot and a red dot. All phenomenal content that falls within the timeframe of S1 is experienced as simultaneous. S2 begins 25 ms after the onset of the 50 ms interval and ends 35 ms after the flashing of the red dot begins. If the phenomenal content that falls within the timeframe of S2 is experienced as simultaneous, and if, within W, the content that corresponds to S1 and S2 are experienced as successive, it seems clear that the perceiving subject will start to see the red dot when S2 begins near the midpoint of the interval and 35 ms prior to the flashing of the red dot. As Eagleman and Sejnowski have shown in regard to the flash-lag illusion, the neuronal processing that takes place in the first 80 ms after stimulus onset will determine how we visually experience movement and other temporal events.

Time Consciousness

One strategy for solving perplexities associated with the specious present is to appeal to subpersonal (brain level) explanations. A different strategy is to suggest that the idea of the specious present is not clear to begin with, or more precisely, that it is not a solution but is rather what needs to be explained. To say that a perceiving subject experiences a short bit of the past and future, and not just a knife-edge present is not to say precisely how this is possible. One approach to providing an explanation addresses two closely related questions: What

does experience have to be like for it to deliver a coherent experience of something like a melody? What does experience have to be like if at the same time it delivers a coherent sense that it is my experience that is developing over time? The latter question asks about the unity of consciousness itself and is related to the issue of personal identity.

Discussions of the specious present are based on the idea that consciousness involves an act–content structure, and on this view consciousness has a unity because at any one point in time its diverse contents fall under a single act of awareness. This is expressed in two assumptions that can be found in many discussions of the specious present. First, the perception of succession requires a momentary and indivisible, and therefore durationless act of consciousness. Second, succession is experienced only when a temporal spread of content is apprehended together, simultaneously, in consciousness. In some cases, these assumptions are adopted as a way to set the problem up, and then one or the other is abandoned in the solution (e.g., Broad). Consider a solution in which the acts of consciousness are not momentary. On this model the content that is apprehended by the temporally extended act of consciousness runs concurrent with the duration of the act. At any one moment, however, the act does not apprehend an extended stretch of content, so that the act is not an awareness of content that lasts longer than itself. To return to the example of a melody, this seems to suggest that corresponding to a succession of notes, there is an enduring act of consciousness in which, at each moment, I am aware of only one moment of the melody – something like a one-to-one, constant correspondence. This, however, is no solution, since it restates precisely the problem we started with. At any moment we experience one note of the melody; the question is how do we experience the succession of notes as a succession so that we experience the melody *per se*?

A more radical solution has been proposed, and that is to give up the act–content structure altogether (Dainton). The act–content structure that is often predicated of consciousness suggests that experience is necessarily like perception, or that even perception of the world, which is surely an embodied process, consists of an internal mental perception of contents. Consider the stream of consciousness as a set of experiences without an act–content structure. The experiences are lived through in a prereflective way that is prior to the act–content distinction. On this model, the unity of consciousness over time is due to relations between experiences that are intrinsic to consciousness itself. The specific relation at stake is the partial 'overlapping' of the phases of consciousness. The stream of consciousness is composed of interrelated phenomenal contents that are intrinsically conscious; that they are overlapping means that they are coconscious (they are experienced together) in a diachronic ensemble. The notion of diachronic coconsciousness of successive contents can be contrasted with synchronous coconsciousness of simultaneous contents. The latter is transitive – if I am aware of A and B as simultaneous, and of B and C as simultaneous, then I am aware of A and C as simultaneous. Diachronic coconsciousness is not transitive – A and B may be experienced together and B and C may be experienced together, but A and C are not experienced together. So in diachronic coconsciousness, conscious contents are experienced as present together, but in contrast to some versions of the specious present, not as simultaneous.

What does it mean to be present together? Dainton's overlap model appeals to the dynamic properties of the flowing character of consciousness. Phenomenal contents are not marked by intrinsic properties of being past or present; they appear as intrinsically dynamic, flowing together, and they do so because of the overlapping of experiences. The fact that experiences overlap, however, does not mean that I experience the overlap. I do not hear notes overlapping notes. The concept of overlap is meant to explain why, when I hear a melody, I hear a succession of notes, but in a way that they are dynamically interrelated. The notes qua physical entities sounded in the environment have their own objective temporal order. This order is reflected in consciousness, and in this sense we could say that the melody itself does some of the work in shaping our experience; but consciousness adds something more insofar as in experience the notes take on the intrinsic temporal character of consciousness, so that consciousness in its overlapping structure preserves the sense of successive order. We hear a melody because consciousness also does some of the work: as phenomenal contents the notes appear to cohere in a flow.

If we accept the overlap model, however, there are still some things that are not explained. The concept of the overlap itself remains somewhat obscure. Moreover, it is not clear how the concept of overlap explains the sense of conscious flow. If we say that the phenomenal contents have an intrinsic flow structure, is that anything more than saying that consciousness itself just has an intrinsic flow structure? In addition, there is no account of the anticipatory aspects of experience, or of self-awareness. These are the kinds of issues addressed by a phenomenological model of time consciousness (e.g., Husserl).

The phenomenological model keeps the act–content distinction. On this view the specious present describes the way the content is structured in experience, but the operations of the act of consciousness gives the explanation of how the specious present is generated. In this case, the specious present is not the solution to questions about the temporality of consciousness, but is taken to be one of the things that require explanation. On the phenomenological model, the act of consciousness is characterized by retentional and protentional functions, and these functions generate the prereflective flow-structure of consciousness. That is, my conscious experience includes a prereflective sense of what I have just been thinking (or perceiving, or remembering, etc.) and a prereflective sense that this thinking (perceiving, remembering, etc.) will continue in either a determinate or indeterminate way.

Consciousness always includes a narrowly directed intentional grasp of the now of whatever is being experienced in the moment, for example, the current note in a melody, the current word in an uttered speech, or the current phase of any enduring object. But this primary impression never happens in isolation; by itself it is an abstraction which cannot deliver an ongoing sense of an enduring object. Accordingly, consciousness normally includes two other structural aspects: the retentional aspect, which provides us with a consciousness of the just-elapsed phase of the enduring object, thereby providing past-directed temporal context; and the protentional aspect, which in a more-or-less indefinite way anticipates something which is about to be experienced thereby providing a future-oriented temporal context for the primary impression. For example, if we are listening to music, the retentional aspect of consciousness keeps the intentional sense of the previous notes or measures available even after they are no longer audible. Furthermore, as I listen I have some anticipatory sense of where the melody is going, or, at the very least, that the melody is heading toward some indeterminate conclusion. The protentional aspect of the act of consciousness also allows for the experience of surprise. If I am listening to a favorite melody and suddenly hear a wrong note, I am surprised or disappointed only because I have some kind of anticipation of hearing the correct note. If the melody is cut off prematurely, I experience a sense of incompleteness, precisely because consciousness involves an anticipation of what the imminent course of experience will provide, even if this remains relatively indeterminate.

According to this kind of account, retention is not a particular thing in consciousness that we hear; rather, through the retentional aspect we experience the just-past musical tones as just past, as part of the specious present. Furthermore, there is no simultaneity between the retentional aspect of consciousness (which is a current feature of consciousness) and that which is retained (which is just past). The just-past tones do not remain present in consciousness, like some reverberation; rather, they are experienced as something that has just happened, and so precisely as just past. Consciousness retains the sense of what has just been experienced, not by retaining the event itself, but by its tacit or prereflective awareness of the just-past phase of consciousness.

If we look at someone running across a field, our perception is not restricted to a durationless snapshot now of her movement – if it were, we would sense no movement at all. Perceptually, it is not as if the person suddenly appears out of nowhere in each new moment. On the one hand, we want to say that we actually perceive the person running rather than that we perceive her present position and then add to that the recollection of where she was a moment ago. We do not engage in an act of comparative remembering in order to establish the temporal context of her current position. On the other hand, it is not the case that all the previous parts of her movement remain perceptually present in the same way as her current position. If that were the case, she would perceptually fill the entire space she has just traversed. The past

movements do not remain visually present in some vague ghostly manner. Retention does not keep a set of fading images in consciousness. Rather, at any moment what we perceive is embedded in a temporal horizon. What I see is part of or a continuation of, or a contrasting change from what went before, and what went before is still intentionally retained so that the current moment is seen as part of the whole movement. Consciousness retains the just past with the meaning or significance of having just happened.

Consider the following diagram (Figure 2). The horizontal line CDEF represents a temporal object such as a melody of several notes. The vertical lines represent abstract momentary phases of an enduring act of consciousness. Each phase is structured by three functions:

- the primary impression (pi), which allows for the consciousness of an object (a musical note, for example) that is simultaneous with the current phase of consciousness;
- retention (r), which retains previous phases of consciousness and their intentional content; and
- protention (p), which anticipates experience which is just about to happen.

In the now-phase there is a retentioning of the previous phase of consciousness. The just-past phase includes a retentioning of prior phases. This sets up a retentional continuum that stretches back over prior experience (rCDE or more precisely, r (E + r[D + r{C}]). Importantly, the continuity involved in retention has two aspects. The first provides for the intentional unification of consciousness itself since retention is the retention of previous phases of consciousness. But since the prior phases of consciousness contain their respective primary impressions of the previously sounded notes, there is also established a continuity of the experienced object.

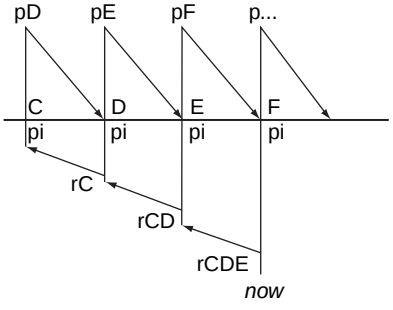


Figure 2 Husserl's model of time consciousness.

Again, consider the uttering of a sentence.

> Laura and David are going to Africa in order to . . .

When in uttering this sentence I reach the word 'Africa' I am no longer saying the previous words, but I still retain a sense of what I have just said. For a sentence to be meaningful, for a speaker or listener, the sense of the earlier words must be kept in mind in some fashion when I am uttering the later words. Retention (a feature of what cognitive scientists would call working memory) keeps the intentional sense of the words available even after the words are no longer audible. Built into this retentional function is the sense that I am the one who has just said these words, or more generally, I am the one who has just experienced this. The experience does not become part of a free-floating anonymity, nor does it seem to belong to someone else; it remains, for me, part of my stream of consciousness. In addition, at the moment that I am uttering 'Africa,' I have some anticipatory sense of where the sentence is going, or, at the very least, that the sentence is heading to some kind of ending. This sense of knowing where the sentence (the thought) is heading, even if not completely definite, seems essential to the experience I have of speaking in a meaningful way. It helps to provide a sense that I am speaking in a sentential fashion, and not speaking a meaningless set of phrases.

In addition, this protentional aspect of consciousness, like the retentional aspect, involves an implicit self-awareness. I am not only consciously anticipating the remainder of the sentence, or the melody, or whatever, but I am anticipating my experience of that which is about to happen. My anticipatory sense of the next note of the melody, or of where the sentence is heading, or that I will continue to think, is also, implicitly, an anticipatory sense that these will be experiences for me, or that I will be the one listening, speaking, or thinking. My experience of the passing or enduring object is at the same time a nonobservational, prereflective awareness of my own flowing experience. This retentional–protentional self-awareness delivers a sense that this experience is mine – that I am the one who is listening to the melody or uttering the sentence.

Husserl's phenomenological model distinguishes between longitudinal intentionality (Längsintentionalität), which delivers the sense that the experience is my experience, and 'transverse intentionality' (Querintentionalität), which is the sense that I have of the enduring or changing object. Transverse intentionality depends on longitudinal intentionality, that is, I can grasp the enduring or changing object only because I am retaining or anticipating my experience of it. For this reason phenomenologists claim that consciousness of something always involves a tacit or prereflective self-awareness, which is constituted by this retentional–protentional temporal structure.

A Hybrid Model

It may be possible to restructure this phenomenological model in a way that gives up the act–content distinction and is consistent with the overlap model. In this way the overlap model gains some important features of the phenomenological analysis – the distinction between longitudinal and transverse intentionalities, the idea of a prereflective self-awareness, and a clearer account of what constitutes the overlap.

The notion of overlap signifies that there is shared content from one specious present to the next – a different way to say this is that specious presents dynamically blend into each other. This blending is accomplished automatically by the implicit retentional–protentional aspects of the conscious flow itself. My experience of a melody is not just of one note and then another, but rather of notes that seem to arise and gradually drift off in a consistent, connected, and meaningful sequence as I am anticipating the notes to come. In the specious present the just-past content is retained as just past and the content that has not yet formed is anticipated as such. All of this happens at the first-order level of phenomenal experience. An additional act structure does not have to be built on top of the phenomenal content in order to execute intentional functions. The intentionality of time consciousness is built into or, more precisely, is the structure of phenomenal consciousness and is what accounts for both its coherence and its flow. The retentional–protentional structure that explains how specious presents blend into each other, and why my experience of things prereflectively includes self-awareness, is not a feature of conscious acts that themselves flow through conscious experience, but of the conscious flow itself. It is not a feature that belongs to consciousness because consciousness may involve an act of perception in one instance, or an act of memory, or imagination, or judgment, in another. It is rather an intrinsic part of experience itself.

This does not mean, however, that this temporal structure is in some way an absolute for consciousness, or that consciousness cannot lose this temporal structure, or that the temporality of consciousness is formally the same in every case. This phenomenological temporal sense is based on retentional and protentional processes that ultimately need to be cashed out in terms of neurological processes that may be affected by a variety of contingencies. This goes beyond phenomenological analysis in the direction of neuroscience, although the phenomenology of time-consciousness should help us to understand what such neuronal processes accomplish. Recent research (e.g., by Karmarkar and Buonomano) shows that on the scale of milleseconds, neuronal activation in sensory areas is more holistic than linear in such a way that recent stimuli remain encoded in the neural network. The short-term history of the system has an effect on the way the system deals with new information. For example, a 50-ms stimulus followed by a 100-ms stimulus is not registered as the sum of the two. Rather, the earlier stimulus has an effect on the processing of the 100-ms interval; the temporal information is shaped by the context of the entire pattern. This data is consistent with a dynamical systems explanation of the retentioning function at the level of neuronal activation. If the dynamics of neuronal processing are changed by drug ingestion, for example, or by changes in bodily metabolism, this may show up in the temporal structure of our experience.

These retentional and protentional aspects of experience, then, should not be thought of as characteristics of a free-floating consciousness. As aspects of consciousness they may be regarded as reiterations of processes that happen in the

brain, and more generally in the body as they are expressed in motor behavior as it is situated in the environment. They are ubiquitous in both conscious and nonconscious processes. The effect of retained information, for example, is clearly demonstrated in priming effects in perception and cognition. Anticipation characterizes both conscious and nonconscious motor control processes. The current kinematics of the graphic production involved in writing a certain letter are influenced nonconsciously by the anticipation of the next letter to be written and will differ depending on what the next letter is. Both aspects, also, as suggested above, play a role in the constitution of a minimal sense of self and may underlie one's sense of agency for action, which may in turn be disrupted if protentional or anticipatory aspects (predictive coding in the motor system, for instance) go awry in, for example, schizophrenic symptoms of delusions of control.

More generally, disruptions of the temporal structure of consciousness can be found in a variety of psychopathologies, including nonpsychotic affective disorders, where the basic temporal structure is maintained, but experience is felt to be either accelerated (as in mania) or decelerated (as in depression), in depersonalization, which is characterized as a temporal disintegration, as well as in schizophrenia where the temporal structure may itself be compromised, as manifested in cognitive dysmetria.

See also: Perception: Subliminal and Implicit; Perception: The Binding Problem and the Coherence of Perception; Self: Personal Identity; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Andersen H and Grush R (in press) A brief history of time consciousness: Historical precursors to James and Husserl. Journal of the History of Philosophy.

Broad CD (1923) Scientific Thought. Paterson, NJ: Littlefield, Adams, 1959.

Dainton B (2006) Stream of Consciousness: Unity and Continuity in Conscious Experience. London: Routledge.

Dennett DC and Kinsbourne M (1992) Time and the observer: The where and when of consciousness in the brain. Behavioral and Brain Sciences 15: 183–201.

Eagleman DM and Sejnowski TJ (2000) Awareness motion integration and postdiction in visual awareness. Science 2036: 287.

Gallagher S (1998) The Inordinance of Time. Evanston: Northwestern University Press.

Grush R (2006) How to, and how not to, bridge computational cognitive neuroscience and Husserlian phenomenology of time consciousness. Synthese. DOI 10.1007/s11229-006-9100-6.

Husserl E (1991) On the Phenomenology of the Consciousness of Internal Time (1893–1917), Brough JB (trans.). Dordrecht: Kluwer Academic Publishers.

James W (1890) Principles of Psychology. New York, NY: Dover Publication, Inc., 1950.

Karmarkar UK and Buonomano DV (2007) Timing in the absence of clocks: Encoding time in neural network states. Neuron 53: 427–438.

Kelly R [Clay ER] (1882) The Alternative: A Study in Psychology. London: Macmillan.

Mabbott JD (1951) Our direct experience of time. Mind 60: 153–167.

McTaggart J (1908) The unreality of time. Mind 17: 456–473.

Pöppel E (1988) Mindworks: Time and Conscious Experience. New York: Harcourt Brace Jovanovich.

Varela FJ (1999) The specious present: A neurophenomenology of time consciousness. In: Petitot J, Varela FJ, Pachoud B, and Roy J-M (eds.) Naturalizing Phenomenology. Issues in Contemporary Phenomenology and Cognitive Science, pp. 266–314. Stanford, CA: Stanford University Press.

Vogeley K and Kupke C (2007) Disturbances of time consciousness from a phenomenological and a neuroscientific perspective. Schizophrenia Bulletin 33(1): 157–165.

Biographical Sketch

Shaun Gallagher is a professor of philosophy and cognitive sciences and a senior researcher at the Institute of Simulation and Training at the University of Central Florida, and a research professor of philosophy and cognitive science at the University of Hertfordshire. He is the author of *How the Body Shapes the Mind* (2005), *The Phenomenological Mind* (2008, with Dan Zahavi), *Brainstorming* (2008), *The Inordinance of Time* (1998), and *Hermeneutics and Education* (1992). He has held visiting positions at the MRC Cognition and Brain Sciences Unit at Cambridge University, the University of Copenhagen, and the Ecole Normale Supériure, Lyon. He is a coeditor-in-chief of *Phenomenology and the Cognitive Sciences*.

The Control of Mnemonic Awareness

B J Levy and M C Anderson, University of Oregon, Eugene, OR, USA

Glossary

Inhibition – Either the subtractive process by which the activation level of a mental representation is actively reduced or the resulting state of reduced activation.
Inhibitory control – The engagement of a controllable mechanism to lower the activity of a mental representation.
Response override – The stopping of a strongly habitual, prepotent response by control mechanisms.
Retrieval stopping – A specific instance of response override, where one must prevent a memory from entering conscious awareness by overriding the retrieval process itself.
Selective retrieval – Another instance of response override in memory that occurs when a cue is related to a strong, prepotent memory that is not currently desired. Remembering some other weaker memory in response to that cue requires the rememberer to override retrieval of the prepotent memory.

Introduction

Every person possesses a fundamentally private conscious awareness that defines their sense of their own existence. This sense includes awareness of perceptions from the surrounding environment, internal bodily states, as well as thoughts, ideas, and memories that may enter consciousness at any moment. Although awareness is often steered by stimuli around us and by our relatively automatic responses to these stimuli, consciousness can also be controlled. We can voluntarily bring to mind some past experience unbidden by any reminder, we can willfully change the direction of our thoughts, or we can focus awareness on a single idea to the exclusion of all else. Indeed, goal-directed cognition relies on the capacity to control awareness. Because this sense of awareness is central to who we are and because our capacity to control awareness is strongly connected to being goal-directed agents, scientific theories of consciousness need to explain how such control is achieved. What permits us to think about some things and not others? How do we regulate the focus of awareness?

In this article, we focus on a theoretical hypothesis about how this type of control is achieved: the response override hypothesis. According to this theory, people control awareness of memories and ideas by engaging executive control mechanisms that were originally developed to control overt motor action. In particular, controlling memory may be a special case of a general situation requiring executive control, referred to as response override. In these situations, one must stop a habitual response to a stimulus due to situational demands – an ability that is crucial for voluntary control (see Figure 1). For instance, after knocking over an object one might reflexively reach out to catch the item and stop its fall. If the falling object is a cactus, however, this otherwise useful perceptuomotor reflex must be overridden to prevent this pain-inducing response. This type of control is widely thought to be accomplished by inhibitory processes that suppress the inappropriate response. According to the response override hypothesis, this same inhibitory mechanism operates within the domain of memory to override the retrieval process, providing the mechanism that allows us to control the current contents of conscious awareness.

Two basic memory situations requiring response override have been identified and studied: the need for selection during retrieval and the need to stop retrieval itself. Selection is required when our goal is to recall an event or fact from long-term memory in the face of interference or distraction from

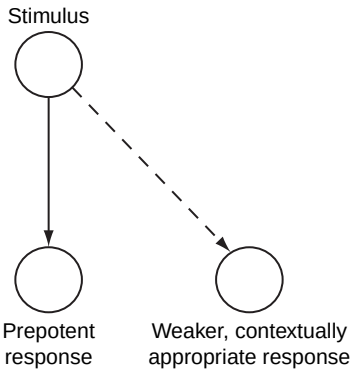


Figure 1 A typical response-override situation. In this figure, a stimulus is associated with two responses, one of which is stronger (prepotent) and the other weaker (indicated by the dotted line). Response override occurs whenever one needs to either select the weaker, but more contextually appropriate response, or to simply stop the prepotent response from occurring. Inhibitory control is thought to achieve response override by suppressing activation of the prepotent response. This basic situation describes many paradigms in research on executive control, including the Stroop and go/no-go tasks.

related traces. The need to stop retrieval arises when we confront a reminder and wish to prevent an associated memory from entering conscious awareness. Both of these processes are necessary for goal-directed cognition as they allow flexible control over whether retrieval is completed and, if so, which memory is retrieved. In both selection and stopping situations, attempts to limit awareness of activated and distracting memories impair memory for those traces later, highlighting an important connection between controlling consciousness and forgetting.

Lessons from Studying Perceptual Awareness as a Model System

The history of psychological research on consciousness has been primarily concerned with perceptual awareness. In these early studies, researchers attempted to isolate situations where a percept enters awareness in order to study the changes – both psychological and neurophysiological – that occur at the boundary between unconscious and conscious perception (see other sections of this volume). Even within perception, research on consciousness has focused primarily on visual perception. For example, studies of binocular rivalry have looked at how conscious awareness flips back and forth between two different stimuli that are independently presented to each eye. Pioneers in this area, such as Cristoph Koch, have argued that this focus has been necessary for initial progress because our understanding of the neurobiological systems underlying sensory processing, particularly vision, is quite advanced compared to those involved in higher-order cognition. Although this approach has been profitable in understanding how conscious perception arises, this research does not address much of what concerns our conscious experience. Our thoughts are not driven solely by external stimuli: they are also influenced by thoughts, ideas, and motivations. Additionally, work on perceptual awareness has focused more on the neural correlates of conscious states and less on how awareness can be controlled. One fundamental goal of theories of consciousness, however, should be to explain how control over consciousness is achieved. Building a more complete model of the internal conscious state requires understanding not only how perceptions become conscious but also how we regulate which memories and perceptions reach awareness and which do not.

Work on conscious perception provides several lessons that can be applied to the study of consciousness within memory. First, there are far more percepts than we can capture within focal attention at any moment. Our brains process much of this perceptual information to some degree, without many of these percepts ever reaching conscious awareness. Critically, attentional control processes are required to select a subset of these stimuli to be represented within consciousness. Similarly, there are far more memories and ideas represented internally than we can be currently aware of and many of these may be at least partially active at any given time. Thus, interactions with internal representations also must be governed by attentional mechanisms that select which memories enter awareness. Second, studies of conscious perception have relied heavily upon subjective reports. It is, in fact, critical for the study of consciousness as these are the most direct measure of consciousness we have. Establishing a mapping between neuronal activity and subjective

reports allows us to move toward physiological measures of consciousness that do not require subjective reports. Third, another critical tool utilized in the study of consciousness involves studying the same stimulus under different types of awareness, so that the constant features of the stimulus remain the same and only the phenomenal conscious awareness changes. These last two points have been influential in directing the research described here toward an explicit investigation of the conscious regulation of awareness in memory, as will be described later. Lastly, studies of conscious perception have emphasized the importance of identifying neural substrates involved in awareness. By identifying which brain regions underlie awareness, we move closer to understanding how consciousness arises. These insights and methods developed in research on perception are reflected in recent work on the control of mnemonic awareness, both in the context of selective retrieval and retrieval stopping.

Selective Retrieval

Our goals often require us to modify the current contents of awareness by redirecting attention to new information relevant to the current task. In the perceptual domain, this requires the selection of a particular aspect of our rich sensory input, to the exclusion of other inputs that may compete for the limited capacity for awareness. In the memory domain, we often need to bring to awareness some particular event or fact that is important for our immediate purposes. In the latter case, the key mental operation for achieving this alteration of the contents of awareness is memory retrieval, and, in particular, selective retrieval (the mnemonic equivalent of selective attention). During retrieval, we use cues relevant to our goals to guide our search for the desired content. Typically, though, these cues are associated with other representations in memory, in addition to the specific content we seek. In fact, often these related memories spring to mind more readily than the desired target. For example, trying to remember what you had for dinner last Tuesday might bring to mind other recent dinners or other evening plans. Similarly attempts to remember a new phone number after moving are often thwarted by the retrieval of the old, no longer relevant phone number from the prior residence. A long history of memory research suggests that when multiple memories are associated with the same cue they compete for access to conscious awareness during retrieval. This type of interference poses a significant problem for the effort to direct consciousness to the desired memory; it requires some form of control to override the nondesired memories. Recent research suggests that this competition for conscious awareness is resolved by inhibitory control processes that suppress distracting memories, similar to the involvement of inhibition in achieving perceptual selective attention. This weakening of related memories allows retrieval of the target, but at the cost of impairing future recall of the nonretrieved competitors. Studying the conditions under which this form of memory impairment occurs thus provides an important behavioral window into the use of inhibitory control mechanisms to manage the redirection of consciousness to new memorial content.

The role of inhibitory control processes in achieving selective memory retrieval has been studied by Michael Anderson and colleagues using a procedure known as the retrieval practice paradigm (see Figure 2). In a typical study, subjects study

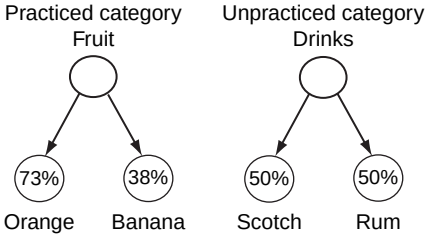


Figure 2 A standard categorical RIF study. Illustrated here are two items from each of two categories that subjects have studied (typically six items are studied from eight categories). In this example, subjects perform retrieval practice on 'Fruit-Orange,' but not on 'Fruit-Banana' (unpracticed competitor) or on any members from the 'Drinks' category (an unpracticed baseline category). The numbers show the percentage of items correctly recalled on the final cued-recall test. As shown here, retrieval practice facilitates recall of the practiced items relative to performance in baseline categories. RIF is reflected in the reduced recall of unpracticed members of the practiced category (Banana), relative to performance in baseline categories (Scotch and Rum).

lists of category-exemplar word pairs (e.g., Fruits-Orange, Fruits-Banana, Drinks-Scotch, Drinks-Vodka) and then practice retrieving some of these items from memory. Specifically, subjects practice half of the studied items from half of the studied categories (just 'Fruits-Orange' from the list above). The categories that are not practiced at all act as a control condition, so they can be used as an estimate of what baseline recall should be in the event that no items are practiced. After a delay, subjects are provided with all of the category cues again and they are asked to remember all of the items they studied earlier. As would be expected, the items that were repeatedly retrieved are recalled better on this final test than are the baseline items; thus, bringing past experiences into consciousness improves the ability to do this again, should the need arise. More interestingly, the unpracticed items from practiced categories (Fruits-Banana) are recalled more poorly than are baseline items. This intriguing finding suggests that when bringing a particular memory into consciousness, other memories that compete for the focus of awareness are inhibited. This finding, that retrieving specific targets induces forgetting of competing memories, has been termed retrieval-induced forgetting (RIF). RIF appears to be an instance in which the need to override automatic retrieval is solved by inhibitory control mechanisms that suppress the distracting content, supporting the effective redirection of consciousness.

Generality of Retrieval-Induced Forgetting

If RIF is a general consequence of attempts to control the redirection of consciousness to new memorial content, then it should be evident in any situation where we attempt to control retrieval. Indeed, RIF is not limited to category-exemplar pairs; rather, it appears to be a general phenomenon of importance to many everyday situations. Studies of learning, as would occur both in and outside the classroom, have shown that retrieving some facts about a topic impairs recall of other facts. For example, after reading textbook descriptions about two topics, being quizzed about some details of one topic causes forgetting of other details concerning that subject, but has no effect on the other studied topic. Similarly, recalling that 7 times 6 equals 42, makes it harder to remember that 7 times 9 equals 63. Beyond forgetting of simple facts, RIF also plays a role during the learning of a foreign language: when novice Spanish-speakers name pictures in Spanish, they subsequently experience difficulty generating the corresponding words in their native language.

Beyond these demonstrations of RIF in learning, this phenomenon is also of importance for other naturalistic situations. Studies of eyewitness memory show that interrogating subjects about specific details of a mock crime, in a manner consistent with actual police interviews, impairs memory for other noninterviewed details concerning the same crime. This suggests that the often numerous interviews performed by police and lawyers may have a profound influence on eyewitness memory for events. Recent work by Malcolm MacLeod and colleagues has even begun to explore whether RIF is responsible, at least partially, for the misinformation effect, originally pioneered by Elizabeth Loftus. Those classic studies showed that misleading information presented after an event, during an interview for example, is often mistakenly remembered as part of the original event. This new line of research suggests that the interview itself is actually critical for the misinformation effect, as the details that are weakened by RIF are the ones that are most vulnerable to subsequent misinformation.

The implications of RIF for social psychological phenomena have also been explored, with research showing that recalling some traits of a person makes it harder to remember that person's other personality traits. This phenomenon has even been extended to help understand how retrieving stereotypical information can lead one to forget stereotype-inconsistent, individuating features of an individual. Conversely, retrieving individuating features can cause forgetting of the stereotypical features. Recent studies also implicate RIF within autobiographical memory, suggesting that selective retrieval may play a role in shaping our own autobiographical history. The breadth of these findings indicate that selective retrieval is indeed involved in everyday cognition and has a profound impact not only on what is consciously remembered but also on what is excluded from awareness, both in the immediate term and in the long-term.

Evidence for Inhibition as the Mechanism that Produces RIF

To understand how consciousness is regulated during retrieval, it is critical to understand the precise mechanism by which selective retrieval is accomplished. Research suggests that RIF is produced by an inhibitory process that targets the competing memory trace itself. In contrast to this view, other researchers have proposed that noninhibitory mechanisms can produce the basic finding of RIF. For example, one noninhibitory account of RIF claims that during the final test phase, the items that were repeatedly practiced (Orange) are so dominant that they leap to awareness and block retrieval of the unpracticed competitors (Banana). By this account, the activation state of the competitor (Banana) has not changed. Instead, the rememberer perseverates on the practiced alternatives that have been strengthened, and this blocks retrieval of the competitor. Other noninhibitory accounts posit that the meaning of the retrieval cue is changed when it is used to practice a subset of its associates (one now thinks of citrus fruits when presented with 'Fruit' as a cue), rendering the cue useless as a means of retrieving noncitrus Fruits that had been previously studied. Importantly, these noninhibitory accounts all attribute the forgetting to some level (e.g., the cue, or the cue-target association) other than the forgotten item itself. By contrast, the inhibitory account makes the unique claim that the memory itself is being suppressed.

Several lines of evidence suggest, however, that selective retrieval inhibits competing items. First, RIF has been shown to be 'cue-independent,' as forgetting is observed even if the item is tested with a novel cue (e.g., Monkey-B_____ for Banana). According to the foregoing noninhibitory explanations, forgetting of competitors should only occur when the originally studied cue is used during the test. This should follow because the source of forgetting, according to those mechanisms, is specific to the original cue: the practiced response becomes so hyperaccessible given that cue, the meaning of the cue changes, or the link between the two items is unlearned. None of these explanations adequately explain why the competitor would be forgotten given an entirely novel cue. The inhibitory explanation, on the other hand, explicitly predicts that the item will be less accessible regardless of how it is tested. Building on this idea, in addition to being harder to recall, the competitors are also harder to recognize. Thus, it appears that the competitors have been reduced in activity.

Another property of RIF is that the forgetting suffered by competitors is not related to the strengthening of the target – a property known as 'strength-independence.' This means that strengthening target items, by itself does not cause forgetting of competitors. For example, if people are merely shown the category-exemplar pairs multiple times without having to retrieve them, similar strengthening is observed for the practiced items, but competitors are unimpaired. According to noninhibitory accounts, such as blocking, any form of strengthening should cause forgetting since strengthening of the practiced items is what leads them to block retrieval of the competitors. This finding demonstrates that RIF is 'retrieval-specific,' a property that is difficult to explain by most noninhibitory accounts. Lastly, the competitors that produce interference during the retrieval practice phase are inhibited more than ones that provide little interference. Thus, forgetting is 'interference-dependent,' suggesting that inhibition is engaged in response to interference from competing items. Again, this finding is difficult to explain by blocking, since the practiced items should block strong and weak competitors alike. Each of these properties strongly supports the claim that RIF is produced by inhibition, suggesting that inhibition plays a critical role in the way that conscious awareness is redirected to new traces in memory.

Neurobiological Basis of Selective Retrieval

Research has begun to explore the neurobiological underpinnings of inhibitory control during selective retrieval. As described earlier, inhibitory control is engaged during selective retrieval to prevent competitors from interfering with retrieval of the desired target. If successful, each successive retrieval practice should render competitors less interfering. Consistent with this, studies of RIF

using functional magnetic resonance imaging (fMRI) have revealed that performing retrieval practice engages both the ventrolateral prefrontal cortex (VLPFC) and the anterior cingulate cortex (ACC). The engagement of lateral PFC is consistent with prior research showing that the resolution of interference during selective retrieval from semantic memory also involves VLPFC. There is also neuropsychological evidence suggesting that patients with damage to lateral PFC experience difficulty resolving proactive interference. Similarly, the involvement of ACC is consistent with a broad range of findings that implicate that region in the detection of conflict. In the retrieval practice paradigm it appears that the competing memories trigger the need for top-down control, via the ACC, in order to resolve competition, which is then implemented by the engagement of inhibitory mechanisms mediated by the lateral prefrontal cortex. Supporting this idea, activity in these frontal regions declines across retrieval practice trials as the weakened competitors require less inhibitory control to be overridden. Critically, people who show greater decline in activity within these regions over trials show more memory inhibition. In addition, the people who show the highest degree of ACC activity during the initial trial are the ones most successful at suppressing. This suggests that subjects who experience the most competition initially are the ones who show the largest decline in lateral PFC activity across trials and the most forgetting.

Studies using EEG have suggested that selective retrieval is associated with a specific event-related potential (ERP) that indexes inhibitory control. In these studies, selective retrieval is contrasted with re-presentation of the studied stimuli, a condition that is known to not produce inhibition (as described earlier). Comparing activity in these two conditions yields an enhanced positive component in the selective retrieval condition over frontal electrode sites. Importantly, this enhanced activity is not due to strengthening of the practiced items, as these two conditions yield comparable facilitation; rather, this retrieval-specific component seems to index the inhibitory process that resolves interference. In fact, the magnitude of this component predicts how much forgetting the subject will experience. This finding suggests again that RIF is not produced by strengthening of practiced items, as is predicted by noninhibitory explanations. Rather, a specific inhibitory component is engaged to suppress the competitors rendering them less interfering. While localization of the source of ERP components is notoriously difficult, the frontal effect observed in these studies corresponds well with the fMRI findings on the importance of lateral PFC during selective retrieval. Thus, lateral PFC seems to subserve the selective filtering that controls which memories enter awareness, consistent with the view that response override mechanisms are central in the regulation of awareness.

Stopping retrieval

Whereas selective retrieval instigates the need to regulate which memories enter awareness, people generally do not form an explicit intention to down-regulate awareness in such situations. In other words, when attempting to remember some event or fact, remembering the target is the primary goal, and regulating interference occurs in support of this goal. However, sometimes stopping retrieval can itself be the person's primary goal. In these instances, we simply wish to stop retrieval from occurring. For example, when glimpsing an image of a loved one who has recently passed away we may marshal our efforts to stop painful thoughts of loss from coming into awareness. During the course of a typical workday we must frequently prevent distracting memories from involuntarily entering awareness and disrupting our current focus. These intrusive memories can be emotional in nature or simply consist of other activities or duties not related to our current goals. In extreme situations, survivors of abuse or combat veterans must exert this type of control in order to prevent traumatic memories from overwhelming their lives. In these instances, there is a clear conscious intention to prevent a memory from entering awareness. Clearly, such motivated retrieval stopping is a critical ability for daily mental functioning and for understanding how the content of consciousness is controlled.

This situation, overriding retrieval of an unwanted memory, has been studied using the Think/No-Think (TNT) paradigm. In these studies, subjects learn pairs of words (e.g., hug-rose, steam-train, broom-house) and are then asked to exert executive control over these memories. On some trials, in what is known as the 'Think' condition, people are asked to try to bring the target word to mind (when you see 'hug' think of 'rose'). For other 'No Think' trials, people are instructed to attend to the cue, but to willfully prevent the unwanted memory from entering consciousness (when you see 'steam' prevent the associated word from entering awareness). An additional set of cue words (e.g., broom) are not shown during this phase in order to provide a baseline measure of how accessible these pairs would be if they were neither retrieved nor suppressed after their initial learning. In the final phase a surprise memory test is given for all of the studied word pairs. Studying the memorial consequences of either thinking of a memory or excluding it from consciousness gives us an objective behavioral window into the mechanisms by which awareness is regulated.

People are, of course, better able to remember the words that they thought about compared to the baseline words, again affirming the idea that bringing memories into awareness improves one's ability to do so again later on. Evidence for inhibitory control arises from the finding that attempting to suppress awareness of response words during 'No Think' trials renders them harder to recall than the baseline items (see Figure 3). This below-baseline recall is present even when subjects are paid for correct answers or when they are misled into believing that the avoided words should be the easiest to remember. Thus, failure to recall does not reflect biases on the part of the person toward not reporting otherwise recallable 'No Think' items. Crucially, the impairment is not observed when the instructions are simply changed so that person only needs to withhold the vocal response, rather than avoid thinking about the memory. This indicates that the attempt to regulate conscious awareness is a critical and necessary component to produce this type of forgetting. Thus, regulating awareness is accomplished through inhibitory control of unwanted memories.

As was the case with RIF, forgetting in the TNT paradigm could be produced through noninhibitory mechanisms. For example, when presented with NT cues, subjects may simply generate alternative associations to distract themselves from

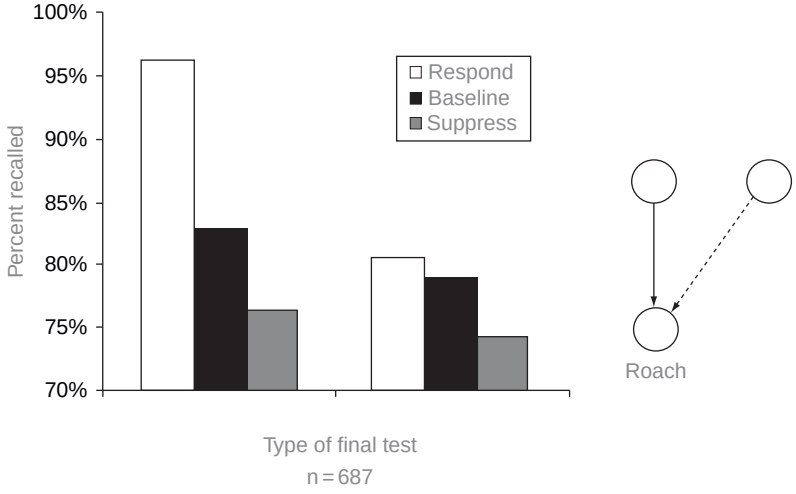


Figure 3 Final recall performance in the TNT procedure. The graph shows the percentage of items that subjects correctly recalled on the final test as a function of whether they tried to recall the item (Think), suppressed the item (No Think), or had no reminders to the item during the TNT phase (Baseline). The left side shows recall when tested with the originally trained retrieval cue (i.e., the Same Probe), whereas the right side shows recall when tested with a novel, extralist category cue (i.e., the Independent Probe). Reproduced from Anderson MC, and Levy BJ (2006) Encouraging the nascent cognitive neuroscience of repression. Behavioral and Brain Sciences 29: 511–513, with permission from Cambridge University Press.

thinking of the learned response. If true, this would mean that the unwanted memory was not so much intentionally pushed out of consciousness, as it was merely replaced by an alternate memory. If people accomplish the task by this form of thought substitution, then these alternative thoughts would become strengthened, potentially blocking retrieval of the response word during the final test phase. In order to rule out this blocking explanation, memory for the word pairs can also be tested with new categorical cues that subjects have not seen earlier in the experiment. Doing this yields similar forgetting, suggesting again that forgetting is 'cue-independent.' Other studies have also established that avoiding these memories also makes them harder to recognize, further confirming that these avoided memories have been inhibited.

Neurobiological Basis of Stopping Retrieval

Neuroimaging studies have found that attempts to stop retrieval are associated with increased activity within the lateral PFC, including both dorsolateral and ventrolateral regions (see Figure 4). Supporting the idea that lateral PFC is critical for suppression, individual differences in the magnitude of DLPFC activation are positively correlated with the amount of forgetting observed (see Figure 5). In addition to lateral PFC, suppression attempts activate a frontoparietal network of regions, including ACC, intraparietal sulcus, and the lateral premotor cortex, that is often observed in studies where subjects must prevent unwanted motor actions. The strong overlap between the network activated by retrieval stopping and motor stopping supports the claim that overriding is accomplished generally by a common system, regardless of whether the output being suppressed is motor or memorial in nature.

In addition to regions that are considered to be the source of the inhibitory signal, such as DLPFC, interest has also been taken in identifying the sites of inhibition: regions that are modulated by control. This goal dovetails with the approach described earlier within conscious perception, where an emphasis is placed on identifying candidate regions necessary for conscious awareness and on studying how these are influenced by tasks

Figure 4 fMRI results from Anderson et al. Plotted above are the brain regions that significantly differed in activation between the Suppression trials and Respond trials during the TNT phase (n = 24). Areas in yellow were more active during Suppression trials than during Respond trials, whereas areas in blue were less active during Suppression ($p < 0.001$). The white arrows highlight the reduced hippocampal activation in the Suppression condition. From Anderson MC, Ochsner K, Kuhl B, Cooper J, Robertson E, Gabrieli SW, Glover G, and Gabrieli JDE (2004) Neural systems underlying the suppression of unwanted memories. Science 303: 232–235. Reprinted with permission from AAAS.

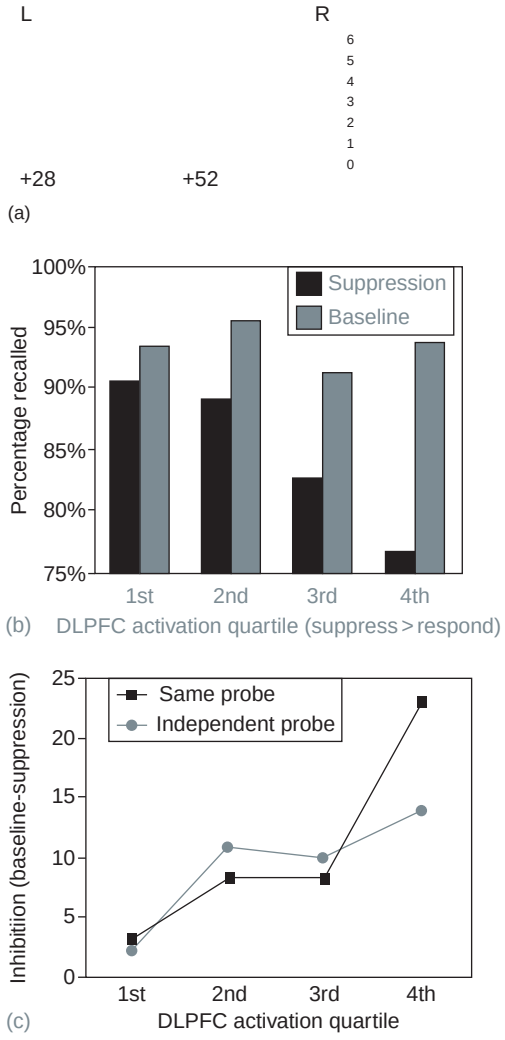


Figure 5 Successful recruitment of DLPFC predicts behavioral inhibition. (a) Shown here are the regions that correlate with the magnitude of the suppression effect observed on the final memory tests (the white arrows indicate the DLPFC). (b) Memory inhibition effects for four subject groups, differing in DLPFC activation. Subjects with greater DLPFC activity (on the right side) show reduced recall of No Think items, but do not differ from other subjects on their recall of Baseline items. (c) Magnitude of the suppression effect on both the same probe and independent probe tests for each DLPFC group. From Anderson MC, Ochsner K, Kuhl B, Cooper J, Robertson E, Gabrieli SW, Glover G, and Gabrieli JDE (2004) Neural systems underlying the suppression of unwanted memories. Science 303: 232–235. Reprinted with permission from AAAS.

that manipulate conscious awareness. Here, in contrast to motor stopping situations where the targets of inhibition would be motor systems, the likely candidate region would be one known to be involved in memory retrieval. It is well established that the medial temporal lobe, particularly the hippocampus, is involved in both the encoding of new experiences and the conscious retrieval of these memories later, especially for recently acquired memories. Moreover, it seems to play an especially critical role in memory for the type of richly detailed episodic memories that form the foundation of our conscious mental life. Interestingly, when subjects attempt to override retrieval, the hippocampus is down-regulated below the activity observed during trials where subjects perform retrieval. More impressively, it is also down-regulated below the level of activity observed during baseline periods where the subject has no task to perform other than to stare passively at a fixation cross. Thus, attempts to regulate conscious awareness by suppressing unwanted memories result in reduced hippocampal activation, produced by inhibitory control processes mediated by the frontoparietal regions described above. These findings support the view that response override systems mediated by the lateral prefrontal cortex can be targeted at structures involved in memory to prevent retrieval, consistent with the response override hypothesis.

Electrophysiological studies have also arrived at similar conclusions. Attempting to prevent an unwanted memory from entering awareness is associated with an early ERP component – the No-Think N2 – that arises over frontal sites and resembles, in topography and timing, the motor N2 typically observed in motor response suppression tasks such as the stop-signal and go/no-go paradigms. Importantly, when subjects are specifically instructed to suppress the unwanted memory directly whenever it comes to mind – and not simply generate distracting thoughts – the magnitude of this component predicts later memory inhibition effects, as would be expected based on the response override hypothesis of memory regulation. In addition to this early component, electrophysiological studies have also observed a late left parietal component that was specific to learned Think items. The timing and topography of this component are consistent with the widely studied parietal episodic memory (EM) effect, which has been linked to the subjective experience of consciously recollecting a past event. Crucially, this component is entirely absent during the No Think trials. Indeed, the EM

component was reduced in magnitude to the level observed for word pairs that were never learned. Thus, suppression completely eliminated this late retrieval-related component, suggesting that executive control can stop conscious recollection very effectively. Thus, ERP evidence is consistent with the model generated from fMRI studies suggesting that inhibitory processes mediated by lateral PFC accomplish control by actively downregulating the hippocampus and, as a consequence, conscious recollection.

Stopping Retrieval as a Laboratory Model of Freudian Suppression

The TNT paradigm provides a useful model for studying the psychological control process that Sigmund Freud referred to as suppression. According to Freud, suppression involved willfully preventing an idea from entering conscious awareness. This is precisely what subjects are asked to do in the TNT task, which suggests that results from those studies may be relevant to clinical issues related to suppression. Importantly, this new research provides empirical evidence about an issue that has largely been treated as scientifically intractable. While suppression has been difficult to study empirically, these new studies indicate that engaging suppression has a clear influence on subsequent access to these avoided memories.

Caution is necessary, however, in applying these laboratory studies to issues of clinical importance. The fact that a subject can forget a neutral word is not evidence that people can inhibit rich, episodic memories of traumatic experiences, as claimed by Freud and others. Recent studies have made inroads in establishing the viability of this paradigm for studying more naturalistic forgetting. In particular, evidence suggests that more complex memories of emotionally arousing events can also be inhibited. The original demonstrations of forgetting have now been replicated with emotionally negative memories even when the stimuli were naturalistic photographs (e.g., the scene of a major car accident), stimuli that are likely to be more vivid and, one might expect, harder to forget. In these studies, forgetting is typically as robust, if not more so, for negative stimuli as compared to neutral stimuli. Thus, even complex, vivid, emotional memories can be inhibited. This suggests that this psychological process has relevance for real-life memory suppression, which likely involves negatively charged memories. More research will be needed to further establish the ecological validity of this process. It is important to note, however, that the studies discussed here do not speak to the existence of unconscious repression, where unwanted thoughts are automatically pushed out of awareness. Indeed, the processes studied in this work are deliberate and intentional.

Individual Differences in Stopping Retrieval

There is considerable variability in how effective individuals are at recruiting inhibition to control which memories enter consciousness. Some people show dramatic forgetting of the avoided memories, while others actually seem to remember these items better despite their attempts to keep them out of mind. Much of this variability appears to be due to individual differences in executive control abilities. As described earlier, overriding retrieval of an unwanted memory is a specific example of a response override task, which is widely regarded as involving executive control. Therefore, it seems reasonable that variations in this ability would predict how successfully people inhibit unwanted memories. In support of this conclusion, neuroimaging studies have shown that the subjects who most strongly recruit dorsolateral PFC, a region thought to be critical for response override ability, show the most inhibition on the final test (see Figure 5). Similarly, complex working memory span tasks can be used as a proxy for executive control abilities, as these tasks require considerable executive control. Subjects load items into working memory and then respond to intervening items, which interfere with the maintained information and disrupt rehearsal. People with a high working memory span show large inhibition effects in the TNT procedure, whereas people with low working memory spans show facilitation of these avoided memories. Additionally, memory control ability differs across various populations known to vary in executive control. For example, older adults, who suffer a disproportionate loss of executive control abilities relative to other cognitive functions, show

difficulty inhibiting in the TNT paradigm. On the basis of this evidence, it seems that differences in memory suppression are at least partially attributable to differences in executive control more broadly. If so, these findings provide further support for a linkage between general response override mechanisms and the regulation of memorial awareness, and further point to clear individual differences in how effectively people regulate awareness.

Identifying conscious awareness of intrusive memories

The earlier sections describe a framework for understanding how inhibitory processes are engaged in order to maintain control over which memories enter awareness. Recent work has examined more explicitly how this type of control is related to consciousness. As described earlier, research on conscious perception has emphasized the use of subjective reports to distinguish between different conscious experiences of a stimulus. The same general approach can be taken within memory by having subjects report when a memory enters awareness. Recent research using the TNT paradigm has begun to employ subjective reports in order to address how conscious awareness of unwanted memories relates to the observed memory impairment. As in typical TNT studies, subjects learn word pairs and then later practice either thinking of the associated response or preventing it from coming to mind. The difference in these studies is that after each trial during the TNT phase, subjects make a subjective rating about whether or not they thought of the response word during the previous trial. Specifically, subjects report whether they 'never' thought of the response, thought of it 'briefly,' or thought about it 'often' during the time the cue word was on the screen. This procedure provides a means of distinguishing between trials where the subject is successful at suppressing awareness of the unwanted memory (i.e., they report never thinking about it) and trials where the unwanted memory 'intrudes' into conscious awareness, even if this awareness is fleeting.

This type of binary distinction may at first seem too gross, but, in essence, that is exactly what the study of consciousness requires. While the activation level of a given memory can be conceived of as a continuous measure, the distinction made in research on consciousness is between representations that have either passed threshold and entered awareness or that have not. It is this transition from a below-threshold to an above-threshold representation that is critical for identifying neural correlates of consciousness. Prior studies of consciousness in memory have generally lacked a proper way to identify when a memory transitions from being unconscious to entering conscious awareness and it has been difficult to induce a situation where a subject can monitor for their occurrence and be likely to actually experience a specific memory entering awareness.

As described earlier, forgetting in the TNT paradigm appears dependent on subjects attempting to exclude a memory from conscious awareness. Using subjective reports, however, provides a direct measure of the regulation of awareness. Using this approach has provided further evidence that attempts to regulate awareness produce behavioral inhibition on the final test. In these studies people report frequent intrusions initially as they struggle to successfully inhibit the unwanted memories. However, with practice, subjects down-regulate the frequency of intrusions, so that with practice intrusions become quite rare. Critically, people who are best able to down-regulate the frequency of intrusions are the same ones that show memory impairment on the final test. Thus the ability to regulate awareness and overcome intrusions with practice predicts the likelihood of forgetting those items on the final test, directly linking inhibitory control with the regulation of conscious awareness.

Neuroimaging evidence also supports a strong coupling between the down-regulation of intrusions and successful inhibition. In particular, overlapping regions within DLPFC predict both measures of inhibitory control ability – the down-regulation of conscious intrusions and inhibition as measured on the final memory test. Thus, these two measures show similar variability across individuals and, furthermore, this variability is produced by common neural substrates. Thus we have strong evidence suggesting that attempts to prevent a memory from entering awareness inhibit the avoided memory, making it less intrusive on subsequent trials and less memorable later even after suppression attempts have ceased.

Fascinatingly, when a person experiences a conscious intrusion of the unwanted memory, increased activation is observed in the lateral parietal cortex – a region that has previously been implicated in reflexive orienting to abrupt onsets in the perceptual world. In these studies, there is no abrupt perceptual onset, but the intrusion of the unwanted memory can be considered to be an abrupt internal onset that draws attention toward this newly activated memory. Thus, internally oriented attention can involuntarily focus attention on an abrupt onset in memory just as startling events in the perceptual environment can capture attention. Critically, this same region is not activated during Think trials, so it does not reflect retrieval itself or the representation of the memory within awareness. Rather, it reflects involuntary, reflexive retrieval that occurs when a memory pops into mind without an intention to retrieve it. This suggests that a common brain region may be engaged when a sudden perceptual event captures attention (and thus shifts awareness) as when a sudden mnemonic event diverts the focus of mnemonic awareness. This underscores potentially important commonalities in systems that challenge the effective control of perceptual and mnemonic awareness.

Controlling access to working memory

Another excellent example of common principles governing the regulation of perceptual and mnemonic awareness comes from research on the role of inhibitory control in regulating which aspects of perception gain access to working memory. In a task devised by Adam Gazzaley, subjects view pictures of faces and scenes and are asked to either attend to the faces, attend to the scenes, or to passively view the stimuli without attending specifically to either stimulus type. During each trial, subjects view a series of faces and scenes and then after a brief retention interval are asked to judge whether a test image was in the previous set. Neuroimaging studies using this task have focused on activity within the cortical regions that are specialized for processing each specific type of stimuli: the fusiform gyrus for face stimuli and the parahippocampal gyrus for scene stimuli. These studies find that attending to a stimulus class enhances activity within that region (i.e., attending to faces engages the fusiform gyrus) relative to when the same stimulus is passively viewed. In contrast, when subjects are asked to ignore a stimulus class, activity within the region devoted to processing that stimulus type is less active than when the stimuli are passively viewed. This suggests that ignoring a specific category of perceptual stimuli reduces activity within the brain region known to represent that stimulus. This clearly supports the idea that inhibitory control can be engaged in a top-down fashion to modulate representations. In this paradigm, older adults have difficulty with inhibiting processing of the currently irrelevant stimulus, while showing no impairment at content-selective enhancement. Like the TNT task, people in this task also regulate conscious awareness by selective attention mechanisms. This control appears to be implemented by the downward modulation of activation in the cortical region involved in processing the ignored stimulus, similar to the hippocampal modulation observed during suppression of episodic memories. Thus, inhibitory control may gate access to consciousness for both external, perceptual information and internal, memorial information.

Studies investigating thought suppression

A parallel body of research examines the ability to suppress unwanted thoughts, as opposed to episodic experiences. In these studies, pioneered by Daniel Wegner, people are asked to keep a specific thought (e.g., a white bear) out of mind for several minutes. During this delay, people are typically asked to continuously speak their thoughts aloud, with no specific instructions about what they should think about. They are further instructed that during this time they should monitor awareness for the presence of the unwanted thought and indicate whenever it comes to mind (e.g., by pressing a button or ringing a bell). Afterward, people engage in another think-aloud period where they are free to think any thoughts they wish. The typical finding from this paradigm is that the people who are asked to avoid a specific thought during the prior phase are more likely to think of the avoided thought. Furthermore, people often have difficulty keeping the thought out of mind during the suppression phase. From these results,

Wegner and colleagues concluded that attempting to avoid an unwanted thought results in it being 'ironically' more accessible later and, therefore, thought suppression is ultimately a futile endeavor. This paradigm has proven to be profitable in clinical research, particularly in relation to obsessive– compulsive disorder, depression, and anxiety.

These findings, however, appear to be at odds with the conclusions drawn from studies using the TNT procedure, namely, that people can successfully suppress unwanted memories. While these seem to be contradictory findings, it is also possible that they represent two different situations where people attempt to control awareness, but with differing results. While people are quite unsuccessful at suppressing in the White Bear studies, this does not mean that people lack the ability to regulate conscious awareness. As the TNT paradigm demonstrates, they can be effective at engaging inhibitory control to prevent an unwanted memory from entering consciousness under the right circumstances. Thus, both paradigms appear to capture situations where thought suppression is employed in naturalistic settings. Further research is needed to specify what factors determine success at regulating awareness.

Conclusion

Much of our conscious experience is driven by perceptual stimuli in the environment and by relatively automatic retrieval processes that respond to those environmental cues. Such automatic retrieval often enables us to retrieve appropriate behavior and interact with the world in an effortless manner. Many situations, however, require us to exert control over memory in order to behave flexibly. Moreover, many hallmarks of conscious awareness, such as our ability to adapt, to change our minds, and to think creatively require us to control retrieval and dictate which memories enter awareness.

In this article, we reviewed evidence for the idea that people exercise control of the contents of mnemonic awareness by engaging executive control processes that have developed in service of behavioral regulation. In particular, we suggest that controlling mnemonic awareness, at its most basic level, involves controlling the retrieval process, which may profitably be viewed as a special case of the broader problem of response override. The parallels between motor response override and mnemonic override may be observed at both the functional level and neurobiological levels: at the functional level, two main functions that engage inhibitory control over motor actions – selection and stopping – also engage inhibitory control in memory retrieval; at the neurobiological level, a common region for response override can be observed in the lateral prefrontal cortical regions that subserve motor inhibition and memory inhibition. A key difference, however, concerns the neural regions targeted by inhibitory control; for motor inhibition, motor cortical structures are affected, whereas for memory inhibition, brain systems involved in conscious recollection of the past (the hippocampus) or conscious perception of the present (the fusiform face area or the parahippocampal place area) are downregulated. Thus, evidence from the behavioral and the neural level point to the existence of mechanisms that actively diminish processing of that which we wish to exclude from awareness – mechanisms whose behavioral footprints may be observed in the later forgetting of those memories. If correct, this view suggests the intriguing principle that controlling the type of content we allow to enter awareness is a matter of controlling internal cognitive actions (like retrieval) that generate that content – an ability that is grounded in fundamental processes that have evolved in service of controlling what we do (and do not do) in the world around us.

See also: Consciousness and Memory in Amnesia; Memory: Errors, Constructive Processes, and Conscious Retrieval.

Suggested Readings

Anderson MC (2003) Rethinking interference theory: Executive control and the mechanisms of forgetting. Journal of Memory and Language 49: 415–445.

Anderson MC and Green C (2001) Suppressing unwanted memories by executive control. Nature 410: 366–369.

Anderson MC and Levy BJ (2006) Encouraging the nascent cognitive neuroscience of repression. Behavioral and Brain Sciences 29: 511–513.

Anderson MC, Ochsner K, Kuhl B, et al. (2004) Neural systems underlying the suppression of unwanted memories. Science 303: 232–235.

Anderson MC and Spellman BA (1995) On the status of inhibitory mechanisms in cognition: Memory retrieval as a model case. Psychological Review 102: 68–100.

Anderson MC and Weaver C (2008) Inhibitory control over action and memory. In: Squire L (ed.) The Encyclopedia of Neuroscience. http://www.elsevierdirect.com/brochures/ens/index.html. Elsevier.

Bergström ZM, Velmans M, De-Fockert J, and Richardson-Klavehn A (2007) ERP evidence for successful voluntary avoidance of conscious recollection. Brain Research 1151: 119–133.

Depue BE, Banich MT, and Curran T (2006) Suppression of emotional and non-emotional content in memory: Effects of repetition on cognitive control. Psychological Science 17: 441–447.

Depue BE, Curran T, and Banich MT (2007) Prefrontal regions orchestrate suppression of emotional memories via a two-phase process. Science 37: 215–219.

Gazzaley A, Cooney JW, McEvoy K, Knight RT, and D'Esposito M (2005) Top-down enhancement and suppression of the magnitude and speed of neural activity. Journal of Cognitive Neuroscience 17: 507–517.

Kuhl BA, Dudukovic NM, Kahn I, and Wagner AD (2007) Decreased demands on cognitive control following memory suppression reveal benefits of forgetting. Nature Neuroscience 10: 908–914.

Levy BJ and Anderson MC (2002) Inhibitory processes and the control of memory retrieval. Trends in Cognitive Sciences 6: 299–305.

Levy BJ and Anderson MC (2008) Individual differences in the suppression of unwanted memories: The executive deficit hypothesis. Acta Psychologica 127: 623–635.

Wenzlaff RM and Wegner DM (2000) Thought suppression. Annual Review of Psychology 51: 59.

Biographical Sketch

Dr. Benjamin Levy is a postdoctoral fellow at Stanford University. He received his PhD from the University of Oregon and his BS from the University of California, Davis.

Dr. Michael Anderson is a professor of cognitive neuroscience at the School of Psychology at the University of St Andrews, Scotland, where he is the director of the Memory Control Laboratory. He has previously served as a professor of cognitive neuroscience at the University of Oregon, has been a visiting scholar at Stanford University and the University of California, Berkeley, and has served on the editorial boards of the Journal of Experimental Psychology: Learning, Memory, and Cognition, Memory and Cognition, and Perspectives in Psychological Science. He received his PhD in cognitive psychology and neuroscience from the University of California, Los Angeles.

Development: Consciousness From Birth to Adulthood

P D Zelazo, University of Minnesota, Minneapolis, MN, USA

Glossary

Intentionality – The property of 'aboutness' or directedness that can be said to characterize the relation between conscious thoughts and their content.
Levels of consciousness – Dissociable varieties of conscious experience generated by different degrees of the reflective reprocessing of information.
Minimal consciousness – The first-order sensory awareness attributed to newborn infants and to adults engaging in implicit information processing.
Prefrontal cortex – The anterior third of the brain.
Psychological distance – Cognitive separation from the immediate perceptual and behavioral environment through the use of reflection and symbolic representation.
Recursive consciousness – Consciousness resulting from one degree of reflection on the contents of minimal consciousness. The term recursive is used here in the sense of a computer program that calls itself.
Reflection – Awareness and conscious consideration of one's own sensations, perceptions, thoughts, and behavioral tendencies. Reflection may be brought about by the reprocessing of information through neural circuits involving the prefrontal cortex.
Theory of mind – Thinking about one's own or others' mental states.

Introduction

The contents of human consciousness obviously change with age, and our interactions with the world clearly influence our ideas about self, subjectivity, volition, and other minds, among many related constructs. However, it is more difficult to determine whether the quality of phenomenological experience is similarly transformed, and whether the cognitive and behavioral functions of consciousness change as children develop. Research on this topic remains controversial, but there is a growing body of evidence consistent with the suggestion that the structure and functions of conscious experience do in fact develop, as children come to reflect on themselves and their subjective experiences. This development may be tied to the growth of neural systems involving the prefrontal cortex, which is critically important for self-awareness and the conscious control of thought and action. The prefrontal cortex continues to mature into adulthood, raising the possibility that the development of consciousness is similarly protracted.

Does Consciousness Develop?

Most children show their first signs of an objective self-concept midway through the second year of life – they use personal pronouns, display self-conscious emotions such as shame and embarrassment, and recognize themselves in mirrors. In the mirror self-recognition paradigm, an experimenter surreptitiously puts rouge on a toddler's nose and then exposes the child to a mirror. Typically, children first exhibit mark-directed behavior in this situation between about 18 and 24 months of age, and this behavior has been interpreted as evidence of a transformation from the sensory consciousness of infancy to the self-consciousness of childhood and beyond. On this view, infants are limited to awareness of present sensations until an enormous transformation (usually presumed to be neurocognitive in nature) occurs that simultaneously adds multiple dimensions to the qualitative character of experience: self-other, past-future, etc.

More recently, and in light of research on the behavioral and neural correlates of consciousness

in adults, developmental psychologists have explored the possibility that consciousness develops more gradually through a series of not just two but several dissociable levels. Developmental increases in the highest level of consciousness that children are able to experience may help explain well documented age-related increases in autobiographical memory, theory of mind, the complexity of children's explicit knowledge structures, and the conscious control of thought, action, and emotion.

The Emergence of Consciousness in Ontogeny

One of the most vexing questions concerning the development of consciousness is when – at what age – does a human being first become conscious? Some scientists have argued that consciousness, including the conscious perception of pain, first emerges during the third trimester (at around 28–32 weeks gestational age), because that is when there is first evidence of functional neural pathways connecting the thalamus to sensory cortex. These thalamocortical connections are established as early as 24 weeks gestational age, but the first evidence of functionality does not occur until several weeks later, as indicated by sensory evoked potentials recorded in preterm infants. Other neural events also occur at about this time, including the appearance of bilaterally synchronous electroencephalographic (EEG) patterns of activation (bursts) and EEG patterns that distinguish between sleep and wakefulness. There are also changes in fetal behavior, such as the emergence of clear heart rate increases to vibroacoustic stimuli, as well as the first evidence of habituation to such stimuli. Of course, one can only speculate about the implications of these events for fetal subjective experience, but it is clear that a 30-week-old fetus is similar in most respects to a full-term newborn.

Simple sentience – feeling – must be mediated by some kind of consciousness, however minimal. In the absence of evidence of self-reflection, however, one might suppose that infant consciousness is indeed limited to awareness of present sensations. The construct of minimal consciousness is intended to capture the simplest kind of first-order consciousness on the basis of which more complex forms of consciousness are constructed (through degrees of reprocessing). Minimal consciousness is intentional (i.e., it has content), and it motivates approach and avoidance behavior – a feature essential to its evolution. However, it is unreflective and present-oriented, and it makes no reference to an explicit sense of self; these features develop during the course of childhood. While minimally conscious, one is conscious of what one sees (i.e., the object of one's experience), but one is not conscious of seeing what one sees or that one (as a self) is seeing what one sees. Moreover, because minimal consciousness is tied to ongoing stimulation, one cannot recall seeing what one saw. Minimally conscious information processing may be described as implicit, as when we drive a car without full awareness. In this example, our overlearned driving behavior is elicited automatically by our fleeting, present-oriented consciousness of environmental stimuli, although if we encounter a problem, then we may reflect on our experience and question the relation between what we experience and what we do. Young infants, in contrast, may be limited to the first-order processing of ongoing intero- and exteroceptive stimulation, without recourse to reflection.

Beyond Minimal Consciousness

Studying consciousness beyond its first emergence remains difficult, but if one assumes the existence of minimal consciousness at birth, then the problem of the explaining subsequent developments of consciousness becomes vastly more tractable. In other words, if minimal consciousness is taken as a theoretical primitive, then it is possible to explain the development of more complex forms of consciousness by hypothesizing that minimal consciousness comes to figure in more complex functional relations. One such approach, associated with the levels of consciousness model, is to suggest that the simple sensory consciousness of the young infant develops through an iterative process of reflection whereby the contents of minimal consciousness are recursively reprocessed via recurrent thalamocortical circuits involving regions of the prefrontal cortex. Each degree of reprocessing results in a higher level of consciousness, and this in turn allows for the integration of more

information into an experience of a stimulus before a new stimulus is experienced; it allows the stimulus to be considered relative to a larger interpretive context. On this view, the prefrontal cortex may not be necessary for simple sensory consciousness, but it plays a key role in reflective consciousness, including various degrees of self-awareness.

Prefrontal Cortex Is Important for Self-Awareness

The role of the prefrontal cortex in self-awareness is revealed by case studies of patients with prefrontal cortical damage, and by neuroimaging studies of healthy adults. Many classic case studies, such as the case of Phineas Gage, involve patients who are grossly insensitive to the consequences of their behavior. For example, they may make disastrous financial decisions, and have severe difficulty maintaining personal relationships. This insensitivity, in itself, suggests a failure to reflect on plans and behavior, as well as a failure to learn from mistakes. But these patients also have difficulty reflecting on the nature of their deficits – they blame others for their persistent problems. This lack of insight cannot easily be attributed to general cognitive impairments because most aspects of cognition remain intact. Indeed, when encouraged to adopt a third-person perspective on their behavior (e.g., as a role-playing exercise), patients with prefrontal damage may demonstrate considerable understanding of their condition. In one case, a patient who was having difficulty at work as a result of his injury was asked to participate in a role-playing exercise in which he played the role of therapist. This patient gave excellent therapeutic advice but then promptly refused to heed this advice himself.

An even more striking disorder of self-awareness is seen in patients with Capgras syndrome, a memory disturbance associated with prefrontal damage (usually to the right hemisphere) and in which patients come to believe that a person or a place has been duplicated. This disturbance is isolated in that it occurs in the context of memory that is otherwise quite good. One such patient returned to his family after recovering from his injury, but was convinced that this family was a second, different family, similar in most respects to the family he had before his injury but approximately one year older. Capgras syndrome has been interpreted as evidence that the prefrontal cortex is important for the self-reflective integration of memories and current experience into a more coherent explanation of one's personal history.

Case studies such as these are supplemented by a growing body of neuroimaging research that identifies a key role for the prefrontal cortex in episodic memory, consideration of the relevance of stimuli to oneself, thinking about one's own and other's mental states (i.e., theory of mind), awareness of contingencies during associative learning, and various other, related phenomena.

The Slow Development of Prefrontal Cortex

Recent research has confirmed that the prefrontal cortex is one of the last regions of the brain to reach maturity during human development. The slow course of the development of the prefrontal cortex has been charted in several ways, using measures of gray matter volume, cortical thickness, synaptic density, white matter volume, and functional activity, among other measures. This research shows that the prefrontal cortex develops rapidly in early childhood, with important changes occurring at particular ages (e.g., at the end of the first year of life, between 3 and 6 years, and around puberty), and then continues to develop into adulthood. Gray matter, for example, does not reach adult levels of volume in the dorsolateral prefrontal cortex until at least the end of adolescence, and myelination in this region continues into the 20s or possibly 30s. Given the important role that the prefrontal cortex plays in more reflective aspects of consciousness, it seems plausible that the growth of networks involving the prefrontal cortex might have important consequences for the structure and function of children's consciousness.

The Development of Reflection

Several major developmental transitions in children's behavior might parsimoniously be interpreted in

terms of the development of reflection. Toward the end of the first year of life, for example, most infants exhibit a cluster of new behaviors. They speak their first words, begin to use objects in a functional way, point proto-declaratively at objects, and start searching in a more flexible way for hidden objects (e.g., passing Piaget's A-not-B task), among other milestones. The rapid appearance of these new behaviors suggests a fundamental change in the way children relate to their environment, and one possibility is that the behaviors are made possible by the emergence of a new level of consciousness – 'recursive consciousness.' In recursive consciousness, the contents of minimal consciousness at one moment are reprocessed via thalamocortical loops involving the prefrontal cortex, allowing the toddler to label the initial contents of minimal consciousness. Because a label can be decoupled from the experience labeled, the label provides an enduring trace of that experience that can be deposited into both long-term memory (allowing epidosic recollection) and working memory. The contents of working memory (e.g., representations of hidden objects) can then serve as explicit goals to trigger action programs indirectly so that the toddler is no longer restricted to responses triggered directly by minimal consciousness of an immediately present stimulus.

Consider how such a change in consciousness – the development of recursive consciousness – might account for changes in children's search for hidden objects. In Jean Piaget's famous A-not-B task, for example, an object is hidden conspicuously at one of two locations (location A) and infants are allowed to retrieve it. This is repeated for several trials, and then the object is hidden conspicuously at the other location (location B) and infants are allowed to search for it. Many 9-month-old infants search incorrectly, returning to location A even though they just watched the object being hidden at location B. By 11 months of age or so, infants are much more likely to search correctly in this situation. Recursive consciousness would allow children to label the object's current location and keep this location in mind (i.e., in working memory) instead of relying on a more superficial gloss of the situation that is associated with reaching toward location A.

The way in which reflection would permit the top-down control of behavior is depicted graphically in Figure 1. In Panel 1, minimal consciousness mediates between an object in the environment (objA) and a response. ObjA triggers a salient, low-resolution 'description' from semantic long-term memory, and this description (or IobjA, for 'intentional object') then becomes an intentional object of minimal consciousness, by way of which it automatically triggers the most strongly associated action program in procedural long-term memory or elicits a stored stimulus-reward association. Location A, for example, may have been associated with interesting activity (e.g., a hiding event) or a reward (e.g., retrieving an object), and so, when seen, may elicit reaching toward that location.

In Panel 2, reflection occurs prior to responding. In this case, the contents of minimal consciousness are fed back into minimal consciousness (at a subsequent moment) where they can be related to a label (descA) from semantic long-term memory. This descA can then be decoupled from the minimally consciousness experience that was labeled, and it can be deposited into working memory where it can serve as a goal (G1) that triggers an action program even in the absence of objA, even though IobjA would otherwise trigger a different action program. In this way, the reflective reprocessing of information prior to action permits the top-down, cognitive control of behavior despite interference from prepotent response tendencies.

An Emerging Sense of Self

The change in children's consciousness hypothesized to occur at the end of the first year of life may be just the first of a series of age-related increases in children's ability to reflect on their experiences and consider additional aspects of stimuli and the context in which these stimuli occur. Another change may occur during the second half of the second year and account for the apparent emergence of an awareness of self. More specifically, reflection on the contents of recursive consciousness would allow children to consider not only an object in the environment but also the relation between an object and a description of themselves, including their behavioral potential.

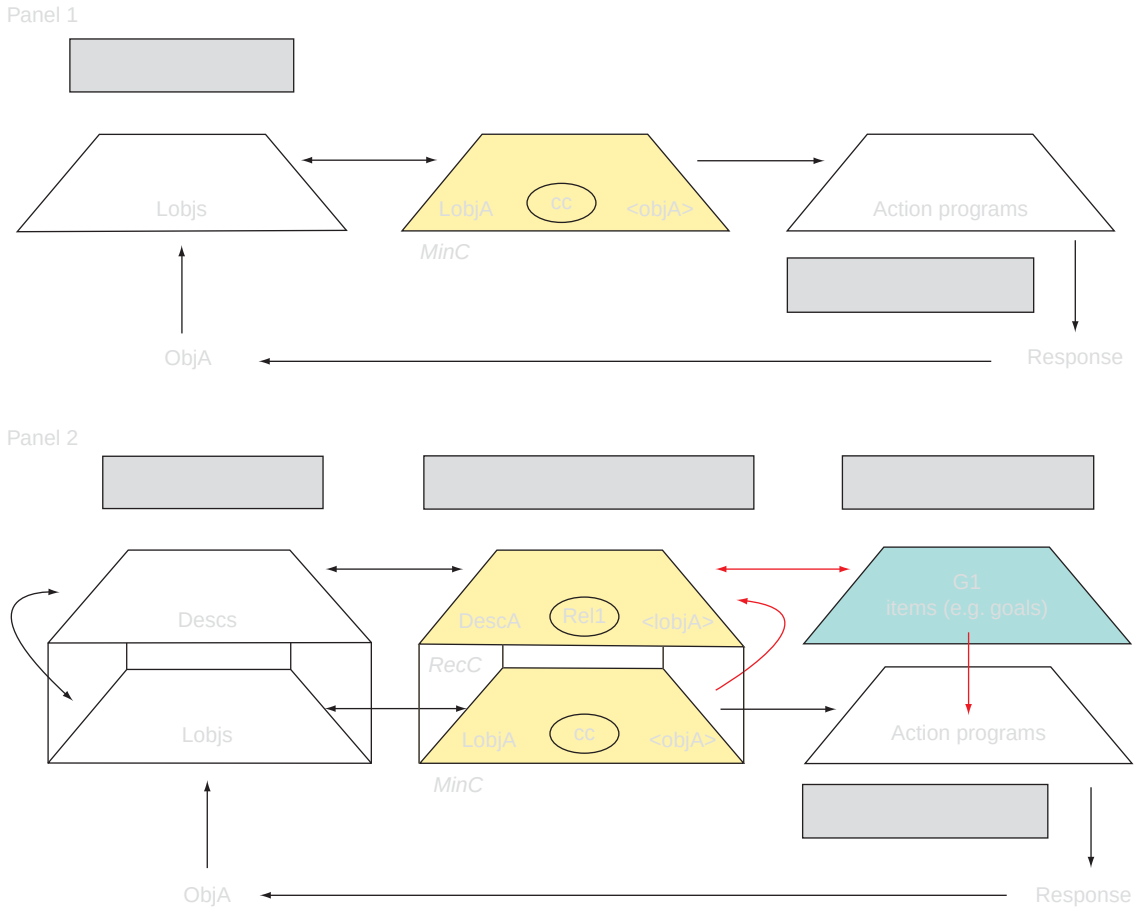


Figure 1 The implications of reflection (levels of consciousness) for search. Panel 1: Automatic action on the basis of unreflective consciousness. An object in the environment (objA) triggers an intentional representation of that object (lobjA) in semantic long-term memory (LTM); this lobjA, which is causally connected (cc) to a bracketed objA, becomes the content of consciousness (referred to at this level as minimal consciousness or minC). Panel 2: Action on the basis of one degree of reflection. Following minC processing of the objA, the contents of minC are then fed back into minC via a reentrant feedback process, producing a new, more reflective level of consciousness referred to as recursive consciousness or recC. The contents of recC can be related (rel) in consciousness to a corresponding description (descA) or label, which can then be decoupled from the experience labeled and deposited into working memory (WM) where it can serve as a goal (G1) to trigger an action program in a top-down fashion from procedural LTM. Reprinted from Zelazo PD (2004). The development of conscious control in childhood. Trends in Cognitive Sciences 8: 12–17, Copyright, with permission from Elsevier.

Development of More Complex Experiences of Self and Other

Two-year-olds' experience of themselves in time remains limited, however, even 3-year-olds perform poorly on measures of delayed self-recognition. In one paradigm, children play a game during which an experimenter surreptitiously places a sticker on their heads. About 3 min later, children are presented with a video image of the marking event. Whereas the majority of 4-year-olds reach up to touch the sticker, most 3-year-olds fail to do so. According to some authors, 3-year-olds maintain a succession of present-oriented representations of self, and they cannot compare these representations or integrate memories with current experiences – much like the patients with Capgras syndrome.

Three-year-olds also resemble patients with prefrontal cortical damage in other ways. First, they give prudent advice to others but then fail to heed that advice themselves, possibly due to difficulty appreciating the personal relevance of

objective knowledge. In one study, 3-year-olds were asked whether the experimenter should take a small reward now or wait and get a larger reward later. Children usually recommended that the experimenter wait and take the larger reward. When children of the same age were given the same choice for themselves, however, they typically chose the smaller, immediate reward. Four-year-olds made similar recommendations for self and for other.

A second way in which 3-year-olds resemble patients with prefrontal damage is in their well documented difficulty with theory of mind. In one standard theory of mind task, children are shown a familiar container (e.g., a Smarties box) and asked what it contains. Subsequently, the container is opened to reveal something unexpected (e.g., string) and children are asked to recall their initial incorrect expectation about its contents: "What did you think was in the box before I opened it?" In order to answer this question correctly, children must be able to recollect (or reconstruct) their initial false belief and consider it in relation to an outdated perspective. Most 3-year-olds respond incorrectly, stating (for example) that they initially thought that the box contained string.

A third example of the resemblance between preschoolers and prefrontal patients is in their difficulty using complex rules to control their behavior. In the Dimensional Change Card Sort, children are shown two bivalent, bidimensional target cards (e.g., depicting a blue rabbit and a red boat), and they are told to match a series of test cards (e.g., red rabbits and blue boats) to these target cards first according to one dimension (e.g., color) and then according to the other (e.g., shape). Regardless of which dimension is presented first, 3-year-olds typically perseverate by continuing to sort cards by the first dimension after the rule is changed. Moreover, reminiscent of the "curious dissociation between knowing and doing" so often noted in patients with prefrontal damage, they do this despite responding correctly to explicit questions about the post-switch rules. For example, children who should be sorting by shape (but persist in sorting by color) may be asked, "Where do the rabbits go in the shape game? And where do the boats go?" Children almost always answer these questions correctly. When they are then told to sort a test card ("Okay, good, now play the shape game: Where does this rabbit go?"), however, they persist in sorting by color.

This type of dissociation lends itself well to the suggestion that 3-year-olds consciously represent the postswitch rules at one level of consciousness (which allows them to provide verbal answers to the explicit knowledge questions), and consciously represent the preswitch rules at that same level of consciousness (which allows them to keep the preswitch rules in working memory to guide their sorting). Because they fail to reflect on their representations of the two rule pairs from a higher level of consciousness, however, they cannot consider them in contradistinction and make a deliberate decision about which pair of rules to use.

By contrast, 4-year-olds, like adults, seem to recognize immediately that they know two ways of construing the stimuli. At this age, children may spontaneously reflect on their multiple perspectives on the situation, consider them from a higher level of consciousness, and integrate them into a relatively complex rule structure: "If we're sorting by shape, then if it's a red rabbit, it goes here. . ." Children may do something similar in theory of mind tasks, where they need to be able to say to themselves, in effect, "From my perspective, there are sticks in the box, not crayons; but I first thought that there were crayons, not sticks." Indeed, it is now well established that children's performance on the Dimensional Change Card Sort is correlated with their ability to reflect on their own and others' mental states in tasks assessing theory of mind, even when age and general intellectual ability are controlled.

Labeling a Subjective Perspective Encourages Reflection on that Perspective

Although 4-year-olds seem capable of reflecting on incompatible perspectives on a single situation, they do not always do so. One recent line of research using the Flexible Item Selection Task has examined the extent to which labeling 4-year-olds' perspectives facilitates their reflection on those perspectives. In the Flexible Item Selection Task, children are shown sets of three items designed so one pair matches on one dimension, and a different pair matches on a

different dimension (e.g., a small yellow teapot, a large yellow teapot, and a large yellow shoe). Children are first told to select one pair (i.e., Selection 1), and then asked to select a different pair (i.e., Selection 2). To respond correctly, children must represent the pivot item (i.e., the large yellow teapot) according to both dimensions. Four-year-olds generally perform well on Selection 1 but poorly on Selection 2, indicating inflexibility. When 4-year-olds are asked to label their perspective on Selection 1 (e.g., ''Why do those two pictures go together?''), however, their performance on Selection 2 improves dramatically. This finding is consistent with the possibility that labeling their initial construal of the pivot item caused them to make that subjective perspective an object of consideration, requiring them to step back from that perspective in what has been referred to as psychological distance, and allowing them to adopt a different perspective on Selection 2.

Relatively little is known about the development of consciousness beyond childhood, but laboratory measures of the prefrontal function reveal improvements in children's behavior at least until late adolescence. In the Stroop color word task, for example, people are shown color words (e.g., the word 'red') printed in nonmatching colored ink (e.g., blue ink) and required to name the color of the ink. Performance on this task improves until at least age 17 years. Similar improvements are seen on various measures of adaptive decision making, and there are suggestions that these age-related improvements may occur in ways that recapitulate changes seen earlier in childhood. For example, in one study of risk taking and delay of gratification, younger adolescents were more likely than older adolescents to make less prudent judgments for themselves than for other people – just like the dissociations seen in preschoolers and prefrontal patients.

The Rise and Fall of Reflection

Performance on laboratory measures of prefrontal function appears to reach a ceiling by late adolescence, failing to show further improvement, although it is possible that new measures will prove more sensitive to subtle changes occurring beyond this age. Even existing measures, however, are sensitive to changes that occur at the end of the lifespan, when difficulty exercising top-down, conscious control over thoughts, actions, and emotions reappears. For example, compared to young adults, elderly adults show a decline in performance on the Stroop task and on versions of the Dimensional Change Card Sort. Other well-known correlates of aging, such as increased forgetting and unwanted intrusions in one's speech, may be attributable, to some extent, to difficulty reflecting on one's representations vis-à-vis the current context. Indeed, one study found that performance on the Dimensional Change Card Sort in children, young adults, and elderly adults was predicted by independent estimates of the conscious (vs. automatic) influences on memory generated by the process dissociation procedure. Thus, the development of prefrontal function appears to follow an inverted-U-shaped curve when considered across the lifespan, and it is possible that this development is associated with the rise and fall of reflective reprocessing.

Conclusion

The discrete levels of consciousness suggested by developmental research may be useful for understanding the complex, graded structure of conscious experience in adults. Reflection changes the structure of experience, and age-related changes in the possibility of reflection would have consequences for the quality of subjective experience: the addition of higher levels would result in a richer, more detailed experience. Variations in the degree to which one reflects on one's experience may account for differences in episodic recollection, the complexity of one's cognitive structures, and the possibility of the conscious control of thought, emotion, and action.

The highest, most developmentally sophisticated level of consciousness that one is able to muster (e.g., when one's expectations are violated or when one encounters a problem) may change with age. Within these age-related constraints, however, individuals may be more or less reflective, either as a matter of cognitive style or as a function of circumstance (e.g., when tired or distracted or performing routine, automatic operations). Age-related constraints on levels of consciousness may be tied to the

maturation of regions of the prefrontal cortex, although anything that interferes with the rapid reprocessing of information through circuits involving the prefrontal cortex would be expected to decrease the likelihood of reflection. Further research on the development of the prefrontal cortical function may reveal in more detail the mechanisms underlying the iterative reprocessing of information, and shed light on the relation between subjective experience and behavior.

See also: Philosophical Accounts of Self-Awareness and Introspection; Self: Personal Identity; Self: The Unity of Self, Self-Consistency; Theory of Mind (Neural Basis).

Suggested Readings

Dalton TC and Bergenn VW (2007) Early Experience, The Brain, and Consciousness: An Historical and Interdisciplinary Synthesis. New York: Lawrence Erlbaum Associates.

Kagan J (1981) The Second Year: The Emergence of Self-Awareness. Cambridge: Harvard University Press.

Karmiloff-Smith A (1992) Beyond Modularity: A Developmental Perspective on Cognitive Science. Cambridge, MA/London: MIT Press.

Lewis M and Brooks-Gunn J (1979) Social Cognition and the Acquisition of Self. New York: Plenum.

Lewis M and Ramsay D (1999) Intentions, consciousness, and pretend play. In: Zelazo PD, Astington JW, and Olson DR (eds.) Developing Theories of Intention: Social Understanding and Self-Control, pp. 77–94. Mahwah, NJ: Lawrence Erlbaum Associates.

Nelson K (2005) Emerging levels of consciousness in early human development. In: Terrace HS and Metcalfe J (eds.) The Missing Link in Cognition: Origins of Self-Reflective Consciousness, pp. 116–141. New York: Oxford University Press.

Perner J and Dienes Z (2003) Developmental aspects of consciousness: How much of a theory of mind do you need to be consciously aware? Consciousness and Cognition 12: 63–82.

Povinelli DJ (2001) The Self: Elevated in consciousness and extended in time. In: Moore C and Lemmon K (eds.) The Self in Time: Developmental Perspectives, pp. 73–94. Mahwah, NJ: Lawrence Erlbaum Associates.

Rochat P (2003) Five levels of self-awareness as they unfold early in life. Consciousness and Cognition 12: 717–731.

Stuss DT (1991) Disturbance of self-awareness after frontal system damage. In: Prigatano GP and Schacter DL (eds.) Awareness of Deficit after Brain Injury: Clinical and Theoretical Issues, pp. 63–83. New York: Oxford University Press.

Stuss DT (1991) Self, awareness, and the frontal lobes: A neuropsychological perspective. In: Strauss J and Goethals GR (eds.) The Self: Interdisciplinary Approaches, pp. 255–278. New York: Springer-Verlag.

Stuss DT, Gallup GG Jr., and Alexander MP (2001) The frontal lobes are necessary for 'theory of mind.' Brain 124: 279–286.

Wheeler M (2000) Varieties of consciousness and memory in the developing child. In: Tulving E (ed.) Memory, Consciousness, and the Brain: The Tallinn Conference, pp. 188–199. London: Psychology Press.

Zelazo PD (2004) The development of conscious control in childhood. Trends in Cognitive Sciences 8: 12–17.

Zelazo PD, Gao HH, and Todd R (2007) The development of consciousness. In: Zelazo PD, Moscovitch M and Thompson E (eds.) The Cambridge Handbook of Consciousness, pp. 405–432. New York: Cambridge University Press.

Biographical Sketch

Philip David Zelazo received his BA (Hons.) from McGill University and his PhD (with distinction) from Yale University. From 1992 to 2007, he taught at the University of Toronto, where he held a Canada Research Chair in Developmental Neuroscience.

He is currently the Nancy M. and John E. Lindahl Professor at the Institute of Child Development, University of Minnesota, and the codirector of the Sino-Canadian Centre for Research in Child Development, at Southwest University, China. Professor Zelazo's research, which centers on the development and neural bases of executive function (or the conscious control of thought, action, and emotion), has been honored by numerous awards, including a Boyd McCandless Young Scientist Award from the American Psychological Association, a Premier's Research Excellence Award from the Government of Ontario, and a Canada's Top 40 Under 40 Award. He is a fellow of the Canadian Institute for Advanced Research (Experience-based Brain and Biological Development Program), he serves on the board of directors of the Jean Piaget Society, he is a member of several editorial boards (*Child Development*, *Emotion*, *Cognitive Development*, *Journal of Cognition and Development*, and *Monographs of the Society for Research in Child Development*), and he is the coeditor (with Morris Moscovitch and Evan Thompson) of *The Cambridge Handbook of Consciousness* (2007).

Emotion and Consciousness

S Sher and P Winkielman, University of California, San Diego, CA, USA

Glossary

Affect – The activation of valence-based representations reflecting the organism's assessment of 'goodness–badness' across multiple neural/psychological systems.

Affective priming – Experimental procedures in which subjects are exposed to valenced stimuli ('primes') and their effects on responses to simultaneously or subsequently presented other stimuli ('targets') are measured. For example, exposure to subliminal angry face primes may, on average, lower evaluations of the attractiveness of subsequently presented shapes or willingness to try a novel drink.

Attention – Process(es) resolving capacity limits in perceptual and cognitive processing, in which some stimuli or ideas are in effect 'selected' over others for preferential processing and/or awareness.

Attribution – An implicit or explicit determination of the source or cause of an observed event, that is, in the present context, a mood or emotional state.

Binding problem – When inputs are analyzed separately along multiple feature dimensions, the problem of knowing which features go with which is known as the binding problem. In visual perception, solving the binding problem allows one to perceive a yellow sun against a blue sky, rather than the other way around. In affective processing, the problem is to determine which affective reactions go with which eliciting stimuli.

Blindsight – An ability to guess, with above-chance accuracy, certain properties of visual inputs of which the subject is not consciously aware. Blindsight has principally been explored in patients with damage to primary visual cortex, though there have also been attempts to demonstrate blindsight in normal subjects. Similar phenomena (e.g., 'deaf hearing,' 'numb touch') in other sensory modalities have also been investigated.

Consciousness – Heated argument surrounds all attempts at definition; arguably indefinable. In experimental practice, 'consciousness' is commonly taken to refer to representations of information with several linked properties, including wide availability of information to cognitive systems and response modalities; flexibility of representation, learning, and response in novel circumstances; and expressed convictions about the subjective 'reality' of the representation.

Emotion – Finer-tuned, qualitatively differentiated variants of affect (e.g., anger, disgust), motivating more specific reactions in more specific situations.

Implicit measures – Measured effects of external stimuli or internal states on physiological reactions or behaviors not involving explicit description of those stimuli or states. Implicit and explicit measures may partly or entirely tap into different pathways of information processing, and correspondingly may at times yield discrepant pictures of the subject's psychological state.

Metaconsciousness – Higher-order knowledge about one's own (past or present) conscious states. Metaconscious awareness comes in many shades and varieties; it may involve simple awareness that one is in the conscious state one is in, or more sophisticated beliefs about the relations between the conscious state and other states and stimuli.

Neglect – A neurological condition in which, owing to brain damage on one side of the brain, the patient shows a marked deficit in attention to and awareness of objects on the

opposite side of perceptual space. This spatial neglect may manifest in multiple reference frames. For example, the left half of visual space, or the left halves of objects located in both halves of visual space, may be neglected. Neglect often resolves over time into a phenomenon known as 'extinction,' in which objects on the neglected side are missed only when another object is simultaneously presented to the nonneglected side.

Subliminal – Refers to perceptual input, which, because of its presentation parameters, is not consciously accessible – that is, the subject would be unable to report explicitly about its properties if he or she tried. Subliminal stimuli can be delivered in different modalities (sight, hearing, touch, etc.), and subliminality is usually achieved by some combination of low stimulus intensity; very brief stimulus presentation; and forward or backward masking, in which another 'masking' stimulus, nearby in space and time, helps to obliterate awareness of a briefly exposed stimulus. Subliminality is generally verified in psychological experiments by subjective tests (e.g., ''Did you see any face?'') and/or objective tests in which subjects guess about some property of the stimulus (e.g., ''A face was flashed – guess its gender'').

Valence – A single, simple encompassing dimension of evaluation, ranging from negative ('bad') to positive ('good').

Introduction

''How do you feel right now?'' is a question we often ask (sometimes meaningfully) and answer (sometimes honestly). Yet our answers to this particular question, unlike other questions, rarely elicit the rejoinder: ''How do you know?'' ''Do you know?'' is another question that might reasonably be asked, but usually is not. This article summarizes and comments on empirical research that attempts to seriously pose these unorthodox questions: How do you know how you feel right now. . .if you know?

Put in more sophisticated, if perhaps less revealing, terms, we describe research on the relationship between emotion (or affect) and consciousness. 'Affect' is usually taken to refer to the simultaneous activation of valence representations (i.e., representations of goodness or badness) across multiple neural/psychological systems. 'Emotion' refers to more fine-tuned variants of affect (anger, sadness, contempt, guilt), which motivate more specific reactions (screaming, weeping, turning up one's nose, apologizing) to more specific situations (injury, loss, perceived inferiority of another, transgression). We will sometimes use the terms 'affect' and 'emotion' interchangeably, but, as we note below, the consciousness requirements for specific forms of emotional processing may conceivably be more stringent than those for more general forms of affective response.

Definitions of 'consciousness' vary widely in their pretensions and limitations, ranging from the austere but simplistic (e.g., 'what the subject reports') to the evocative but tautological (e.g., 'what the subject experiences'). For our purposes, it will suffice to observe that some instances of human information processing, but not others, are characterized by a complex syndrome of interrelated properties: (1) widespread and coordinated availability of information to a broad range of response systems, including those involved in verbal reports, button-presses, etc.; (2) flexibility in dealing with and learning from novel situations; and (3) the reported feeling that the current experience is in some sense 'real.' In exploring conscious and unconscious emotional phenomena, we will be content to contrast cases in which these properties are jointly and robustly present (which we will call 'conscious') from cases in which these properties appear to be jointly absent (which we will call 'unconscious'). To explore these phenomena systematically, our treatment is structured into three sections, which address three different levels of consciousness which may or may not be associated with an emotion.

Emotional states need not be triggered by simple stimuli in simple ways. But when they are, we can pose three questions about the relationship between emotion and consciousness: (1) Is the

subject conscious of the stimulus that triggers the emotion? (2) Is the emotion itself conscious? And (3) if the subject is conscious of both the eliciting stimulus and the elicited emotion, is he or she aware of the connection between stimulus and emotion – that is, does the subject accurately attribute the emotion as effect to the stimulus as cause?

Affective Processing of Unconscious and Unattended Stimuli

A large literature has investigated affective reactions to stimuli of which the subject is unaware, using methods drawn from cognitive psychology, neuropsychology, and neuroimaging. In a typical experiment using a paradigm called 'subliminal affective priming,' developed by Robert Zajonc and his colleagues, the subject looks at a screen on which an unfamiliar and affectively ambiguous target shape – for example, an ideograph from an unknown language – is presented. The subject's task is to form an overt evaluation of the target – for example, to rate the attractiveness of the ideograph on a numerical scale. Unbeknownst to the subject, however, a photograph of a face is very briefly flashed on the screen just before the target shape appears at the same location, replacing it. Under the right timing parameters, the target shape has the effect of 'masking' the briefly flashed face – that is, preventing it from entering consciousness. Nonetheless, numerous studies have found that subjects evaluate the target shape more favorably, on average, when the subliminal face that preceded it had a happy rather than an angry or fearful expression. Despite the fact that, as indexed by subjective reports or objective forced-choice tests, the subject appears to be unaware of the faces, the facial expressions are found to bias evaluations of subsequent target shapes. As discussed in the section 'Conscious and unconscious emotion,' subliminal facial expressions have also been found to have effects on behaviors other than overt evaluation, including risk-taking in gambling situations and consumption behavior.

Evidence from affective blindsight further corroborates the notion that affective properties of a visual stimulus can be partly computed in the brain even when the subject is not conscious of the stimulus. Extensive damage to the primary visual area of the cerebral cortex may leave a patient phenomenally blind over part or all of their visual field. Nonetheless, when forced to guess about some property of a visual stimulus presented in their blind field – for example, ''Is the line segment vertical or horizontal?'' – such patients may make inspired guesses, a phenomenon known as 'blindsight' – for example, correctly guessing the orientation of the unseen line segment far over 50% of the time. Recently, some cases of 'affective' blindsight have been reported. When images of emotionally expressive faces are presented to these patients' blind fields, they are sometimes nonetheless able to guess, with high accuracy, the affective valence of the faces. Neuroimaging studies of these patients have associated their inspired guesses about the image's affective properties with changes in amygdala activity. Parallel findings have emerged from neuroimaging studies of normal subjects, with higher amygdala activation upon exposure to emotionally charged – and especially to fearful – faces, even when these faces are not consciously perceived.

Thus at least some affect-related processing takes place for stimuli that, because of brain damage or brief stimulus exposures, fail to reach awareness. An important related question concerns the role of attention in affective evaluation. It is widely accepted that human perception is subject to steep capacity limits, such that enhancing processing in one part of the perceptual field – by allocating attention to one object or spatial region – impairs, in certain respects, the processing of other parts of the field.[1] But there has long been controversy about the relationship between attention and awareness – and in particular about the fate of nonattended perceptual input. Is attention to a stimulus necessary for conscious awareness of that stimulus? Further, can attention influence processing of an unconscious stimulus? Cognitive experiments on change blindness – in which subjects are oblivious even to gross changes between

[1]It is useful to distinguish between attention to objects or locations, on the one hand, and attention to different properties of a fixed object at a fixed location, on the other. The studies summarized here involve the allocation of attention to different objects or locations, though in some cases these shifts of attention are confounded with shifts of attention between different kinds of properties.

one visual display and another – and inattentional blindness – in which subjects, attending to one part of a display, seem to have no awareness of a stimulus that pops up in another part of the display – have been advanced as evidence that we have no consciousness of what we do not attend to. However, the interpretation of these results, as well as the larger question of the relationship between attention and awareness, remains hotly disputed. Furthermore, whether or not awareness itself requires attention, recent behavioral and neuropsychological studies in cognitive psychology suggest that attending to a location influences the processing of stimuli at that location, even when these stimuli remain unconscious.

Affective processing of perceptual input has often been conceived as 'automatic' – that is, proceeding independently of attention and cognitive strategies, perhaps along neural pathways distinct from those which culminate in consciousness of the visual and other perceptual features of the input. Some evidence for the relative independence of affective processing from attention comes from studies in which affect-laden (especially unpleasant) stimuli appear to grab attention to themselves – or to resist experimental manipulations that tend to diminish attention – suggesting that their affective properties may to some extent be 'preattentively' computed (though the evidence along these lines is somewhat mixed).[2] This research, arguing for affective processing without attention, generally adapts attention-related paradigms widely used in cognitive psychology, including visual search, 'attentional blink,' and attentional cueing tasks. However, evidence that affective processing can be modulated by attention has been accumulating.

The relationship between emotion and attention has recently come under intensive focus in neuroimaging experiments using functional magnetic resonance imaging (fMRI). As noted earlier, increased amygdala activation to emotional (especially fearful or angry) faces has been found in a number of fMRI studies, using both supraliminal (above-threshold) and subliminal (below-threshold) faces. A key question addressed by recent studies concerns the degree to which this differential amygdala activation for emotional faces persists when the faces are unattended. In these studies, subjects see displays containing faces and other objects. In different conditions, the task requires subjects either to attend to and make a judgment about the faces, or to attend to and make a judgment about the other objects in the same displays. The question is whether the observed amygdala response to the faces is reduced, eliminated, or otherwise altered when attention is directed away from the faces. (Measures are generally taken to ensure that shifts of attention are not accompanied by movements of the eyes.) The picture emerging from these studies is still a somewhat complex work in progress, with conflicting results between different studies using different attention tasks. But some general conclusions can be drawn. While several studies failed to find any attentional modulation of affect-related amygdala activation, numerous other studies have observed such modulations. In one influential study by Luiz Pessoa and colleagues, none of the affect-related activations found when attention was focused on happy or angry faces survived when attention was allocated to other items in the same displays. The results of this and a number of other neuroimaging studies indicate that attention can modulate amygdala responses to emotional faces under some conditions, and sometimes dramatically – but it is not entirely clear how general these effects of attention are.

This developing literature is actively seeking to identify factors that will explain and thus reconcile the different results found in different experiments. These potentially relevant factors include: how demanding the distracting task is, with evidence that more demanding distracting tasks may more severely limit activation from unattended faces; the location of the face images in the visual field, which may influence the visual pathways through which the face information flows; the extent to which the tasks involve active suppression of irrelevant stimuli, as opposed to mere inattention; and the role of individual differences in different subject populations (e.g., people with

[2] In general, it is important to note that the relationships between affect and attention, and between affect and awareness, are two-way streets: Just as attention and awareness may influence affective processing, so affective processing may change the probability that a stimulus will reach the threshold for awareness and whether our attention is drawn to it.

anxiety disorders). It is important to remember that the levels of amygdala activation revealed in an fMRI study supply only a crude measure of what the amygdala may or may not be computing. Other studies have used event-related potentials, recorded at the scalp, to try to separate out the effects of attention on different components of the brain's response to emotional stimuli. In addition, while these studies have focused on affective processing of face stimuli, different kinds of affective stimuli may show different kinds and degrees of attentional modulation. For example, faces and words differ in evolutionary significance, in visual properties such as spatial frequency, in how they are learned, and in their neural coding; attentional constraints on affective processing in the two cases may (or may not) be correspondingly different. In any event, while the roles of the various factors described above remain to be fully sorted out, it has by now repeatedly been shown that, under suitable conditions, the allocation of attention can substantially modulate the amygdala's response to emotional faces.

Another relatively underexplored problem in the relationship between attention and emotion concerns affective binding. According to Anne Treisman's feature integration theory, one essential function of spatial attention is to solve the visual system's 'binding problem.' In distributing your attention across a visual display, you may perceive that redness, greenness, a circle, and a square are all present. However, according to the theory, the formation of a coherent integrated percept – knowing that the circle is red and the square is green, rather than the other way around – may require focused attention. That is, attention focused at a location may be needed to bind the colors, shapes, and other visual features at that location together. Insofar as affective properties can be processed in the absence of focused attention, feature integration theory raises two questions about the relationship between emotion and attention. First, when the affective value of a stimulus depends on the precise combination of the simple visual features in a display (e.g., a brown face next to a green leaf may elicit different affective reactions than would be sparked by a green face next to a brown leaf), must the items be individually attended and their features globally integrated in order for affective responses to be formed? Or might affective processing proceed in part along a different pathway in which focused attention is not required for crude feature integration, at least for familiar stimuli? Second, how do we bind affective evaluations to the objects that elicit them? This corresponds to a more complex case of the problem of affective attribution, described below. In this case, multiple stimuli are present in the visual array, potentially triggering multiple distinct affective evaluations. The problem now is to decide which evaluation goes with which stimulus. Even if attention is not absolutely necessary in the formation of evaluations, it may (or may not) be necessary for their assignment among the different objects we consciously perceive.

Finally, a fascinating study by John Marshall and Peter Halligan illustrates the possibilities for some forms of affective processing when attention and awareness are jointly impaired. These researchers studied a 'neglect' patient who, as a consequence of right-hemisphere brain damage, persistently failed to attend to left visual space and to report awareness of objects appearing there. This patient also had a partial left visual field defect, but the experiment was conducted under free viewing conditions in which the patient, by moving her head, could put 'left space' (e.g., the left half of an object) into her right visual field. The patient was repeatedly shown two house drawings, vertically arrayed. In one of the drawings, flames emerged from a window on the left side of the house, but the two drawings were otherwise identical. While the patient, ignoring the left-side flames, judged the two houses to be exactly the same, she almost always decided (when forced to make what struck her as an entirely arbitrary choice) that she would prefer to live in the house without the flames. A burning house is not an appealing place to live, apparently even when the flames are neither attended not fully perceived.

Taken together, the research summarized above indicates that affective processing of a stimulus can take place in the absence of awareness of that stimulus. Nonetheless, while it is unlikely that focused attention is necessary for all forms of affective processing, such processing can be,

under conditions which have yet to be fully mapped out, strongly affected by, and perhaps dependent on, attention.

Conscious and Unconscious Emotion

In the previous section we considered whether the external triggers of an emotion need to be consciously represented. But what about the emotional response itself? Are emotions necessarily conscious? If not, how do conscious and unconscious emotions differ? To answer these questions, it is worth recalling that 'emotion' is typically defined as a coordinated response to a significant valenced event across several components – perceptual, cognitive, motivational, expressive, bodily, and experiential. Therefore, the key empirical question in this area is which components of a large-scale affective reaction can be activated, but remain unconscious.

For many writers, the most critical element of emotion is the intrinsically conscious, subjective, experiential component. In other words – the essence of feelings is 'the feeling.' Indeed, it is initially difficult even to think about anxiety without conjuring the phenomenal experience of apprehension, worry, loss of control, or impending doom, along with the subjective sense of trembling and sweating. Similarly, it appears that the very essence of love is the subjective feeling of care, attachment, and warmth toward another. Conscious feelings are also powerful motivators of behavior. Few clinical patients complain of unconscious anxiety or depression (though their partners might). Similarly, few people would drink or take recreational drugs if they made them only unconsciously happy or relaxed. Othello's poignantly conscious jealousy poisons his mind and eventually drives his destruction, whereas Romeo is motivated by conscious love and compassion. Unsurprisingly, understanding conscious feelings plays a central role in both research and clinical practice. Thus, emotion researchers have spent many years delineating the various meaning dimensions of conscious feelings (e.g., appraisal and attribution theories), whereas philosophers have explored their phenomenological structure (e.g., Husserl, Brentano, Sartre, Solomon). On the practical end, psychiatrists have focused on pharmacological interventions into affective neurochemistry that alter conscious experience (e.g., benzodiazepines like Valium and SSRIs like Prozac).

However, the idea that emotion can also be unconscious has a long history in emotion research. It goes back at least to Darwin, who described many 'instinctive' (fast, rigid, involuntary) emotional behaviors and speculated about their origins in our remote evolutionary ancestry. Early prototypes of complex emotion presumably evolved to spur appropriate reactions to positive or negative events. Accordingly, many basic behavioral reactions associated with human emotion are widely shared by animals, including reptiles and fish. The evolutionarily old neurocircuitry and neurochemistry underlying basic emotional reactions (fear, liking) is wired into subcortical brain structures, such as the amygdala, nucleus accumbens, hypothalamus, and even brainstem parabrachial nucleus and pons. Indeed, the most direct and effective neural manipulations of basic emotional reactions involve electrical or chemical intervention into subcortical structures. For example, Kent Berridge and colleagues showed that brain microinjections of drugs that activate opioid receptors in subcortical nucleus accumbens elicit increased 'liking' responses for sweetness. Importantly, these effects do not depend on 'higher-order' neural machinery as they occur even in decorticated animals. Similarly, in anencephalic infants, whose brains lack nearly all of the forebrain, including the entire neocortex, sweet tastes still elicit positive facial expressions whereas bitter tastes elicit negative facial expressions. In short, affective neuroscience highlights the role of subcortical structures in basic emotional reactions. This raises the possibility that some causes of human emotion, and perhaps even some emotional reactions themselves, might not be accessible to full-blown conscious awareness.

Data from psychological studies with normal subjects support this possibility. As mentioned in the section 'Affective processing of unconscious and unattended stimuli,' there is now extensive evidence that affect, and perhaps even emotion-like states, can be triggered by stimuli of which

the subject is unaware. But can the emotional reaction itself be unconscious? Experimental evidence suggests that, at least under some circumstances, people are unable to report any shift in conscious emotion even as a consequential behavior appears to reveal the presence of a covert affective reaction. For example, in one series of studies by Winkielman and colleagues, participants were unobtrusively exposed to several subliminal happy or angry facial expressions. Immediately after the subliminal elicitation of affect, participants reported their conscious feelings (mood and arousal) and also consumed and rated a novel beverage. The ratings of conscious feelings were unaffected by subliminal faces. However, participants consumed more beverage after happy rather than after angry faces, and rated the beverage more favorably. Not only was their overt behavior indicative of affective change, but follow-up studies using psychophysiological measures, such as affective startle and facial EMG, revealed that responses of low-level approach/avoidance systems were influenced in an affect-congruent way by the subliminal faces. In short, these results suggest the possibility of genuinely unconscious affect, in the sense of a valenced (positive–negative) reaction that is strong enough to alter behavior and physiology, but of which people are not subjectively aware.

To be sure, there remain many open questions about the conceptualization and mechanisms of unconscious emotion. In addition to probing for neural substrates of conscious and unconscious emotional reactions, ongoing research in several laboratories is examining whether the critical property of unconscious motivating states is simply positive–negative valence (unconscious affect), or whether there are unconscious states that drive behavior in differentiated fashion associated with specific emotions (fear, anger, disgust, sadness, etc). It is also worth exploring the tricky possibility that what sometimes presents as 'unconscious emotion,' as suggested by the failure of emotion self-reports, might sometime represent a failure of constructing or updating an appropriate higher-order self-description of an emotional state (i.e., a problem with meta-awareness, as we discuss shortly). It will take further research with clever designs to address these possibilities. But for now, it seems likely that at least in some conditions not only the processing of emotion triggers, but also to some extent emotional responding, may unfold without reaching full awareness.

Thinking about Feelings and Their Causes

When an emotion and the stimulus that elicits it are both consciously accessible, it is an open question whether the subject will appreciate the causal connection that links them. The broken air conditioner in the museum may suffice to explain the art critic's vague discomfort, but what will prevent the critic from attributing this unease to the paintings on display? More generally, we daily keep ourselves busy diagnosing the causes of our emotional welfare and – especially – our emotional ills. How competent are we at such diagnosis? Do emotions emerge into consciousness tagged with their sources of origin? (The affective influence of unconscious stimuli, summarized above, would suggest that this could not universally be the case.) Or may affective attribution best be conceived as a complex matching problem in which we often guess and sometimes err?

A large body of research has repeatedly demonstrated that the problem of affective attribution is not solved flawlessly. Researchers commonly employ 'mood manipulations' – uncomfortable temperatures, cramped postures, sad music, etc. – to modulate subjects' overall mood states in more or less subtle ways. Just as the scowls of unseen faces depress attractiveness ratings for subsequent novel shapes (see section 'Affective processing of unconscious and unattended stimuli'), so broad shifts in mood, caused by consciously accessible stimuli and conditions, tend to bleed into subjects' affective evaluations of different objects that subsequently come within the focus of attention. The prior fear-related induction of arousal may make a potential romantic partner seem more enticing, and on relatively gloomy days respondents tell survey researchers that their overall life satisfaction is relatively low.

To be sure, such effects need not automatically be interpreted in terms of defective attribution. A grey day may render more salient the greyer aspects of one's existence. However, there is

abundant experimental evidence that such mood manipulation effects can derive from, or at least be strongly modulated by, active processes of attribution. For example, in a classic study by Norbert Schwarz and Gerald Clore, a telephone interview about life satisfaction was conducted. As noted above, survey respondents give higher life satisfaction ratings when they are contacted on sunny than on gloomy days. However, when the weather was mentioned explicitly by the interviewer just before the life satisfaction question was asked, life satisfaction ratings were no longer affected by ambient gloom. Schwarz and Clore interpret this finding in terms of their 'feelings-as-information' model, according to which affect experienced while attending to an affectively ambiguous stimulus is, by default, assumed to convey information about the stimulus – unless an alternative, and more plausible, attribution for the affect is brought to the subject's attention. Life satisfaction is a particularly nebulous concept, which takes on different forms when viewed from different perspectives, and hence can accommodate radically discrepant interpretations under different conditions. According to the model, when the lousy weather outside is explicitly highlighted before respondents are asked about life satisfaction, they attribute the lousy feelings they are currently experiencing to the weather and hence not to the disappointments that taint their life satisfaction. Note that, in this model, the subjective feeling of affect is itself assumed to be information-bearing: it is not taken to be a mere inert correlate of some other underlying computation. But the information the feeling carries is not obvious, and it is subject to systematic misconstrual under suitable experimental conditions.

However, such misconstrual should be viewed as the exception rather than the rule. According to the feelings-as-information model, after all, feelings are informative. The use of one's present affective state for default attribution is a 'heuristic' (similar to others studied in research on judgment and decision making), a rough but useful rule-of-thumb that generates reasonable judgments under typical conditions. For instance, if you feel vaguely uneasy when you enter a new apartment, you might be better off, on average, renting a different one. Because conscious deliberation is slow, effortful, and often relatively insensitive to fine-grained nuance, deferring by default to less informationally transparent but faster and more sensitive affective reactions may often, many researchers believe, be a wise general policy. Indeed, some evidence from experimental studies of neurological patients suggests that impairments in affective processing may lead to systematically poorer choices in 'rational' choice domains like decision making under risk.

It is important to note that a given situation potentially raises multiple problems of affective attribution at multiple levels of processing, and the extent to which their solutions overlap is a nontrivial question. The attribution 'implicit' in a classical conditioning experiment – where learning systems may (as John Garcia demonstrated) 'blame' recently ingested food, rather than a red flashing light, for presently experienced stomach discomfort – need not dovetail, in process or output, with the attribution a subject makes 'explicit' in filling out a questionnaire. To more fully capture the complexity of typical real-world attribution problems, imagine an experiment (1) involving a subtle unpleasant mood manipulation, (2) in which multiple affectively ambiguous items are simultaneously exhibited to the subject, and (3) in which multiple probes – some involving explicit measures like ratings of attractiveness, others involving implicit measures like skin conductance response (SCR) or heart rate – are employed to assess affective reactions to each of the stimuli presented. Affective reactions and attributions may or may not run in parallel across the different levels of processing revealed by different kinds of probe. Even if the subject consciously decides upon a single primary source of current mood – this painting, perhaps, and not that sculpture, is the main culprit – we cannot assume that implicit measures of affect like SCR will draw the same distinctions – with, for example, a raised SCR for the painting, subsequently viewed, but not for the sculpture. The ways in which processes of considered affective attribution may interact, or fail to interact, with less consciously accessible and controlled implicit processing of affective stimuli is an important question in research on attribution. At present, researchers are actively contesting the role of conscious awareness in affective conditioning – in which an initially neutral

'conditioned stimulus' inherits the affective properties of an attractive or aversive 'unconditioned' stimulus with which it is paired. While there is substantial diversity of opinion, some researchers believe that effective conditioning of affect requires consciousness of the conditioned–unconditioned pairing, which would imply that 'higher-level' representations can feed down to influence seemingly uncontrolled responses.

A related limitation of the preceding discussion is that it artificially separates the process of causal attribution from the causal process for which the attribution is being made. That is, we have presupposed the existence of a separate and prior causal relation between external stimulus and internal affect; holding this relation fixed, we then ask whether the subject's judgment of casual attribution corresponds with this actual relation. In this way, the subject's causal judgments about his or her affective states are taken to be just as isolated from those affective states as they would be if the subject were instead making causal judgments about the origins of another person's affective states. This idealized division and comparison of the causal judgment and the causal process judged is often useful. However, when the affective state which the subject contemplates is ongoing, it is a misleading simplification. Because our various mental states intimately and dynamically impinge on and interact with each other, beliefs about our own states can readily become self-fulfilling or self-defeating. Explicit affective attributions fall within this class of problematic beliefs. An initial affective attribution may proceed to make itself true or make itself false: now that I have attributed my present gloominess (set initially into motion by the overcast sky overhead) to this odd black sweater, the sweater itself may begin to make me feel gloomy. This may now feed back into and reinforce or otherwise modify the process of attribution, rendering it partly accurate. Furthermore, in addition to the outcome of an explicit attribution process, the mere process of stepping back to explicitly think about one's affective states may have significant effects on those states.

It is important not to overstate this important point. There are, in particular, striking and well-documented parallels between the ways people respond to cues about their own affective states and the ways people respond to cues about the affective states of others. The self-fulfilling, -defeating, and otherwise-altering effects of explicit affective judgment should not be viewed as a chaos that completely engulfs and disfigures the normal process of attributive inference. Nonetheless, the process and outcome of affective judgment can have significant effects on the affective states being judged.

Jonathan Schooler has argued for a tripartite distinction between the unconscious, the conscious, and the 'metaconscious.' This distinction derives from the claim that one can have a subjective experience without knowing that one is having it. Awareness that one is in an experiential state is then said to render that state (transiently) 'metaconscious.' On this view, the transition from consciously being in state X to being in a state where one also metaconsciously knows that one is in state X requires additional computational steps that are only intermittently executed and are potentially fallible. It is important to appreciate that 'metaconsciousness' may come in many varieties and can vary in scope and intensity – it is not an all-or-nothing proposition. There is a distinction, while experiencing a red image, between (1) having an inarticulate awareness of one's experience; (2) privately articulating to oneself (in English) "Here, I see red"; and (3) explicitly entertaining a sophisticated causal account of the experience (e.g., 'this present experience is the product of light of a certain wavelength reflecting off a surface with certain properties, impinging my retinas, and triggering a specific cascade of axon potentials in my brain, associated with this experience'). While certain problems of interpretation may arise at the far extreme – where dim metaconsciousness needs to be differentiated from none – the concept is useful over a broad range, where metaconscious contents (thoughts referring to current or past conscious states) can patently vary in articulateness, sophistication, and veridicality.

As noted earlier, explicit metaconscious attributions of affective content ("My present annoyance is at the smug expression on your face") can feed back and modify both the affective state and the process of affective attribution that explains it. In addition, Schooler and his colleagues have reported some evidence suggesting that merely engaging in

metaconscious processing can alter the affective state itself. In one study, their subjects listened to Stravinsky's *The Rite of Spring* – a striking and jarring piece of music, which the researchers assumed would be affectively ambiguous. Subjects who were asked to continuously rate their happiness while listening to the piece subsequently reported being less happy than other subjects. The authors speculated that focused hedonic monitoring may limit the attention devoted to the experience itself, and may diminish sensitivity to subtle and nuanced aspects of the experience which resist clear articulation (just as reflecting on elusive qualities of a fine wine may disrupt our perception and memory of those very qualities). While further investigation is needed in this area, this and other evidence – for example, a number of studies showing that mood inductions have altered effects in the presence of a mirror, which presumably encourages the subject to more closely monitor his or her own overt reactions – indicates that the intensity and direction of affective metaconsciousness can systematically alter conscious affective states.

Summary

This article has described multiple dissociations between consciousness and emotion. Stimuli of which we are unaware can elicit affective states of which we are aware. In some cases, the elicited affective states, while systematically modifying our physiology and behavior, may fail to reach full awareness. Furthermore, even when both the eliciting stimulus and the emotion are conscious, we may be unaware of the relationship between them. Yet, while affect can in these ways be multiply dissociated from awareness, affective reaction and conscious awareness do not occupy distinct and hermetically sealed compartments, our conscious selves looking on ineffectually as mere observers of our approaches and avoidances, only able to guess what and why. Rather, our deliberate allocation of attention can strongly modulate the affective effects of valenced stimuli, and the process and products of metaconscious reflection about affective experience can partly reshape that experience.

Or, returning to our original question: How do you know how you feel right now. . .and do you know? The research described here shows that these questions are not just perversely skeptical. You may not always know how you feel or what causes you to feel that way. And your knowledge of how you feel derives at least in part from general-purpose patterns of reasonable but fallible inference from imperfectly informative cues. Nonetheless, what you think you know with regard to how you feel, about what, and why, can exert a profound influence on how you feel, about what, and why.

Suggested Readings

Barrett LF, Niedenthal P, and Winkielman P (eds.) (2005) Emotion and Consciousness. New York: Guilford.

Berridge KC (2003) Pleasures of the brain. Brain and Cognition 52: 106–128.

Clore GL and Huntsinger JR (2007) How emotions inform judgment and regulate thought. Trends in Cognitive Sciences 11: 393–399.

Kihlstrom JF, Mulvaney S, Tobias BA, and Tobis IP (2000) The emotional unconscious. In: Eich E (ed.) Cognition & Emotion. New York: Oxford University Press.

Marshall JC and Halligan PW (1988) Blindsight and insight in visuo-spatial neglect. Nature 336: 766–767.

Murphy ST and Zajonc RB (1993) Affect, cognition, and awareness: Affective priming with optimal and suboptimal stimulus exposures. Journal of Personality and Social Psychology 64: 723–739.

Ohman A, Flykt A, and Lundqvist D (2000) Unconscious emotion: Evolutionary perspectives, psychophysiological data and neuropsychological mechanisms. In: Lane RD, Nadel L, and Ahern G (eds.) Cognitive Neuroscience of Emotion, pp. 296–327. New York: Oxford University Press.

Pegna AJ, Khateb A, Lazeyras F, and Seghier ML (2005) Discriminating emotional faces without primary visual cortices involves the right amygdala. Nature Neuroscience 8: 24–25.

Pessoa L (2005) To what extent are emotional visual stimuli processed without attention and awareness? Current Opinion in Neurobiology 15: 188–196.

Schooler JW (2002) Re-representing consciousness: Dissociations between experience and meta-consciousness. Trends in Cognitive Sciences 6: 339–344.

Schooler JW, Ariely D, and Loewenstein G (2003) The pursuit and assessment of happiness can be self-defeating. In: Brocas I and Carrillo JD (eds.) The Psychology of Economic Decisions, vol. 1, pp. 41–70. Oxford: Oxford University Press.

Vuilleumier P (2002) Facial expression and selective attention. Current Opinion in Psychiatry 15: 291–300.

Winkielman P and Berridge KC (2004) Unconscious emotion. Current Directions in Psychological Science 13: 120–123.

Winkielman P, Berridge KC, and Wilbarger JL (2005) Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. Personality and Social Psychology Bulletin 1: 121–135.

Zajonc RB (2000) Feeling and thinking: Closing the debate over the independence of affect. In: Forgas JP (ed.) Feeling and Thinking: The Role of Affect in Social Cognition, pp. 31–58. New York: Cambridge University Press.

Biographical Sketch

Shlomi Sher is a postdoctoral scholar in psychology at the University of California, San Diego. He received his BA in mathematics from Harvard University and his PhD in psychology from Princeton University. He is interested in conceptual and experimental problems in the study of consciousness, perceptual attention, rationality, and moral judgment.

Piotr Winkielman is a professor of psychology at the University of California, San Diego. After undergraduate study at the University of Warsaw and the University of Bielefeld, he received his PhD in social psychology from the University of Michigan and postdoctoral training in social neuroscience at the Ohio State University. His current research explores the relation between emotion, cognition, body, and consciousness using psychological and psychophysiological approaches. His research has been supported by the National Science Foundation and National Alliance for Autism Research and he has served as an associate editor of *Psychological Review* and *Emotion*.

Ethical Implications: Pain, Coma, and Related Disorders

C Schnakers, University of Liege, Liège, Belgium
M-E Faymonville, Centre Hospitalier Universitaire Sart Tilman, Liège, Belgium
S Laureys, University of Liege, Liège, Belgium

Glossary

Analgesic – Any member of the diverse group of drugs (colloquially known as painkillers) used to relieve pain (achieve analgesia). Analgesic drugs act in various ways on the peripheral and central nervous systems; they include paracetamol (acetaminophen), the nonsteroidal anti-inflammatory drugs (NSAIDs) such as the salicylates, narcotic drugs such as morphine, synthetic drugs with narcotic properties such as tramadol, and various others.

Functional neuroimaging – The use of neuroimaging technology to measure an aspect of brain function, often with a view to understanding the relationship between activity in certain brain areas and specific mental functions.

Hydranencephaly – A type of cephalic disorder. This is a rare condition in which the cerebral hemispheres are absent and replaced by sacs filled with cerebrospinal fluid.

Opiates – Named so because they are constituents or derivatives of constituents found in opium. The major biologically active opiates found in opium are morphine, codeine, thebaine, and papaverine.

Pain – Defined as an unpleasant sensory and emotional experience associated with real or potential tissue damage.

Primary somatosensory area – In the human cortex it is located in the postcentral gyrus (parietal lobe). It is the location of the primary somatosensory cortex, the main sensory receptive area for the sense of touch.

Sedative – A substance that depresses the central nervous system, resulting in calmness, relaxation, reduction of anxiety, sleepiness, and slowed breathing. At high doses or when they are abused, many of these drugs can cause unconsciousness (see hypnotic) and even death.

Thalamus – It constitutes the main part of the diencephalon. The thalamus is believed to both process and relay sensory information selectively to various parts of the cerebral cortex, as one thalamic point may reach one or several regions in the cortex.

Introduction

"Pain is defined as an unpleasant sensory and emotional experience associated with real or potential tissue damage." As this definition of the International Association of Pain Specialists points it out, pain is a subjective first-person experience. Pain assessment is directly based on the patient's verbal report. Pain scales were developed to be used by the patient himself as, for instance, the Visual Analogical Scale (VAS), one of the most used scales which allows to assess the intensity of pain in communicative patients. This method is inadequate for patients who cannot functionally communicate (verbally or nonverbally). Detecting and treating signs of pain represents an important medical and ethical stake especially in patients who are not communicative, as it is the case for many patients recovering from coma. Progress in acute neurocritical care has led to an increase in the number of patients surviving severe brain injury. Whereas some recover quickly, others take more time and pass through different states of unconsciousness (i.e., coma, vegetative state) before partially (i.e., minimally conscious state) or fully recovering awareness. In patients with an altered state of consciousness, it is necessary to

use indirect means of assessments such as behavioral observation or physiological measurements. Responses such as facial expression, movements of the limbs, vocalizations and modification of heart rate or breathing is often considered to detect the presence of a painful experience. Numerous standardized pain scales were developed. The aim of this article is to (1) review the remnant brain activity evoked by noxious stimuli in altered states of consciousness, particularly in vegetative (VS) and minimally conscious (MCS) states; (2) discuss the problems encountered in pain assessment in these noncommunicative patients; and (3) review the ethical concerns as regards to their potential feeling of pain.

Pain Assessment in Noncommunicative Patients

Severely brain-injured patients are unable to communicate their feelings and possible pain experiences. Numerous pain scales were developed for assessing noncommunicative subjects such as newborns or demented elderly. However, few of these scales were properly validated. The most validated scales are the Neonatal Infant Pain Scale (NIPS) and the Faces, Legs, Activity, Cry, Consolability (FLACC) Pain Assessment Tool used for assessing pain in newborns (see Table 1). The Pain Assessment in Advanced Dementia Scale (PAINAD) and the Checklist of Nonverbal Pain Indicators (CNPI) are used for assessing pain in the demented elderly (see Table 1). These pain scales mainly include the observation of grimaces, cries, negative verbalizations, body movements, changes in breathing patterns, and consolability.

Some of these clinical parameters are observed during the behavioral assessment of patients in altered states of consciousness (see Table 2). The behavioral observation remains the gold standard to detect conscious perception in response to various stimuli such as auditory, visual, tactile, as well as noxious. Considering noxious stimulation, three types of motor responses to pain are usually considered: stereotypical responses (i.e., slow generalized flexion or extension of the upper and lower extremities), flexion withdrawal (i.e., the limb moves away from the point of stimulation), and localization responses (i.e., the nonstimulated limb must locate and make contact with the stimulated body part at the point of stimulation). These responses are respectively linked, based on current understanding and theory, to brainstem, subcortical or cortical activity, respectively. Localization response to pain is the only motor response considered indicative of conscious perception. Clinically, these behaviors are studied by applying pressure to the fingernail, to the temporomandibular joint, the supraorbital nerve, or to the ear. For instance, the Glasgow Coma Scale, which is the most used coma scale in the world, uses the pressure to the fingernail. The literature nevertheless suggests that pressure of the finger nail bed with a pencil as was first proposed by Teasdale and Jennett (1974) falsely lowers the level of responsiveness. No study has assessed which stimulation is the more efficient to elicit localization of pain. Determining which stimulation is the most powerful to detect signs of conscious perception is a real challenge. Indeed, previous studies showed a high rate of misdiagnosis (37%–43%) among patients diagnosed as being in VS underlying the difficulty to detect signs of consciousness. Moreover, various rates of misdiagnosis were observed according to the scale used for the behavioral assessment. As these scales used different noxious stimulations,

Table 1 Selection of behavioral scales which assess acute pain in noncommunicative patients

Population	Behavioral scale
Infants	NIPS: Neonatal Infant Pain Scale
	FLACC: Faces, Legs, Activity, Cry, Consolability Observational Tool
	PIPP: Premature Infant Pain Profile
	CRIES
	CHEOPS: Children's Hospital of Eastern Ontario Pain Scale
Demented elderly	PAINAD: Pain Assessment In Advanced Dementia
	CNPI: Checklist of Nonverbal Pain Indicators
	DOLOPLUS 2
	ADD: The Assessment of Discomfort in Dementia Protocol
	PACSLAC : Pain Assessment Checklist for Seniors with Limited Ability to Communicate

Adapted from Schnakers C and Zaster ND (2007) Pain assessment and management in disorders of consciousness. Current Opinion Neurol, 20, 620–626.

Table 2 Behavioral coma scales with assessment of pain

Scale name	Clinical assessment
Glasgow Coma Scale	Encompasses three components: eye (E), verbal (V), and motor (M) response to external stimuli. Procedure: As regards to motor responses, pressure is applied to the fingernail bed with a pencil resulting in either flexion or extension at the elbow. If flexion is observed, stimulation is then applied to the head and neck and to the trunk to test for localization. Number of stimulations: Not explicitly mentioned. Quotation: Best response observed ranging from M1 to M6.
Coma Recovery Scale – Revised	Includes auditory, visual, motor, verbal, communication, and arousal subscales. Procedure: A deep pressure applied to the nail beds of each extremity. Number of stimulations: Two times each side. Quotation: considered as present if behavior is observed two out of four times.
Full Outline of Unresponsiveness	Consists of four components (eye, motor, brainstem, and respiration) with a maximal score of four. The motor component combines decorticate and withdrawal flexion responses. Procedure: The painful stimulus is applied to the temporomandibular joint or supraorbital nerve. Number of stimulations: Not explicitly mentioned. Quotation: Best response observed.
Coma/Near-Coma Scale	Assesses responses to auditory, visual, verbal, tactile, olfactory, threat, and noxious stimulation. Procedure: Two stimuli are used to assess response to pain: (1) firm pinch on finger tip; (2) robust ear pinch/pull. Number of stimulations: Three times each side. Quotation: Score of 0–4 according to the number and the type of response observed.

Adapted from Schnakers C and Zaster ND (2007) Pain assessment and management in disorders of consciousness. Current Opinion Neurol, 20, 620–626.

further studies could assess which of these simulations is the more interesting in the detection of conscious perception and, therefore, in the detection of pain.

Some coma scales also study grimaces but none in response to noxious stimulation. Even if grimacing is considered as a pain indicator, for instance, in pain assessment scales employed in demented elderly, the Multi Society Task Force on PVS did not consider it as a necessary sign of conscious perception. Patients showing no sign of consciousness except grimaces to noxious stimuli can therefore be diagnosed as being in VS. However, clinical data considering the proportion of VS patients only showing this behavior are warranted. Moreover, until now, no functional neuroimaging study has investigated the neural processing of pain in these patients as previous studies did not involve patients presenting grimaces in response to pain. Additional research is needed to better understand the brain processing underlying this apparent indicator of painful experience. Similarly to grimaces, other parameters such as vocalization and verbalization or, less often, changes in breathing rate are part of some behavioral consciousness scales but never in response to pain. Finally, the consolability (i.e., number of tactile or auditory stimulations needed to reassure the stimulated patient) is not considered in any known coma scale.

To summarize, existing coma scales do not specifically assess possible pain perception in noncommunicative patients recovering from coma. This is why we have recently developed a scale to assess pain in severely brain-injured patients, the Coma Pain Scale (CPS). This scale consists in the observation of motor, verbal, and visual responses, facial expression, and pain anticipation. In a previous version, the CPS also included the subscale 'breathing,' which was excluded considering the difficulty in assessing this parameter without appropriate monitoring devices. Each subscale of the CPS is scored from 0 to 3 according to the response complexity (the total score is 15). A pilot study was performed in 24 severely brain-injured patients (63 ▫ 14 years, 15 men, 10 traumatic, 8 chronic) diagnosed as VS (n ¼ 11) or MCS (n ¼ 13). The results showed a good correlation between the CPS and other validated pain scales such as the PAINAD, the CNPI, the NIPS, and the FLACC, suggesting that, in parallel to other scales, the CPS assessed pain. However, on the contrary to these pain scales, the CPS scores were significantly

different according to clinical entity (i.e., VS and MCS), suggesting that the CPS is better adapted for the assessment of pain in patients recovering from coma. Finally, a good interrater agreement was observed. The CPS seems therefore to be a promising tool for assessing pain in severely brain-injured patients in altered states of consciousness. Further investigations are undergoing and the obtained results will need to be compared with functional neuroimaging data.

Pain Processing in Coma and Related Disorders

It is known that pain is mediated by a widely distributed cerebral network. The neural correlates of pain involve the lateral and medial pain systems. The lateral pain system includes the lateral thalamus, primary and secondary somatosensory cortex (SI and SII), parietal operculum, and insula. The medial pain system involves the medial thalamus, anterior cingulated cortex, amygdala, hippocampus, hypothalamus, locus coeruleus, and periaqueductal gray matter.

In fact, the emergence of pain perception is composed by sensory–discriminative, cognitive–evaluative, and motivational–affective central systems. Indeed, the thalamus (which participates to the increase of arousal following a noxious stimulation) and midbrain (more exactly, periaqueductal matter) are thought to be involved in the modulation of reflex responses to pain stimulus. Primary and secondary somatosensory cortex participate to the sensory–discriminative aspects of pain processing, whereas cingulate, insula, orbitofrontal, and medial prefrontal cortices are considered to be involved in the affective aspect of pain processing. Moreover, interconnectivity between the periaqueductal matter and orbitofrontal cortex may be key to cognitive–emotional responses associated with pain. Additionally, a recent study of Boly *et al.* showed that activity in the anterior cingulate cortex and insula just before pain stimulation can increase pain perception.

Recently, residual central pain processing existing in altered states of consciousness such as VS and MCS was investigated by the use of functional neuroimaging. Laureys *et al.* compared cerebral activation to high-intensity noxious electrical stimulation of the median nerve at the wrist in 15 VS patients (12 nontraumatic, mean time postinsult was 1 month) and 15 healthy volunteers. Noxious stimulation activated contralateral thalamus, midbrain, and primary somatosensory cortex in every vegetative patient, possibly suggesting a partially preserved sensory–discriminative pain processing. Kassubek *et al.*, who used a similar methodology in 7 VS patients (all anoxic, mean time postinsult was 1.5 years), confirmed the activation in primary somatosensory cortex but also found an activation in secondary somatosensory, insular, and anterior cingulate cortices, which is considered critical in the affective and cognitive processing of pain. However, brain connectivity studies conducted by Laureys *et al.* showed that primary somatosensory cortex was functionally disconnected from secondary somatosensory, bilateral posterior parietal, premotor, polysensory superior temporal, prefrontal cortices, as well as from anterior cingulate cortex. The observed primary cortex activation is therefore suggested to be isolated from higher-order associative cortical activity considered crucial in the conscious perception of the stimuli as well as from areas involved in the affective and cognitive pain processing. Finally, in brain death, noxious stimuli do not lead to any neural activation whatsoever (see Figure 1).

Figure 1 Painful stimuli activate a widespread network of cortical areas encompassing the anterior cingulate cortex considered to be involved in the affective component of pain perception (arrow). Patients in a vegetative state not only show subcortical activation (i.e., brainstem and thalamus) but also of primary somatosensory cortex (circle). However, this area is disconnected from the rest of the gray matter and is hence considered to be insufficient to lead to conscious perception of pain. In brain death, noxious stimuli do not lead to any neural activation whatsoever. Adapted by permission from Macmillan Publishers Ltd: [Nature Reviews Neuroscience] (Laureys et al., 2005), copyright (2005).

The cerebral activation to pain is different in MCS patients. Boly et al. showed brain activation similar to controls in response to noxious stimuli in five patients in a MCS. This activation involved the anterior cingulate area, which suggests that the patients could perceive the unpleasant aspect of painful stimulation. Even if other studies are needed to confirm these results, this study suggests a sufficient cortical integration for conscious perception and hence a conscious pain experience in MCS patients. Further studies will need to investigate the level of pain perception in these patients. Indeed, many cognitive components such as long-term memory (particularly, the ability to remember previous pain experiences) play a role in the experience of pain. Few studies have investigated the residual cognitive functioning in altered states of consciousness. Recently, Owen and coworkers showed a brain activation similar to controls in a severely brain-injured patient who was instructed to imagine herself playing tennis or going for a walk in her home. As semantic as well as autobiographical information encoded in long-term memory were needed to perform this task, this result suggests that some high-level cognitive treatment could be preserved in patients recovering from coma, even in the presence of low behavioral levels (this patient exclusively showed brief visual fixation).

Ethical Considerations

As discussed, neuroimaging data seem to indicate a brain activation to pain in MCS similar to controls, involving the anterior cingulate cortex. The data suggest that these MCS patients could perceive pain. On the contrary, VS patients showed a functionally disconnected brain activity, suggesting the absence of an integrated pain perception. Considering these results, adequate analgesic treatment has to be provided in MCS patients. The issue is much more complicated in VS patients. Given the high rate of misdiagnosis (37%–42%), if we decide not to administer analgesic treatment in the presence of a potential painful experience (e.g., contractures or fractures), there is a real probability for not treating a patient erroneously diagnosed and, hence, for not treating a patient who perceives pain. As regards to the ethical principles of beneficence and nonmaleficence, clinicians want to be certain that an individual is not suffering when making clinical decisions about treatment or the end of treatment. Therefore, medical staff has to provide pain treatment and comfort to all patients, even noncommunicative patients diagnosed as being in a VS.

In a medicolegal context, the question may be different as the patient has to be categorized as perceiving or not perceiving pain. However, as pain is a subjective first-person experience mediated in part by beliefs or emotions, we cannot be sure whether patients in an altered state of consciousness perceive pain or not. In our view, considering the current levels of clinical and scientific uncertainty, pain treatment should be considered in all patients in a VS or MCS. However, current clinical guidelines do not share this view and do not propose the use of analgesics in VS. For instance, Terry Schiavo died from dehydration without administration of opiates as she was diagnosed as VS by the High Court's experts.

On the contrary, a systematic use of analgesics in VS could have undesirable sedative effects leading to an underestimation of the state of consciousness. Under-use of analgesics could however also lead to an underestimation of consciousness. Indeed, the presence of intense pain may diminish already trifling cognitive and motor abilities and could, then, lead to diagnostic error. Adequately assessing and monitoring pain and pain therapy hence represents a real clinical challenge.

In our view, much more research is needed in order to propose evidence-based guidelines. But such researches represent major ethical challenges. For some scientists noxious stimuli cannot be applied to patients unable to give informed consent. Monitoring pain in severely brain damaged patients represents such an important humane, affective, and social problem that it warrants further study to better understand the underlying cerebral dysfunction of VS and MCS. In fact, to exclude investigations of residual perception of pain in these patients would be ethically unwarrantable. We propose an ethical framework balancing on patients' protection and inclusion in research protocols and medical advances.

Conclusions

VS and MCS patients can, by definition, not communicate their feelings and possible pain perception. Behavioral coma scales developed for assessing the consciousness level of severely brain-injured patients integrate some parameters used for detecting pain perception in noncommunicative patients. These scales are nevertheless not sufficient for specifically assessing pain in VS or MCS patients. Moreover, some of the studied parameters such as physiological changes are not sufficient to discern a conscious painful experience. Indeed, studies in general anesthesia showed that autonomic measurements (i.e., heart rate, respiratory frequency, blood pressure, pupillary diameter, and skin conductance) are not reliable indicators of pain. Future studies should hence focus on methodologies for adapted pain assessment relevant to this patient population. A standardized and validated behavioral pain scale for altered states of consciousness will allow (1) at a scientific level, to better specify the behavioral pattern of VS patients (e.g., prevalence of grimaces) and to determine the possible prognostic value of these behaviors and (2) at a clinical level, to monitor pain treatment in order to avoid sedative effects as well as under-uses of analgesics.

As regards to neuroimaging, on the contrary to MCS, VS does not suggest an integrated cortical pain processing. However, there are still some debates regarding whether conscious perception of pain may be mediated by subcortical areas. Indeed, a subcortical system comprising the basal ganglia, medial and midline thalamic nuclei, substantia nigra, ventral tegmental area, superior colliculi, midbrain, and pontine reticular formation has been proposed by Merker and coworkers as sufficient to mediate the organization of consciousness. Consistent with this theory, the responses to noxious stimulation of children with hydranencephaly seem sometimes purposeful and similar to those of intact children. Preterm neonates or adolescents with cortical parenchymal injury mount biobehavioral responses to pain that are indistinguishable from those of normal controls. This suggests that the mechanisms of conscious sensory perception are not entirely dependent on cortical activity. However, most neuroscientific studies point to a key role of cortico-cortical and thalamo-cortical interaction in the emergence of conscious experiences. This interaction seems to be absent in VS patients. Therefore, the question of whether pain perception and suffering are present in patients with an altered state of consciousness has certainly to be further investigated in noncommunicative VS and MCS patients.

Finally, in our view, as regards to our current scientific knowledge on brain processing in altered states of consciousness, the possibility of pain perception should be considered in all patients recovering from coma. In the future, researches integrating behavioral and neuroimaging data will be warranted to establish clear guidelines for treating pain in these patients and therefore to increase the quality of life or of the end of life in this challenging noncommunicative population.

Acknowledgments

This research was funded by the Belgian National Funds for Scientific Research (FNRS), European Commission, James McDonnell Foundation, Mind Science Foundation, French Speaking Community Concerted Research Action, Fondation Médicale Reine Elisabeth, and University of Liège.

See also: Animal Consciousness.

Suggested Readings

Anand KJS (2006) Pain. Clinical updates. International Association for the Study of Pain 4(2): 1–4.

Boly M, Faymonville ME, Peigneux P, et al. (2005) Cerebral processing of auditory and noxious stimuli in severely brain injured patients: Differences between VS and MCS. Neuropsychological Rehabilitation 15(3–4): 283–289.

Boly M, Balteau E, Schnakers C, et al. (2007) Baseline brain activity fluctuations predict somatosensory perception in humans. Proceedings of the National Academy of Science 104: 12187–12192.

Boly M, Faymonville ME, Schnakers C, et al. (2008) Perception of pain in the minimally conscious state with PET activation: An observational study. Lancet Neurology 7: 1013–1020.

Chatelle C, Vanhaudenhuyse A, Mergam N, et al. (2007) Mesurer la douleur chez le patient non communicant. Revue MÉdicale de Liège 62(4): 1–9.

Demertzi A, Vanhaudenhuyse A, Bruno MA, et al. (2008) Is there anybody in there? Detecting awareness in disorders

of consciousness. Expert Reviews of Neurotherapeutics 8: 1719–1730.
Faymonville ME, Laureys S, Degueldre C, et al. (2000) Neural Mechanisms of Antinociceptive Effects of Hypnosis. Anesthesiology 92(5): 1257–1267.
Fins JJ, Illes J, Bernat JL, et al. (2008) Neuroimaging and disorders of consciousness: Envisioning an ethical research agenda. American Journal of Bioethics-Neuroscience 8: 3–12.
Huskisson EC (1982) Measurement of pain. Journal of Rheumatology 9: 768–769.
IASP (1994) Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms. Task force on taxonomy. Seattle, WA: IASP Press.
Kassubek J, Juengling FD, Els T, et al. (2003) Activation of a residual cortical network during painful stimulation in long-term postanoxic vegetative state: A 15O-H2O PET study. Journal of the Neurological Science 212: 85–91.
Kupers R, Faymonville ME, and Laureys S (2005) The cognitive modulation of pain: Hypnosis- and placebo-induced analgesia. Progress in Brain Research 150: 251–269.
Kupers R and Kehlet H (2006) Brain imaging of clinical pain states: A critical review and strategies for future studies. Lancet Neurology 5: 1033–1044.
Laureys S and Boly M (2008) The changing spectrum of coma. Nature Clinical Practice Neurology 4: 544–546.
Laureys S, Faymonville ME, Peigneux P, et al. (2002) Cortical processing of noxious somatosensory stimuli in the persistent vegetative state. Neuroimage 17: 732–741.
Merker B (2007) Consciousness without a cerebral cortex: A challenge for neuroscience and medicine. Behavioral and Brain Sciences 30: 63–81.
Schnakers C, Giacino J, Kalmar K, et al. (2006) Does the FOUR correctly diagnose the vegetative and minimally conscious states? Annals of Neurology 60(6): 744–745.
Schnakers C and Zasler ND (2007) Pain assessment and management in disorders of consciousness. Current Opinion Neurol 20: 620–626.
Zasler ND, Horn L, Martelli MF, and Nicholson K (2007) Post-traumatic pain disorders: Medical assessment and management. In: Zasler N, Katz D, and Zafonte R (eds.) Brain Injury Medicine: Principles and Practice. New York: Demos Publishers.

Biographical Sketch

Caroline Schnakers graduated as a neuropsychologist from the University of Liège in 2003. Dr Schnakers joined the Coma Science Group in 2002. She has expertise in the behavioral assessment of consciousness levels in severely brain-injured patients. She validated the French version of the Coma Recovery Scale – Revised and employs in electrophysiological methods (more exactly, electroencephalogram and cognitive evoked potentials) investigating its potential in the detection of early signs of consciousness in coma survivors. She received her doctorate in psychological sciences in 2008 and is involved in the Belgian federal network for the care of VS and MCS patients. She has recently developed a freely available DVD for educational purposes on the diagnosis and assessment of chronic disorders of consciousness. Dr Schnakers is supported by the 'Fonds Léon Fredericq,' The European Commission and the Belgian National Funds for Scientific Research (FNRS).

Marie-Elisabeth Faymonville graduated as a medical doctor medicine from the University of Liège (UL) in 1977. She specialized in anesthesia-intensive care in 1981 and obtained in 1983, a PhD in clinical sciences. She developed in 1992 a new technique in anesthesia: hypnosedation. Her scientific approach allowed to promote hypnosis as a clinical tool, which is particularly interesting in modern medicine (especially, in chronic pain and in palliative care). Her research activity is focused on the investigation of the neuroanatomic mechanisms of different states of consciousness, including hypnosis. Since 2004, she has been leading the Pain Clinic and the Palliative Care Unit of the University Hospital of Liège.

Steven Laureys is a senior research associate at the Belgian National Fund of Scientific Research (FNRS) and Clinical Professor head of clinics at the Department of Neurology of the Liège University Hospital. In 1993, he graduated as a medical doctor from the Vrije Universiteit Brussels, Belgium. While specializing in neurology he took up a research career and obtained his MSc in pharmaceutical medicine, working on pain and stroke using *in vivo* microdialysis and diffusion MRI in the rat (1997). Drawn by functional neuroimaging, he moved to the Cyclotron Research Center at the University of Liège, Belgium, where he obtained his PhD (2000) by studying residual brain function in coma, vegetative, minimally conscious, and locked-in states. He is board certified in neurology (1998) and in palliative and end-of-life medicine (2004). He is a recipient of the William James Prize (2004) from the Association for the Scientific Study of Consciousness (ASSC) and the Cognitive Neuroscience Society (CNS) Young Investigator Award (2007). He recently published *The Boundaries of Consciousness* (Elsevier, 2005) and The Neurology of Consciousness (Academic Press 2008). He nowadays leads the Coma Science Group at the Cyclotron Research Centre at the University of Liège, Belgium.

Folk Theories of Consciousness

B F Malle, Brown University, Providence, RI, USA

Glossary

Folk theory of mind – Ordinary people's conceptual framework within which they interpret (human) behavior and mental states. The term 'theory of mind' is sometimes used to refer to the psychological processes by which this interpretation occurs (e.g., empathy, inference, simulation).

Reason explanations – The primary mode of ordinary people's explanations of intentional behavior, in which they try to cite the agent's own reasons (typically beliefs and desires) for which he or she acted.

Causal history of reason explanation – The second mode of ordinary people's explanations of intentional behavior, in which they try to cite factors that had a causal influence on the actor's reasons (e.g., culture, personality, or unconscious states) but are not themselves the reasons for which the person acted.

Privileged access – The hypothesis that humans have a unique way of registering and learning about their own mental states (typically introspection) that is not available to other people.

Incorrigibility – The hypothesis that humans cannot be wrong in their awareness of or beliefs about some of their own mental states.

Phenomenal consciousness – Mental states that have an experiential quality, a particular feeling of having or being in them (e.g., emotions, pain, sensations such as smell).

Why Folk Theories?

Some natural phenomena can be safely assumed to exist independently of human observers, and the science of such phenomena need not concern itself with lay perceptions and beliefs. When we study gold, penguins, and diabetes, all that counts for science is what gold, penguins, and diabetes are really like; people's folk theories about these phenomena will be secondary, perhaps erroneous, or a pale reflection of extant scientific knowledge. The folk theories could still be important for other purposes, such as for an understanding of misperceptions – those about diabetes, for example, that stand in the way of successful treatment.

Other natural phenomena exist because of human observers, and the way people think about these phenomena partially constitutes their nature. The science of such phenomena must therefore consider the folk theories that people hold, not only as a source of inspiration but also as part of the actual object of investigation. For example, a scientific study of 'marriage' as a psychological, sociological, and political phenomenon must pay close attention to ordinary people's concept of marriage. Likewise, a scientific study of what 'intentional actions' are must be solidly grounded in people's folk theory of action. Even if a comprehensive theory of action may in the end go beyond this folk theory, to be a theory about 'action' (and not, say, about bodily movement), the scientific theory must build on the folk theory.

Consciousness belongs squarely to the second group of phenomena; in fact, it is one of its prototypical members, because the very phenomenon of consciousness entails a subjective viewpoint, a person with a mind who experiences and conceptualizes internal and external states. Folk theories of consciousness are therefore far more than a curious sideshow to the 'real' scientific investigation of consciousness. Obviously, the content of these folk theories does not exhaust what there is to know about consciousness; but the folk theories set boundaries to any scientific investigation and inform the conclusions it may reach.

Empirical research on folk theories of consciousness proper has been relatively infrequent, but we will be casting the net more widely, examining how people conceptualize minds in general and thereby gathering important insights into people's concept of consciousness in particular. Research from developmental and social psychology will blend with work in theoretical and empirical philosophy to reveal the nature, functions, and impact of folk theories of consciousness. We will see several misrepresentations of what this folk theory entails as well as boundaries and guideposts that they set for the scientific study of consciousness.

Semantic Preliminaries

To qualify for a folk 'theory' of consciousness, we will demand only that people have a consistent concept of the phenomenon that reasonably corresponds to the modern (scientific and philosophical) concept of consciousness. There need not be any elaborate 'theory' in the literal sense, but only a concept that incorporates identifying features and assumptions about the phenomenon and reveals itself in consistent practices.

Under the plausible assumption that language reflects concepts and language history reflects conceptual history, a brief look at etymology and use of the term 'conscious' provides the first step of investigation.

The primary meaning of 'conscious,' according to the Oxford English Dictionary (OED), is 'aware, know.' This meaning most directly reflects the Latin root, con-sci-ous, from 'scire,' knowing (which is obviously the root for 'science' as well). One of the earliest uses cited in the OED is the following:

> 1620 ABP. USSHER Serm. (1621) 1 Being so conscious vnto my selfe of my great weakenesse.

What is noteworthy about this meaning is that it comes with the prefix 'con-,' which means 'together.' This signifies that one 'also' knows, with others, like others, some fact of the world, and this aspect makes particular sense when we look at the most frequent object (in the OED) of this kind of consciousness: something undesirable, often possessed or created by oneself.

In a second, expanded meaning, the togetherness aspect is absent. 'Conscious' beings are aware of their own existence and, more importantly, have a unique capacity of thinking:

> 1692 BENTLEY Serm. (J), Matter hath no life nor perception, and is not conscious of its own existence.
>
> 1725 WATTS Logic I. ii. }2 Among substances some are thinking or conscious beings, or have a power of thought, such as the mind of man, God, angels.

Finally, there is an introspective meaning of 'conscious' that, according to the OED, is philosophical:

> 1690 LOCKE Hum. Und. II. i. }11 To be happy or miserable without being conscious of it, seems to me utterly inconsistent and impossible.

With these preliminaries in mind, we now introduce extant knowledge on people's folk theory of mind more generally. The concepts of representation that are embedded in this highly evolved and well-studied folk theory will reveal three prototypes of consciousness that people master, which correspond loosely to the above semantic categories of monitoring, thought, and introspection.

The Folk Theory of Mind

One of the major advances in the past few decades of psychological research has been the recognition that humans have a sophisticated and powerful conceptual, cognitive, and affective system for interpreting their conspecifics' behavior. This system is typically called a '(folk) theory of mind' (sometimes 'folk psychology' or 'common-sense psychology') and is rather sophisticated. It consists of (1) processes that filter, group, and integrate certain stimuli (e.g., biological motion, facial expressions, gestures) into core concepts or categories such as 'agent,' 'intention', 'perception,' and 'goal;' (2) conceptual relations among these categories (e.g., that 'perception' leads to 'belief'; that 'desires' and 'beliefs' can function as 'reasons' of 'intentional action'); and (3) a variety of processes that operate on these categories, including simulation, empathy, joint attention, and inference.

Some have argued that the earliest psychological processes within this system (such as affective

tuning or emotional contagion) may operate in the absence of conceptual distinctions. However, the system would not be able to respond adequately to, say, bodily contact or facial expressions of emotion without at least rudimentary categories that are capable of detecting and grouping relevant perceptual input. For example, objects that are self-propelled and behave contingently are classified into the category 'agent;' object-directed movements of such agents are classified into the category 'intentional action.' In none of these cases does the infant or toddler have a concept of 'agent' or 'goal' the way adults 'have' such concepts (are able to reflect on them, provide sketches of definitions, etc.). Whether one calls those early categorizations 'concepts' is less important than the recognition that they germinate full-blown, linguistically differentiated concepts (e.g., of distinct emotions and types of intentions such as promises).

The significance of this system cannot be overestimated. It is not present (or only in the most rudimentary forms) in other primates; its absence (e.g., in autistic and to some extent in schizophrenic individuals) poses enormous challenges for normal social interaction; and its development is tightly connected to infants' and preschoolers' affective bonds with caregivers, with cognitive and emotional self-regulation, with language development, and with moral judgment and moral decision making. Moreover, at least some of its features may be innate and would have evolved in millions of years of adaptation to the complexity of human behavior.

Perhaps the core concept of this system is that of intentional action because it closely ties behavior to mind, and the observable to the unobservable. In its final instantiation, the concept of intentionality has five components (conditions), and four of them are mental states. An action is considered intentional when the agent has a 'desire' for an outcome, a 'belief' that the action would lead to that outcome, an 'intention' to perform the action, 'awareness' of fulfilling that intention while performing the action, and reliable 'skill' to perform the action (Figure 1).

To fully master the concept of intentionality, people must make a number of subtle distinctions. For example, the concepts of 'intention,' 'desire,' and 'goal' are often used interchangeably in the psychological literature, but ordinary people clearly differentiate between them. Moreover, because belief and desire alone are not sufficient for a judgment of intentionality – an intention has to be present as well – people differentiate the action descriptions of doing something 'knowingly' (with belief, but without desire or intention), 'willingly' (with belief and desire but without specific intention), and 'intentionally' (with all three). Many legal systems fail to make these distinctions.

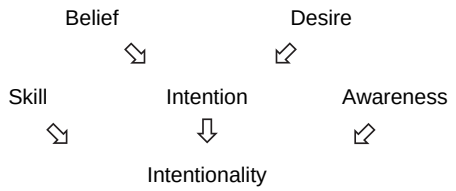


Figure 1 The folk concept of intentionality. Reproduced with permission from Malle BF and Knobe J (1997a) The folk concept of intentionality. Journal of Experimental Social Psychology 33: 101–121.

The awareness component within the folk concept of intentionality provides insight into people's understanding of consciousness. Awareness is a state of mind at the time of action that reflects what the agent thinks she is doing, whether she is intending to perform the currently performed action. This self-referential state is one prototype in people's folk theory of consciousness: monitoring one's own action and recognizing it as fulfilling a particular intention, thus comparing the representation of a current action to a previously formed representation of an action-to-be. This prototype arguably underlies the scientific notion of cybernetic models and is instantiated in simpler forms both in conceptions of nature (teleological processes) and in technology (feedback loops).

Action monitoring is not only a folk notion but has been identified as a key process in human action control. We know, however, from research by Daprati and colleagues that people have more precise conscious access to ongoing actions discrepant with their intentions than to actions conforming to their intentions. As long as the performed movement and its consequences obtain roughly as intended, little updated information about the movement execution reaches consciousness. Thus, the action is experienced as it was intended, not

exactly as it is executed. Put another way, people may not have detailed awareness of their movements (the exact physical motion pattern of their limbs and body) but are perfectly aware of their actions – what they are intending to do irrespective of how it is implemented in detail by the motor system. And this is precisely the level of analysis at which the folk concept of intentionality posits conscious awareness.

There is a second prototype of consciousness inherent in the folk concept of intentionality, and it lies in the transition from beliefs and desires to intentions – in the act of forming an intention or consciously choosing a course of action. People conceptualize intentions as the output of a conscious reasoning process to which desires and beliefs are relevant inputs (e.g., "I intend to A because I want O and A gets me to O"). This conception develops in late childhood. Before the age of 6 or 7, children hold a mixed desire-and-intention concept that is closely tied to action. After this age they reliably distinguish between desire and intention and thus between mere representations of desirable contents (that may or may not be conscious) and the reasoning-based representations of planned and chosen action (that are conscious).

This conception of reasoning toward intentions and actions is the foundation for people's explanations of intentional behavior in terms of the 'reasons for which' an agent decided to act – the relevant beliefs and desires that favored the forming of a particular intention. The study of behavior explanation can provide more detailed insight into this concept of conscious reasons.

Conscious Reasons in Explanations of Behavior

Reason explanations are the dominant mode of explaining intentional behavior. They cite one or more mental states whose content played (in the eyes of the explainer) a significant role in the agent's decision to act as she did:

(1) Anne invited Ben for dinner because she realized she hadn't seen him in a month.

Work by Malle and colleagues has shown that people make two assumptions about reason explanations: The agent must have been aware (at the time of choosing to act) of the cited reason contents (e.g., that she had not seen him in a month), and the reasons must provide rational grounds for the intention to act (i.e., not having seen him in a while rationally supports the dinner invitation if she wants to see him). We focus here on the awareness assumption, because it helps clarify the second prototype of consciousness: that of conscious choice. The force of the awareness assumption can be illustrated by recording people's responses to cases in which such awareness is denied:

(2) Anne asked Ben for change because she didn't have any quarters, but Anne wasn't aware that she didn't have any quarters.

People reject such sentences as not making sense: If not having quarters was Anne's reason for asking Ben for change, she must have been aware of not having quarters. In people's folk theory, agents must be aware of reasons because they support a conscious choice and for the choice to be conscious the agent must have actively considered those reasons and integrated them in the reasoning and choice process. Without being aware of what they believe and want, agents cannot consider these states and cannot make them 'reasons for' choosing a course of action.

Now consider the following explanation:

(3) Anne invited Ben for dinner because she is gregarious.

Here we do not have a reason explanation but a 'causal history of reason explanation.' Such statements help explain the intentional action by reference to a causal background that gave rise to Anne's reasons, even if the precise reasons are not mentioned. Because example (3) is not a reason explanation, it is not absurd to make the following claim:

(4) Anne invited Ben for dinner because she is gregarious but she wasn't actually aware of her gregariousness.

Indeed, people do not reject such denials for causal history of reason explanations, because these explanations do not carry the assumption of awareness. Causal history factors can operate outside the agent's conscious awareness – referring to

factors such as personality, culture, subtle stimulus contexts, repressed desires, and so on. Supporting this point, Malle showed that when people are faced with a list of explanations for a variety of behaviors and are asked to check whether any given explanation was a 'conscious reason' for which the agent acted, they reliably distinguish between reason explanations and causal history of reason explanations.

This research suggests that people have a differentiated conception of the role of consciousness in human action. In line with a prototype of 'conscious choice,' agents have to be consciously aware of their reasons for acting; otherwise the reasons cannot serve as the appropriate inputs to the choice of pursuing a certain action. In line with the initial prototype of consciousness as 'monitoring,' people require conscious awareness during the action phase. In order for the output of the earlier reasoning (i.e., the intention) to be appropriately executed, monitoring and readiness to adjust the action to the intention are required.

Because of what these prototypical functions of consciousness include and exclude, there is no need for people's folk theory to include a specific concept of the 'unconscious.' This does not mean that people do not have beliefs about unconscious processes – in the past hundred years, such beliefs have become part of many cultures' shared knowledge. However, when we speak of a folk theory, we do not refer to the vast sets of beliefs that people may hold; we refer to fundamental concepts and assumptions that are likely to be cross-culturally and historically stable, develop reliably, and fit in well with what we know about the evolutionary functions of the mind.

Language of Mental Activities

To examine the folk conceptions of consciousness further, we can also look at their systematic linguistic manifestations within the broad language of mental activities. The lexicon and semantics of the mind are likely to highlight consciously accessible states and processes because for an object to be picked out by language, the referent must be either publicly perceptible or at least indirectly verifiable. This is Wittgenstein's private language argument turned on its head: If there cannot be reliable terms for entirely private mental states, as he argued, then the mental state terms we do have must somehow connect to shareable aspects of the mind – those states that are both noticed on the inside and have reliable antecedents and consequences (e.g., expressions) to emerge repeatedly and thus become linguistically tractable.

The first thing to note is that the five constituents of the folk concept of intentionality – belief, desire, intention, skill, and awareness – can be reliably found across many different languages and may be universal conceptual primitives of the mind. Wierzbicka's list of such conceptual primitives does not separate intentions from desires, but other linguistic work provides evidence for the unique role of intention. Bybee has shown that, across countless and diverse languages, 'intending' and the future tense are tightly connected, something that we do not see for desire or goal concepts.

Mental words are not names of things; they are sophisticated attempts to capture the complexity of minds within a complex web of language. The vocabulary of the mind therefore gives indications about distinctions people make and broad classes they recognize. Schwanenfluegel *et al.* examined 30 verbs of perception and cognition (e.g., hear, notice, recognize, understand, plan, decide) and asked participants to rate them for their pairwise similarity. In a multidimensional scaling procedure the primary dimension was the distinction between perceptual processes (recognize, observe, discover) and conceptual and logical processes (reason, plan, figure out, estimate). D'Andrade proposed a somewhat more differentiated set of mental categories inherent in the folk model of the mind, namely, perceptions, feelings, desires, beliefs, intentions, and self-control.

A research tradition concerned with 'implicit verb causality' has highlighted a distinction that cognitive linguists have made a while ago: between 'actions' and 'experiences.' Verbs that denote either of these types perform differently in tasks of searching for and assigning causality as well as in providing explanations. However, these verb classes are in fact part of a larger 2 2 scheme that crosses public observability with intentionality (Figure 2) and appears to underlie people's descriptions and explanations of psychological

	Intentional	Unintentional
Observable	Actions	Accidental and uncontrolled behaviors
Unobservable	Choice, reasoning, intention, self-control	Experiences, perception, desires, feelings

Figure 2 Four psychological events people attend to, wonder about, and explain. From Malle BF and Knobe J (1997b) Which behaviors do people explain? A basic actor-observer asymmetry. Journal of Personality and Social Psychology 72: 288–304. ã American Psychological Association, adapted with permission.

events. Within this scheme, we find actions to be intentional while observable and experiences to be unintentional and unobservable. The two prototypes of consciousness can be grouped into the scheme as well – choice is in the unobservable intentional cell and monitoring (as awareness and perception) is in the unobservable unintentional cell. The mental categories identified earlier also fill the cell of unobservable intentional states (intentions, reasoning, self-control), and they add to the cell of unobservable unintentional states (desires, feelings, and perceptions). This last cell, which also contains feelings and other affective states, may constitute a third prototype of consciousness that is discussed next.

Phenomenal Consciousness

Earlier we introduced two prototypical forms of consciousness: The monitoring awareness of one's intentions while acting and the executive conscious choice following a reasoning process that integrates beliefs and desires in support of action. Now we turn to a potential third prototype – what philosophers have called 'phenomenal consciousness.' There is an ongoing debate over whether people really have a folk concept of phenomenal consciousness; so the current literature is very much in flux. What can we say at this point?

In the classic philosophy of mind work, we find a distinction between two main classes of mental states: propositional attitudes (e.g., beliefs, desires, intentions) and phenomenal states (e.g., pain, feeling sad, experiencing the color red, smelling fresh coffee). Propositional attitudes are understood by many to be 'computational,' that is, they could in principle be implemented in other media besides the human brain, such as in computers or aliens. By contrast, phenomenal states are understood to be uniquely implemented in the human brain – though some states may also occur in a variety of other earthly organisms.

Do ordinary people make this distinction as well? Gray, Gray and Wegner asked participants to ascribe a variety of mental states and capacities to different agents, including robots, frogs, chimps, infants, god, and adult humans. The pattern of ascriptions formed a two-dimensional space. One axis, which the authors labeled 'experiences,' was constituted primarily by phenomenal states (e.g., hunger, fear, pain, pleasure); the other axis, which the authors labeled 'agency,' was constituted primarily by states of higher cognition (e.g., self-restraint, moral judgment, memory, and planning). The results allow for somewhat different interpretations as well, because the two mental state sets similarly map onto the distinctions of affect versus cognition and of unintentional versus intentional. It would be interesting to examine where the classic perceptual states fall (seeing red, smelling coffee). If the 'agency/intentionality' interpretation of the second axis is correct, they should be among the experiences (axis 1); if the 'cognition' interpretation of the second axis is correct, they should be among these nonexperiential states (axis 2). Similarly, where would beliefs fall, the markers of propositional attitudes? They belong to axis 2 if this axis represents cognitions but not if it represents agency: and they might belong to axis 1 if they can be phrased in quasi-experiential ways (e.g., ''I feel like we have been here before'').

Knobe and Prinz began with the working hypothesis that people ascribe phenomenal states only to entities that have both brains and bodies, whereas they ascribe propositional states like beliefs or desires to many more entities. Group agents constituted the authors' test case, because they are not physically realized as brain and body but are nonetheless the targets of everyday mental state ascriptions (e.g., reason explanations). The results of six studies suggested that people ascribed beliefs and desires, but not phenomenal states, to

group agents, supporting the authors' hypothesis. However, subsequent research has cast some doubts on these conclusions. First, even though people do not ascribe states such as 'feeling regret' to groups in the abstract, they do ascribe concrete versions of such states to groups, such as 'Acme Corporation is feeling regret about its recent decision.' Also, people from Eastern cultures appear to have far less trouble ascribing phenomenal states to groups.

Huebner used different target entities, asking people to agree or disagree with various statements that ascribe beliefs, pain, and emotions to humans versus robots, each of which has either a biological or a machine brain. In two studies, people ascribed beliefs (e.g., ''He believes that 2 + 2 = 4'') roughly equally to the four entities, whereas they ascribed pain and other emotions (e.g., happy feelings) more comfortably to humans with brains than to any of the other three agents. The data, however, do not show people actually denying the capacity for pain or emotions to nontypical agents. Rather, their means of 2.6–3.3 (on a 1–5 scale) indicate they neither agreed nor disagreed with the ascriptions, essentially withholding judgment.

The philosophical concept of phenomenal consciousness includes not only states of pain and emotion but also sensory experiences such as 'seeing an intense red' or 'smelling a rose.' In a series of studies, Sytsma and Machery examined whether ordinary people's (as compared to philosophers') concept of consciousness is similarly inclusive in this way. Whereas philosophers in the sample ascribed neither seeing red nor feeling pain to a simple robot with camera, arms, and wheels (that easily performed a color discrimination task or received an electric shock), ordinary people ascribed seeing red to it, but not feeling pain or feeling anger. Are sensory experiences thus excluded from people's folk concept of phenomenal consciousness? Not quite.

Follow-up studies in Sytsma and Machery's article showed that people denied some sensory states to the robot (e.g., smelling banana, smelling vomit), and the authors interpreted this finding as showing that people distinguish 'affective' sensory states from simple perception and apparently require an affective component for phenomenal consciousness (as is present in pain but not in seeing). Language habits do confirm this line that people seem to draw: Whenever affect is involved we can use the term 'feeling' to describe the conscious state, and then the typical phenomenal claim that some conscious states 'feel a certain way' follows easily. That seeing red 'feels' a certain way may be more difficult to accept.

One might conclude, then, that people separate feelings, emotions, and other affective states from the rest of the mind's landscape, being reluctant to ascribe them to agents who lack either a biological body or brain (e.g., robots, companies). Until disproven, we may consider that this concept of feeling states a third prototype of consciousness.

Many questions remain. Do people grant robots desires (which often have affect)? Or only goals (which can be construed purely cognitively)? What about mixed concepts – can robots, for example, 'approve' of things? Approval may be defined as an affectively positive state of believing that something is right or appropriate; and such a state crosses the line between phenomenal and propositional mental states. Does the denial of feeling states hold only for primitive machines? Would people, for example, grant affective states to an android whose brain and body is organized very much like a human's but consists entirely of synthetic materials? What if a human's brain and body tissue were radically rearranged and replaced by synthetic cells and organs? And if an android grew up in a human family, would it learn to have emotions and affectively charged sensations? These questions are not merely theoretical. They are important for the granting of moral status, which we discuss next.

Implications for Personhood

An increasingly interesting topic is the relationship between assigning the capacity for consciousness to an entity and its moral (and legal) status. Intelligent, sentient machines are richly illustrated in science fiction novels, but they may be in store for humanity sooner than one might think. People's folk theory of consciousness will be the main criterion by which artificial intelligence or extraterrestrial life will be classified and treated. Science may convince us of such potential agents' objective similarity or difference to humans; but it is people's

folk judgments, their concepts, criteria, and ascription practices that will have the final word.

Might the prototypes of consciousness provide the criteria for identifying 'persons' – those to whom we grant rights and responsibilities? The capacity for conscious monitoring does not seem sufficient for the application of certain moral and legal concepts. Many sophisticated machines have (self-)monitoring capacities without being considered persons. It is most likely the second prototype of consciousness that makes an entity a fully responsible person – the capacity to consciously and rationally consider facts, desires, and their implications and choose a course of action in light of them. This is absent in severe mental illness, sleep walking, hypnosis, infancy, and (more debatably) in many animals, and this may put those agents out of reach for full responsibility (but perhaps not personhood).

What about the (potential) third prototype, that is, phenomenal consciousness? Huebner argues that ascriptions of affective, emotional states to an entity ''recruit hot-empathetic mechanisms that drive us into the realm of moral concern.'' He illustrates this point with the film Blade Runner, in which an individual is classified as human or mere robot by the 'Voigt-Kampff test,' which probes for subtle physiological responses to emotion-arousing stimuli. Robots do not have those emotions; hence, they do not (or should not?) arouse our empathic emotions either.

However, several counter-examples suggest phenomenal states may not be necessary for personhood. 'Data,' an android from Star Trek: The Next Generation, allegedly feels no pain, cannot be nervous, and neither seems to have emotions nor understands human emotions very well. However, most characters within the story treat him as a person, and most viewers presumably do too. Data clearly has the first two kinds of consciousness, and that might suffice for personhood. Phenomenal consciousness may not be sufficient either, if the portrayals of human–machine relationships in Battlestar Galactica are representative. The human characters grant the Cylons (a machine species that developed biological organs) pain and other experiences, but somehow that is not enough. The show gets a bit murky about the criteria that ultimately make humans different – apparently the capacity for love, conscience, and some kind of unique biological origin. One fascinating issue posed in this show is whether a child from a human parent and a machine parent is a person. At first, the human characters within the story deny such a possibility, but just as they develop doubts over time, so may humanity in its real future.

Privileged Access and Incorrigibility

According to Dennett (2005), one assumption of the folk theory of consciousness is that ''consciousness is utterly private, inaccessible to outsiders, somehow at least partly incommunicable'' (p. 27). As is true for many philosophers writing about folk theories, the author musters no empirical data to support this diagnosis. In looking for pertinent research it helps to break up the statement into two separate claims: that consciousness is 'private' (inaccessible to other people) and that it is 'incommunicable.'

The incommunicability claim can be set aside, because it is falsified by vast amounts of mental state language found cross-culturally and by the terrific comfort people have in addressing, expressing, and discussing their own and other people's mental states. There may be subtle and novel aspects of mental states that people find difficult to describe, but these are exceptions; and even in such cases, it is not as if the experiencer had a perfectly clear inner picture and just cannot find the right words for other people. Complex and novel combinations of experiences are inherently difficult to grasp and describe, both in the mind or in the public world.

That people assume their conscious states to be private is more plausible, and there we can examine two versions: People may believe that they find out about their own mental states in ways that are not available to others ('privileged access'); and they may believe that they can never be wrong in assessments of their own conscious states ('incorrigibility').

Privileged Access

According to the awareness assumption for reason explanations, if actors are honest and undeluded,

the reason explanations they offer for their actions cite the very reasons for which they acted – the contentful mental states such as beliefs and desires that generated the intention to act. People can misremember, and observers can muster evidence that proves certain dishonest reason claims wrong; but there is privileged access in that normally only the actor knows the very reasons on the grounds of which and in light of which she decided to act. Partially as a result of this access asymmetry, people explain their own intentional actions with more reasons and fewer causal history explanations than they explain other people's actions.

However, people do not assume that humans have exclusive access to their conscious reasons. They feel perfectly comfortable inferring and ascribing mental states in other people and ascribing reasons to them. They even feel comfortable ascribing reasons to group agents. The apparent epistemic privilege for reason explanations may not be unlike the teacher's privilege in knowing more about the subject matter than the student (including how to locate and expand that knowledge).

Research has documented additional asymmetries between first- and third-person perspectives that go beyond reasons and are suggestive of a privileged access assumption for mental states in general or at least indicate a practice consistent with it. In social interaction, actors attend to and explain observable behaviors far less often than observers do. This pattern is completely reversed for mental states, which actors attend to and explain far more often than observers do. The latter finding shows that actors still have to figure out why they have certain mental states (e.g., certain bodily feelings, emotions, or intrusive thoughts). The privilege of access (knowing about conscious states) does not extend to the privilege of knowing why one has that state – such a privilege of explanation is assumed only for reason explanations of intentional actions.

But do people assume a more fundamental difference in the 'kind' of access that the first person has to her consciousness? Folk psychology does not seem to entail a conception of introspection specific enough to answer this question – or at least no research has documented such a conception yet. The known obstacles to psychotherapy, meditation practice, and everyday mindfulness suggest that ordinary people are not exactly habitual introspectors. The 'mind's eye' that, according to philosophers, ordinary people believe in, often is half closed or suffers from blurry vision. But such introspection amounts to conscious second-order cognition – a state of conscious awareness of a second, separate conscious state (e.g., noticing that you are angry; realizing that you are daydreaming). This state is not the same as everyday conscious experiences. To use an analogy, normally when a camera records reality, it does not record itself in this reality. An unusual case occurs when the camera becomes part of the picture (e.g., when recorded by a second camera), and that corresponds to the case of explicit introspection, when people register themselves as registering some internal states.

More typical than second-order cognition is people having conscious first-order mental states (being angry, daydreaming) that are 'transparent' – that is, one does not need to ask for, search, or infer evidence for that mental state. As the Oracle in *The Matrix* put it, ''No one needs to tell you that you are in love, you just know it, through and through.'' Likewise, the concert audiency need not be told that they hear orchestral music, or the patient that he is in pain. Having or being in such states is enough for actors to become aware of them, and because observers cannot have or be in someone else's states they therefore must take a different (e.g., inferential or simulating) route to become aware of them. This, at least, appears to be the folk assumption.

People appreciate, however, that having or being in a conscious state reveals only that state's content (e.g., the particular sensation of the viola's timbre), not necessarily the reality that the state represents. People understand the difference between subjective appearance and reality, and children learn to grasp this distinction during the preschool years. Clark argues that proper use of verbs of appearance – such as 'seem, look, appear, feel' – reveals a tacit insight into the subjectivity of conscious states. When emphasizing that something seems, looks, or feels a certain way, people notice and reflect on discrepancies between their own perceptual experience and (what they independently believe to be) reality. They compare current sensations to memories, impressions to

knowledge, and one's own impression to another person's. And, of course, in addition to making the appearance-reality distinction, people also ascribe this capacity to others. In recognizing this capacity, people may integrate versions of the first and second prototype of consciousness: monitoring the content of one's experiences, considering multiple representations (e.g., perception and memory), and choosing the presumed 'real' one among them.

Incorrigibility

The transparent character of at least some conscious mental states and their designation as subjective appearances suggests a possible assumption of incorrigibility – the claim that people are infallible in accessing their minds because the data (i.e., the experience) and the truth criterion (whether the experience is really there) are one and the same.

It is safe to say that ordinary people would not claim that all mental states are infallible (clearly, perception, belief, and memory are error prone), or that one's report about an internal state is infallible in describing the exact properties of that state (e.g., describing a particular emotion or attitude can be hampered by uncertainty or the state's complexity). However, experienced phenomenal states – perceptions, bodily states, and feelings – can hardly be doubted or denied 'as being there,' as the experiences that they are. Two people may argue about a particular quality of an object (e.g., whether it is pleasant or not), but each person can – and people routinely do – retreat to a subjective claim such as, "Okay, but that's what it feels to me – it seems unpleasant." The interlocutor has little to reply to such a claim – the 'seeming,' the experience of displeasure, cannot be overruled. At present, we do not have systematic data on the degree to which people regard such cases as truly incorrigible.

Consciousness and Free Will

In discussions of free will, whether as a perceived or real phenomenon, the issue of consciousness often comes along for free. Is conscious will an illusion? Does consciousness really cause behavior? We must ask, then, how people's folk theory of consciousness relates to their folk theory of free will.

There is one innocent way of relating: 'willing' is the state of desiring, wishing, and wanting, and this state is consciously experienced, reflecting the first and third prototypes of consciousness. At stake, however, is the second prototype, the notion of 'choice,' which occurs in light of and because of desires (and beliefs). In ordinary terms, choice is always conscious, and if anything is considered 'free,' it is choice, not will or desire. Folk assumptions about conscious choice are, in the eyes of many scholars, deeply mistaken. For Prinz (2003), "there appears to be no support for the folk psychology notion that the act follows the will, in the sense that physical action is caused by mental events that precede them" (p. 26). Two questions arise: exactly what does the folk concept of free choice assume? And what is the support for the claim that those assumptions are false?

Probing the folk concept of free choice, we need to ask first what this choice is assumed to be free from. Some scholars argue that the folk assumption of free will implies "the replacement of usual causal determination through another, causally inexplicable form of determination" (Prinz, 1997, p. 161), the belief that "willfulness somehow springs forth from some special uncaused place" (Bayer et al., p. 100). However, there is actually no evidence that people's conception of free choice is a claim about uncaused causes. In a short survey, Andrew Monroe and I asked 200 undergraduate students a simple free-response question: "What does it mean to have free will?" Their answers referred to three main concepts: 47% of responses mentioned the role of decision and choice (e.g., "to have a choice in what you do"); 23% mentioned the role of desires (e.g., "ability to act on one's owns needs and desires"); and 24% mentioned the absence of constraints (e.g., "you aren't forced into anything"). Not a single response referred to an uncaused cause.

There is also no evidence that free choice is free from obeying natural laws. People certainly believe that their intentions have a lawful causal force to bring about action (how else could they rely on their intentions to cause actions?), and the choice process itself has to obey the laws of both causality and rationality – the latter dictating a certain way in which the contents of reasons have to combine to provide the grounds for action.

Another charge is that the concept of free will involves a ''renunciation of explanation and cutting short of causal chains'' (Prinz, 1997, p. 162). People's folk explanations of behavior seamlessly explain actions and intentions by reasons and they explain reasons themselves by way of causal history of reason explanations. It would be a rather disturbing thought for ordinary people that their intentions or choices inexplicably pop into being.

So what are free actions free from? According to the abovementioned survey, free actions are sufficiently free form external forces and internal obstacles (e.g., ''It means that you don't have a higher power controlling your actions, and that biological impulses can be overcome.''). Among the external constraints mentioned we find primarily other people, and occasionally legal or moral restraints. Among internal obstacles we find intelligence, disabilities, genetic make-up, and so on.

To summarize, survey data suggest that people consider 'free will' to be a person's capacity to choose to act on the basis of what one wants. From previously reported research on behavior explanations we also know that conscious choices are understood to be grounded in reasons, but those reasons are causally shaped by a variety of factors, such as personality, culture, or context.

Given this modest folk concept of free choice, what might be the arguments to consider the folk concept an illusion? Two types of arguments have become fashionable. One is the argument from the unconscious, and the reasoning is given as follows: Because unconscious processes demonstrably exert a causal influence on behavior (e.g., behavior priming, unconscious visual processing), much or all of behavior that we think is under conscious control may not be. Bargh (2005), for example, writes that ''conscious intention and behavioral (motor) systems are fundamentally dissociated in the brain'' (p. 43). Or more succinctly, ''the real causes of human action are unconscious'' (Wegner & Wheatley 1999, p. 490). But unless one believes that all behavior is brought about by free choice or that conscious reasons are the only causes of behavior, nothing in the argument threatens the folk conception of free choice. As shown earlier, people's explanations of intentional actions are inclusive of processes that lie outside the conscious reasoning process, and nobody will deny that lower-level brain and body processes are important proximal causes of behavior (even intentional behavior). So the folk concept of free choice can happily accommodate unconscious causal factors.

The second type of skeptical argument refers to neuroscience experiments designed to demonstrate that intentional action is directly caused and controlled by brain processes and that intentions are too late in the causal chain or entirely left out of the loop. For example, trans-magnetic stimulation (TMS) was shown to influence participants' decisions to use their right versus left hand in a simple choice task. Participants' behavior was clearly intentional, but they were unaware that the TMS had influenced those decisions. Data such as these, however, are well explained by the notion that TMS triggered an idea or preference or desire and that the person took this preliminary motivational state into account when deciding to act one way or another. Once more, the presence of a causal history of reasons does not vitiate the causal force of reasons themselves. (For discussion of related skeptical arguments, see 'Suggested Readings.')

One of the curious features of recent skepticism about free choice is the distinction scholars make between intentions and (unconscious) brain processes. Ryle would have a field day with this category error. Intentions are of course somehow implemented in the brain (how else could they engage with motor programs?), but causal relations do not just occur at the neural, molecular, or atomic level. The fire that burned down the house was somehow instantiated by complex atomic events, but we do not therefore conclude that the fire was not really causally efficacious in leveling the house.

Summary

''The everyday notion of consciousness is considerably ambiguous and devoid of clear boundaries'' (Lahav, 1997, p. 178). Against this claim, we can see that people's folk theory of consciousness can be divided into three relatively distinct prototypes of conscious mental functioning: monitoring (awareness); choice; and subjective experience.

Each of these prototypes is embedded in the broader folk theory of mind, and the first two are integral parts of the concept of intentional action. The third prototype, subjective experience, is clearly manifest in language (verbs of experience and appearance), and recent research suggests that people's ascriptions of experience may have to meet conditions (in particular, having a brain and body) that do not apply to other mental states (e.g., beliefs and desires). However, whether this folk concept of conscious experience corresponds to the theoretical construct of phenomenal consciousness is still debated.

At least some of the prototypes of consciousness play a critical role in the assignment of personhood, an issue that is relevant to questions of animal rights, legal capacity, and the potential status of future artificial intelligence.

All conscious mental states, whether monitoring, executive, or experiencing, are believed to enjoy some degree of privileged access for the person having those states. However, this privilege does not prevent people from inferring, communicating about, and evaluating other people's mental states. Though not in the bright public lights, minds are not considered solipsistically private.

Recent skepticism about the importance of conscious states has focused on the second prototype, that is, the conscious choice to act. Some scholars have claimed that such 'free choice' is an illusion, but when fairly assessing what the folk concept really entails, all evidence mustered for such a skeptical position thus far complements the notion of free choice without contradicting it. The folk concept of choice operates at the functional level, postulating a rational and causal connection between particular kinds of mental states (beliefs, desires) and an intention to act and leaves plenty of room for nonconscious contributions to behavior. The folk concept is mute about the specific implementation of these states, whereas scientific evidence focuses precisely on such aspects of implementation.

The folk theory of consciousness does not answer the broad philosophical question of what consciousness is. However, it clarifies how we should pose the question of what consciousness is – for example, by distinguishing between monitoring, choice, and experience as three distinct phenomena (and much of philosophy has done so). Furthermore, a careful study of people's folk concepts reveals an appreciation for complexity and subtlety that is often not granted to ordinary people. Ironically, some of these subtleties are missing from scientific theories about consciousness. Many scholars, for example, fail to make the distinction that people make between an agent's reasons for acting and the causal history of those reasons. No skeptical or constructive scientific theory of conscious and unconscious mental functioning should not omit this distinction or, for that matter, any other aspects of the evolved folk conception of consciousness. Science go beyond this folk conception, but it must not overlook it.

See also: Cognitive Theories of Consciousness; Concepts and Definitions of Consciousness; History of Philosophical Theories of Consciousness.

Suggested Readings

Bargh JA (2005) Bypassing the will: Towards demystifying behavioral priming effects. In: Hassin R, Uleman JS, and Bargh JA (eds.) The New Unconscious, pp. 37–58. New York: Oxford University Press.

Bayer UC, Ferguson MJ, and Gollwitzer PM (2003) Voluntary action from the perspective of social-personality psychology. In: Maasen S, Prinz W, and Roth G (eds.) Voluntary Action: Brains, Minds, and Sociality, pp. 86–107. New York: Oxford University Press.

Blakemore S-J, Wolpert DM, and Frith CD (2002) Abnormalities in the awareness of action. Trends in Cognitive Sciences 6: 237–242.

Bybee J (1994) The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World. Chicago: The University of Chicago Press.

Clark A (2008) Phenomenal properties: Some models from psychology and philosophy. Philosophical Issues 18: 406–425.

D'Andrade R (1987) A folk model of the mind. In: Holland D and Quinn N (eds.) Cultural Models in Language and Thought, pp. 112–148. New York: Cambridge University Press.

Dennett D (2005) Sweet Dreams: Philosophical Obstacles to a Science of Consciousness. New York: Basic Books.

Gray H, Gray K, and Wegner D (2007) Dimensions of mind perception. Science 315: 619.

Huebner B (2008) Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? Unpublished manuscript retrieved 24 July 2008. http://www.unc.edu/~huebner/androids.pdf.

Knobe J and Prinz J (2008) Intuitions about consciousness: Experimental studies. Phenomenology and the Cognitive Sciences 7: 67–83.

Lahav R (1997) The conscious and the non-conscious: Philosophical implications of neuropsychology. In: Carrier M and Machamer PK (eds.) Mindscapes: Philosophy, Science, and the Mind, pp. 177–194. Pittsburgh, PA: University of Pittsburgh Press.

Malle BF (2004) How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction. Cambridge, MA: MIT Press.

Malle BF (2005a) Folk theory of mind: Conceptual foundations of human social cognition. In: Hassin R, Uleman JS, and Bargh JA (eds.) The New Unconscious, pp. 225–255. New York: Oxford University Press.

Malle BF, Knobe J, O'Laughlin MJ, Pearce GE, and Nelson SE (2000) Conceptual structure and social functions of behavior explanations: Beyond person-situation attributions. Journal of Personality and Social Psychology 79: 309–326.

Monroe ALE and Malle BF (2008) From uncaused will to conscious choice: The need to study not speculate about, people's folk concept of free will. European Review of Philosophy.

Pockett S, Banks WP, and Gallagher S (eds.) (2006) Does Consciousness Cause Behavior? An Investigation of the Nature of Volition. Cambridge, MA: MIT Press.

Prinz W (1997) Explaining voluntary action: The role of mental content. In: Carrier M and Machamer PK (eds.) Mindscapes: Philosophy, Science, and the Mind, pp. 153–175. Pittsburgh, PA: University of Pittsburgh Press.

Prinz W (2003) How do we know about our own actions? In: Maasen S, Prinz W, and Roth G (eds.) Voluntary Action: Brains, Minds, and Sociality, pp. 21–33. New York: Oxford University Press.

Ryle G (1949) The Concept of Mind. London and New York: Hutchinson.

Schwanenfluegel PJ, Fabricius WV, Noyes CR, Bigler KD, and Alexander HM (1994) The organization of mental verbs and folk theories of knowing. Journal of Memory and Language 33: 376–395.

Sytsma J and Machery E (in press) How to study folk intuitions about phenomenal consciousness. Philosophical Psychology.

Wegner DM (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

Wierzbicka A (1996) Semantics: Primes and Universals. Oxford: Oxford University Press.

Wittgenstein L (1953) Philosophical investigations, Anscombe GEM (trans.). England: Oxford, Basil Blackwell (Original work published in 1953).

Biographical Sketch

Bertram F. Malle studied psychology, philosophy, and linguistics at the University of Graz, Austria. After earning two master's degrees at Graz, Malle continued his studies in psychology at Stanford University, where he received his PhD in 1995. Between 1994 and 2008 he was assistant, associate, and full professor of psychology at the University of Oregon and served there as Director of the Institute of Cognitive and Decision Sciences from 2001 to 2007. Since fall of 2008 he has been professor of psychology in the Departments of Psychology and Cognitive and Linguistic Sciences at Brown University. Malle received the 1995 Society of Experimental Social Psychology Dissertation Award and a National Science Foundation CAREER award 1997–2001. He is currently president-elect of the Society for Philosophy and Psychology. Professor Malle's research examines the cognitive tools that humans bring to social interaction, especially the capacity to recognize intentionality; make inferences about mental states; and explain, predict, and morally evaluate human behavior. He is author of over 50 articles and chapters, coeditor of three published volumes, *Intentions and Intentionality* (2001, MIT Press), *The Evolution of Language Out of Pre-language* (2002, Benjamins), and *Other Minds* (2005, Guilford), and author of the monograph *How the Mind Explains Behavior* (2004, MIT Press).

Free Will

A R Mele, Florida State University, Tallahassee, FL, USA

Glossary

Compatibilism – The thesis that free will is compatible with the truth of determinism.
Determinism – The thesis that the combination of a complete statement of the laws of nature and a complete description of the condition of the entire universe at any point in time logically entails a complete description of the condition of the entire universe at any other point in time.
Distal decisions and intentions – Decisions and intentions about what to do later.
Epiphenomenalism – The thesis that although all mental events are caused by physical events, no mental events cause any physical events.
Incompatibilism – The thesis that free will is incompatible with the truth of determinism.
Libertarianism – The conjunction of incompatibilism and the thesis that people sometimes perform free actions.
Proximal decisions and intentions – Decisions and intentions about what to do immediately.
Readiness potential – A measure of activity in the motor cortex that precedes voluntary muscle motion.

Introduction

If a friend were to ask you whether blop exists, you would ask what 'blop' means. Similarly, anyone who wishes to discover whether free will exists should ask what 'free will' means. The meaning of 'free will' is the topic of the section 'Theoretical background'. This section surveys leading philosophical theories about what 'free will' means and provides a theoretical background for the discussion in subsequent sections of empirical work on free will.

Theoretical Background

Definitions

Free will may be defined as the power to perform free actions. But what does 'free action' mean? Traditional philosophical answers fall into two groups: compatibilist and incompatibilist. Compatibilism and incompatibilism are positions on the relationship between free action and determinism. Determinism is the thesis that the combination of a complete statement of the laws of nature and a complete description of the condition of the entire universe at any point in time logically entails a complete description of the condition of the entire universe at any other point in time.

Compatibilism is the thesis that free action is compatible with – that is, not precluded by – the truth of determinism. Because they attend to what contemporary physics tells us, the overwhelming majority of contemporary compatibilists do not believe that determinism is true; but they do believe that even if it were true, it would leave it open that people sometimes act freely. Their identifying themselves as compatibilists is explained by a lengthy tradition of framing the issue in terms of whether or not determinism precludes free will.

Incompatibilism is the thesis that free action is incompatible with the truth of determinism. In the incompatibilist group, most answers to the question "What does 'free action' mean?" come from libertarians. Libertarianism is the conjunction of incompatibilism and the thesis that some people sometimes perform free actions. Some incompatibilists argue that no one can perform free actions. They argue that even the falsity of determinism creates no place for free action.

Compatibilist Theories

The compatibilist thesis often sounds strange to nonspecialists. When people first encounter the pair of expressions 'free will' and 'determinism' they tend to get the impression that the two ideas are defined in opposition to each other and are mutually exclusive by definition. This is one reason that it is useful to think of free will as the power to perform free actions and to regard free action as the more basic notion – that is, as a notion in terms of which free will is to be defined. Consider the following conversation between two firefighters who have a stingy colleague named Stan. Alice: "Stan gave $20 to a homeless man today." Bill: "Why? Did he hold a gun to Stan's head?" Alice: "No, Stan gave him the money freely." Surely, Alice and Bill do not need to have an opinion about whether determinism (as defined above) is true to have this conversation. If what Alice says is true – that is, if Stan freely gave the person $20 – and free will is the power to perform free actions, then Stan has free will (or, at least, he had it at that time). Even if free will is typically opposed to determinism in ordinary speech, the expression 'he did it freely' seems not to be. And even if 'he did it freely' were typically opposed to determinism in ordinary speech, that would settle nothing. After all, in ordinary speech, deductive reasoning seems to be defined as reasoning from the general to the particular, and that certainly would only jokingly be said to constitute an objection to a logician's definition of deduction (according to which 'Alice is a firefighter; Bill is a firefighter; therefore Alice and Bill are firefighters' is a valid deductive argument). The standard definition of deduction in logic is very useful even if it does not capture ordinary usage. Similarly, a compatibilist definition or analysis of 'he did it freely' may be very useful even if it does not capture ordinary usage.

Compatibilist theories about the meaning of 'free action' emphasize a distinction between deterministic causation and compulsion. If determinism is true, then this author's driving to work today, his writing this paragraph, and so on, were deterministically caused; and so were a certain compulsive hand-washer's washing his hands dozens of times today, a certain delusional person's spending the day trying to contact Martians with his cell phone, a certain crack addict's using crack today, while in the grip of an irresistible craving for the drug, and a certain person's handing over the money in his pocket to gunmen who convincingly threatened to kill him if he refused. But there is an apparent difference. This author is sane, suffers from no uncontrollable addictions, and has received no death threats today. The basic compatibilist idea is that the meaning of 'free action' is such that when mentally healthy people act intentionally in the absence of compulsion and coercion they are performing free actions, and an action's being deterministically caused does not suffice for its being compelled or coerced.

Many compatibilists have been concerned to accommodate the idea that, for example, if this author freely drove to work this morning, he could have done something other than drive to work this morning. They grant that, if determinism is true, then there is a sense in which people could never have done otherwise than they did: they could not have done otherwise in the sense that their doing otherwise is inconsistent with the combination of the past and the laws of nature. But, these compatibilists say, the fact that a person never could have done otherwise in that sense is irrelevant to free action. What is relevant is that people who perform free actions are exercising a rational capacity of such a kind that if their situation had been different in any one of a variety of important ways, they would have responded to the difference with a different suitable action. For example, although this author drove to his office today, he would not have done so if a friend had bet him $500 that he would not stay away from his office all day. This truth is consistent with determinism. (Notice that if a friend had made this bet with this author, the past would have been different from what it actually was.) And it reinforces the distinction between deterministic causation and compulsion. Offer a compulsive hand-washer $500 not to wash his hands all day and see what happens.

Libertarian Theories

Libertarian theories about the meaning of 'free action' divide into three kinds: noncausal,

event-causal, and agent-causal. Most theories about what intentional actions are include a causal condition. Roughly speaking, according to these theories, all intentional actions are events that are caused in a certain distinctive range of ways – either deterministically or indeterministically. For example, it may be claimed that what it is to be an intentional action is to be an event that is suitably caused by motivational and representational states. Noncausal libertarian theories about the meaning of 'free action' reject this idea. Like compatibilists, noncausal libertarians tend to maintain that when mentally healthy people act intentionally in the absence of compulsion and coercion they are performing free actions, but they insist both that the deterministic causation of an action precludes its being a free action and that uncaused events can be intentional actions.

Typical event-causal libertarian theories about the meaning of 'free action' assert that agents never perform free actions unless some of their actions are indeterministically caused by immediate antecedents that are events in the agents. Whereas the laws of nature that apply to deterministic causation are exceptionless, those that apply most directly to indeterministic causation are instead probabilistic. Typically, events like deciding to help a homeless person – as distinct from the physical actions involved in actually helping – are counted as mental actions. Suppose that Aida's decision to help a homeless person is indeterministically caused by, among other things, her thinking that she should help. Given that the causation is indeterministic, she might not have decided to help given exactly the same internal and external conditions. In this way, event-causal libertarians seek to secure the possibility of doing otherwise, which they require for free action, or for fundamentally free action (i.e., free action that does not derive its freedom solely from earlier free actions the agent performed).

Agent-causal libertarian theories about the meaning of 'free action' assert that agents themselves – as opposed, for example, to agents' motivational and representational states – are causes of free actions. According to these theories, the meaning of 'free action' is such that acting freely requires 'agent causation.' Causation may be regarded as a relation between cause and effect. In ordinary event causation – for example, a landslide's crushing a house – both the cause and the effect are events. These events are connected by the relation causation. In agent causation, an agent is connected by the relation causation to an action and that connection is not reducible to a connection between states or events and the action. Whereas most agent-causal libertarians prefer their agent causation straight, some mix it with event causation in a theory about the production of free actions.

Reservations about Compatibilist and Libertarian Theories

Each of the theories about the meaning of 'free action' described above has its detractors. Some theorists view determinism as precluding a kind of flexibility that they take to be required for free action and therefore reject all compatibilist theories about what 'free action' means. Even if they accept the compatibilist distinction between deterministic causation and compulsion, they contend that each precludes the required flexibility in its own way. Other theorists regard uncaused actions as impossible and therefore reject noncausal libertarianism as impossibly demanding. Yet others maintain that although event-causal libertarianism introduces a chance of acting otherwise that is absent in deterministic universes, it does not give agents the sort of control over their actions that free action requires. They argue, for example, that event-causal libertarianism has the undesirable result that it is just a matter of chance that an agent decides on a particular course of action at a given moment rather than deciding on an alternative course of action then. Some theorists raise the same objection to agent-causal libertarianism, and others argue that agent causation is impossible and that agent-causal libertarianism is impossibly demanding.

Do We Ever Act Freely?

Are there free actions? That depends on what 'free action' means. If a typical compatibilist theory about the meaning of 'free action' is correct, then it is very likely that people often perform free actions. After all, there are many mentally healthy people, and it is difficult to deny that they often

perform intentional actions in the absence of compulsion and coercion. And, of course, if free will is simply the power to perform free actions, anyone who performs free actions has free will (at least on some occasions). If any of the libertarian theories about what 'free action' means are correct, it is easier to doubt that people are capable of performing free actions. If there are no uncaused actions, then no theory about the meaning of 'free action' that requires that free – or fundamentally free – actions be uncaused permits us to perform free actions. If the brain, in fact, does not work indeterministically in ways required for free action by event-causal libertarian theories about the meaning of 'free action,' and if those theories are correct, it turns out that no one performs free actions. And if there is no agent causation in the real world (or if agent causation is impossible), agent-causal libertarianism has the same upshot.

Moral Responsibility

Often, theorists who disagree about what 'free action' means, attempt to find support for their view in the closely related sphere of moral responsibility. A common claim about the power to perform free actions – that is, free will – is that it is a necessary condition for being morally responsible for actions one performs, where moral responsibility is understood as a necessary condition for such things as deserved punishment, deserved moral blame, and deserved moral praise or credit. The various theories about the meaning of 'free action' described here have readily recognizable counterparts in the sphere of moral responsibility.

Connecting Free Will to Consciousness

One Empirical Approach to Studying Free Will

Whether compatibilism or incompatibilism is true is a conceptual question – a question about how the concepts of free action and determinism are related. How might scientists proceed if they wish to investigate free will without taking a stand on this conceptual, philosophical question? One way is to study phenomena that are commonly thought to be associated with free will and can, in principle, occur whether or not determinism is true: for example, delay of gratification and decision making. Presumably, if physicists were to discover that determinism is true, we would not conclude that no one has ever successfully resisted temptation or that no one has ever made a decision. Incompatibilists would conclude that no one has ever freely done these things, but that is another matter.

Other things being equal, actions that are consciously performed for conscious reasons would seem to be better candidates for being free actions than actions that are unconsciously performed or actions that agents consciously perform, but not for any reasons of which they are conscious. Like delay of gratification and decision making, consciousness depends neither on the falsity of determinism nor on its truth. A lot of scientific work has been done on what place consciousness may or may not have in the production of intentional actions. In this connection, it is important to distinguish between actual findings and inferences made partly on the basis of those findings. Some findings and inferences are discussed later.

Conceptual Distinctions

Attention to some conceptual distinctions facilitates the discussion. It is often thought that free overt actions – that is, free actions that essentially involve peripheral bodily motion – are products of free, conscious decisions and that these decisions are free in a more basic way than free overt actions are. So decisions require special attention. Some of our decisions are about things to do immediately. They are proximal decisions. Others – distal decisions – are about things to do later. A shy student, Stu, who has been thinking about when it would be best to raise his hand to attract his teacher's attention, decides to raise it now. This is a proximal decision. Later in the day, after thinking about when to start writing a term paper of his, Stu decides to start it next Tuesday. This is a distal decision. The scientific work on decision making that is most closely associated with skepticism about free will focuses on proximal decisions. Some of that work is the topic of subsequent sections.

Deciding to do something should be distinguished from wanting (or having an urge) to do

it. Sometimes we want to do things that we decide not to do. And often, when we want to do each of two incompatible things – for example, meet some friends for lunch at noon and go to class at noon – we settle matters by deciding to do one and not the other. To decide to do something is to form an intention to do it; and just as deciding should be distinguished from wanting, so should intending. Intending to do something seemingly is more tightly connected to action than is merely wanting to do it. Also, just as there are proximal and distal decisions, there are proximal and distal intentions – intentions to do things now and intentions to do things later.

Free Will and Consciousness: Proximal Decisions about Timing

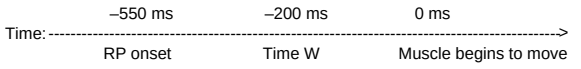
Libet's Studies

Neurobiologist Benjamin Libet, in a body of work that has received an enormous amount of attention, has argued that the brain decides to initiate actions before the person is aware of these decisions. He also contends that there is about a 100 ms window of opportunity for free will to get involved and veto the brain's decision. Libet has many critics and many supporters. Some scientists follow him part of the way. They accept the thesis that our decisions precede our awareness of them, but reject the window of opportunity for free will as an illusion.

Libet's findings are generated by an innovative series of studies. In some of the studies, subjects are regularly encouraged to flex their right wrists whenever they wish. In subjects who do not report any advance planning of their movements, electrical readings from the scalp – averaged over at least 40 flexing actions for each subject – show a shift in readiness potentials (RPs) beginning about 550 ms before the time at which an electromyogram shows relevant muscular activity to begin. These are type II RPs. Subjects who are not regularly encouraged to aim for spontaneity or who report some advance planning produce RPs that begin about half a second earlier – type I RPs. The same is true of subjects instructed to flex at a prearranged time.

Subjects are also instructed to recall where a revolving spot was on a special clock when they first became aware of something, x, that Libet variously describes as a decision, intention, urge, wanting, will, or wish to move. (The spot on this Libet clock moves about 25 times faster than the second hand on a normal clock.) On average, the onset of type II RPs preceded what the subjects reported to be the time of their initial awareness of x (time W) by 350 ms. Time W, then, preceded the beginning of muscle motion by about 200 ms. Libet's findings for type II RPs may be represented as follows:

Libet's findings for type II RPs



(Libet finds independent evidence of what he regards as an error in the subjects' recall of the times at which they first become aware of sensations. Correcting for it, time W is ▫ 150 ms.)

An Inference Examined

One inference that Libet makes on the basis of these findings is that the brain produces a proximal decision or intention to flex about a third of a second before the subject becomes aware of that decision or intention. Is this inference warranted? One alternative hypothesis is that what the brain produces around ▫ 550 ms is a potential cause of a subsequent proximal decision or intention to flex and that the decision or intention emerges significantly later. Libet's findings do not contradict this hypothesis.

How might one get evidence about whether the onset of the type II RPs at ▫ 550 ms is correlated with an unconscious proximal decision or intention to flex or instead a potential cause of a proximal decision or intention to flex? An apt question to ask in this connection is how long it takes a proximal intention to flex to generate muscle activity (a muscle burst). If, in fact, the brain produces proximal decisions or intentions in Libet's study about 550 ms before the muscle burst, then, in his subjects, it takes those decisions or intentions about 550 ms to produce a muscle burst. Is this a realistic figure?

Some reaction-time studies provide relevant evidence. In one relevant study, the mean time

between the sounding of the go signal and the muscle burst was 231 ms. The subjects (who were watching a Libet clock) were instructed to respond as rapidly as possible to the go signal by pressing a button. If detection of the go signal produced a proximal intention to press the button, then the mean time between a subject's acquiring a proximal intention to press and the muscle burst was less than 231 ms. (Detecting a go signal takes time.) And notice how close this is to Libet's time W – his subjects' reported time of their initial awareness of something he variously describes as an intention, urge, wanting, decision, will, or wish to move (▫ 200 ms). Even without putting a great deal of weight on the exact number, one can fairly observe that if the proximal intentions to flex are acquired in Libet's studies, the finding just reported makes it look like a much better bet that they are acquired around time W than that they are acquired around ▫ 550 ms.

Someone might object that in reaction time studies of the kind described, muscle bursts and actions are not produced by proximal intentions, but by something else. It may be claimed, for example, that the combination of the subjects' conditional intentions to press whenever they detect the go signal, together with their detection of it, produces muscle bursts and pressings without the assistance of any proximal intentions to press. But if this claim is accepted, a parallel claim about Libet's studies should be taken very seriously. The parallel claim is that, in Libet's studies, the muscle bursts and actions are not produced by proximal intentions, but by the combination of the subjects' conditional intentions to flex whenever they detect a conscious proximal urge to flex, together with their detection of such an urge. Someone who makes this claim may hypothesize that the onset of the type II RPs at ▫ 550 ms is correlated with a potential cause of a conscious proximal urge to flex. Libet's findings do not contradict this hypothesis.

Window of Opportunity for Free Will

Even if Libet is wrong in thinking that the brain produces proximal intentions or decisions to flex at about ▫ 550 ms, his claim about the 100 ms window of opportunity for free will merits attention. Libet's idea is that free will can only be exercised consciously and, therefore, can only be exercised after his subjects become conscious of the proximal intentions, decisions, or urges to flex (and before it is too late to stop what is in place from generating a muscle burst). He contends that free will can be exercised only in vetoing the decision, intention, or urge of which the person has become conscious. One alternative hypothesis is that Libet's subjects exercise free will in consciously deciding to flex rather than after they become conscious of such a decision (or intention or urge). Given that Libet's findings do not justify the inference that proximal decisions to flex are made before the subjects are conscious of any such decision, they do not contradict the present hypothesis.

Are Conscious Intentions Powerless?

Libet's findings are sometimes said to support the thesis that conscious intentions and decisions do not play any role at all in producing the corresponding actions. It is claimed that they are caused by the same brain events that cause actions and that they are not themselves in the causal chain that results in action. Sometimes the following assertion is offered in support of the preceding one: The subjects' conscious proximal intentions to flex cannot be among the causes of their flexing actions because those intentions are caused by unconscious brain events. This assertion is badly misguided, as attention to the following analogous assertion shows: Burnings of fuses cannot be among the causes of explosions of firecrackers because burnings of fuses are caused by lighting of fuses. Obviously, both the lightings of its fuse and the burning of its fuse are among the causes of a firecracker exploding in normal scenarios. Other things being equal, if the fuse had not been lit – or if the lit fuse had stopped burning early – there would have been no explosion. There is no reason to believe that the more proximal causes of firecracker explosions cannot themselves have causes. Analogously, there is no reason to believe that items that are among the relatively proximal causes of flexing actions cannot themselves have causes and cannot be caused by unconscious brain events.

Might it be that conscious proximal intentions to flex are part of the causal chain leading to the flexing actions of Libet's subjects? Someone who wishes to answer this question should take a step backward and ask another: Is the brain activity

registered by, for example, the first 300 ms of type II RPs – type 300 activity, for short – as tightly connected to subsequent flexing actions as lightings of firecracker fuses are to exploding firecrackers? In fact, no one knows. In the experiments that yield Libet's type II RPs, it is the muscle burst that triggers a computer to make a record of the preceding brain activity. In the absence of a muscle burst, there is no record of that activity. So, for all anyone knows, there were many occasions on which type 300 activity occurred in Libet's subjects and there were no associated flexing actions.

Libet mentions that some subjects, encouraged to flex spontaneously, report that they sometimes aborted or suppressed conscious proximal urges to flex. As he points out, because there was no muscle activation, there was no trigger to initiate the computer's recording of any RP that may have preceded the veto. So, for all anyone knows, type 300 activity was present before the urges were suppressed.

Notice that it is urges that these subjects are said to report and suppress. Might it be that type 300 activity is a potential cause of conscious urges to flex in Libet's subjects and that some subjects make no decision about when to flex – unconsciously or otherwise – until after the conscious urge emerges? And might it be that prior to the emergence of the conscious urge, subjects have no proximal intention to flex? That our urges often are generated by processes of which we are not conscious is not at all surprising. And if we sometimes make effective decisions about whether or not to act on a conscious urge, so much the better for free will. Moreover, Libet's data do not show that the subjects have unconscious proximal intentions to flex before they have conscious proximal intentions to flex. The data do not contradict the hypothesis that what precedes these conscious intentions is a causal process that includes no unconscious proximal decisions or intentions to flex.

Unconscious and Conscious Proximal Intentions

Two points made thus far in this section merit emphasis. Libet's data do not warrant either of the following claims: (1) what happens earlier than, say, ▫ 200 ms in his subjects is causally sufficient for a muscle burst to occur at 0 ms; (2) his subjects have proximal intentions to flex before they think they do. Some related issues merit attention.

Even if claim 2 is not warranted by Libet's data, his idea that we have unconscious proximal intentions should not be lightly dismissed. Such intentions may be at work when, for example, experienced drivers flip their turn indicators to signal for turns they are about to make. In a study in which subjects are instructed to flex whenever they feel like it without also being instructed to report after flexing on when they first became aware of an intention, urge, or decision to flex, would they often be conscious of proximal intentions, urges, or decisions to flex? Might unconscious proximal intentions to flex – and, more specifically, proximal intentions of which they are never conscious – be at work in producing flexing actions in the imagined scenario?

Imagine that someone conducts the experiment just sketched and discovers (somehow) that the subjects were never or rarely conscious of proximal urges, intentions, or decisions to flex. Could it legitimately be inferred that, in Libet's own experiment, conscious urges, intentions, and decisions had no effect on the flexing actions? No. One possibility is that some of Libet's subjects treat their initial consciousness of an urge to flex as a go signal. If they do, the conscious urge seemingly has a place in the causal process that issues in the flexing. Another possibility is that some subjects treat the conscious urge as what may be called a decide signal – a signal calling for them consciously to decide right then whether to flex right away or to wait a while. If that is so, and if they consciously decide to flex and execute that decision, the conscious urge again seemingly has a place in the causal process, as does the conscious decision.

Perhaps it will be suggested that even if a subject treats a conscious urge to flex as a go or decide signal, that urge has no place in the causal process that issues in a flexing action because an unconscious brain event caused the conscious urge. But the inference here has the same form as the badly misguided assertion about conscious intention discussed earlier. An x can be among the causes of a y even if the x itself is caused. Possibly, it will be claimed that by the time the conscious urge emerges it is too late for the subject to refrain

from acting on it (something that Libet denies) and that is why the conscious urge should not be seen as part of the process at issue, even if the subjects think they are treating the urge as a go or decide signal. One way to get evidence about this is to conduct an experiment in which the subjects are instructed to flex at a time, t, unless they hear a stop signal. By varying the interval between the stop signal and the mean time of the completion of a full flex when there is no stop signal, experimenters can try to ascertain when subjects reach the point of no return. (Time t can be a designated point on a Libet clock, and brain activity can be measured backward from t.) Perhaps it will be discovered that that point is reached significantly later than time W. (Of course, some researchers and theorists worry about how seriously subjects' reports of their first awareness of a proximal urge or intention to flex – time W – should be taken.)

Vetoing and RPs

Libet offers two kinds of evidence to support his claim that subjects have time to veto proximal conscious urges to flex. One kind has already been mentioned: subjects say they did this. The other kind is generated by an experiment in which subjects are instructed to prepare to flex at a prearranged time (as indicated by a revolving spot on a clock face) but to refrain from actually flexing. Libet finds that an event-related brain potential (ERP) is produced that resembles a type I RP until about 150–250 ms before the prearranged time. Ironically, this study indicates that a kind of ERP that Libet takes to indicate the presence of an intention to flex is not actually associated with such an intention. Keep in mind that the subjects were instructed in advance not to flex, but to prepare to do so at the prearranged time and to veto this. The subjects intentionally complied with the request. They intended from the beginning not to flex at the appointed time. So what is indicated by the ERP? Presumably, not the acquisition or presence of an intention to flex; for then, at one and the same time, the subjects would have both an intention to flex at the prearranged time and an intention not to flex at that time. And how can a normal person simultaneously intend to flex at t and intend not to flex at t? (Can you intend to close this book when you finish reading this sentence, while also intending not to close it when you finish reading this sentence?) In short, it is very plausible that Libet is mistaken in describing what is vetoed in this experiment as intended motor action.

If the ERP in the veto scenario is not associated with an intention to flex at the appointed time, with what might it be associated? Perhaps a subject's wanting to comply with the instructions – including the instruction to prepare to flex at the appointed time – together with the recognition that the time is approaching, produces a growing urge to (prepare to) flex soon, a pretty reliable causal contributor to such an urge, or the motor preparedness typically associated with such an urge. Things of these kinds are potential causal contributors to the acquisition of proximal intentions to flex in other circumstances. A related possibility is suggested by another finding: the pattern of brain activity associated with imagining making a movement is very similar to the pattern associated with preparing to make a movement. The instructions given in the veto experiment would naturally elicit imagining flexing very soon, an event of a kind suitable, in the circumstances, for making a causal contribution to the emergence of a proximal urge to flex. Finally, the flattening or reversing of the ERP at about 150–250 ms before the prearranged time might indicate a consequence of the subjects' vetoing their preparation.

The veto experiment does not show that people have time to veto conscious proximal urges to flex, unless these subjects actually have such urges. But it does provide grounds for caution about what is indicated by type I RPs, which, again, are very similar to the ERP produced in the veto experiment until about 150–250 ms before time 0. Perhaps the type I RPs (until about □ 150 to □ 250 ms) are correlated with whatever the matching segment of the ERP in the veto experiment is correlated with; and, again, it is extremely unlikely that the matching segment is correlated with an intention to flex, because the subjects, in fact, intend not to flex. One should also be cautious about what is represented by the shorter, type II RPs produced in Libet's main experiment. Possibly, the first half or so of those RPs represents only a potential cause of a proximal intention to flex. Again, Libet's data

do not show that subjects have proximal intentions to flex before they think they do.

Generalizing

A simple observation about generalizing from findings like Libet's is in order. To the extent that free will is being studied in these experiments, it is being studied in the sphere of proximal decisions or intentions about matters that do not normally call for decision making – for example, exactly when to flex a wrist. Generalizing from results obtained in this domain to a view about distal decisions made about important issues in situations of a very different kind would be extremely bold, to say the least. Seemingly, a more suitable place to look for free decisions is in the sphere of distal decisions made about important practical or moral matters.

Epiphenomenalism about Conscious Intentions

Epiphenomenalism

A variety of studies show that, in some circumstances, people are not conscious of some of their actions and, in others, people believe they intentionally did things that, in fact, they did not do. This section reviews some such findings along with some related clinical findings and discusses their bearing on free will. Some background on epiphenomenalism sets the stage.

In philosophy, epiphenomenalism is the thesis that although all mental events are caused by physical events, no mental events cause any physical events. Some scientists appeal to findings of the sort to be reviewed in this section to support what they call 'epiphenomenalism' about intentions. However, what they mean by this word in this connection is not what philosophers mean by it. Brief attention to the difference will help forestall confusion.

Suppose that proximal intentions are caused by physical events, but never cause any physical events. Suppose also that the neural correlates of proximal intentions do, in fact, cause physical events – for example, bodily motions that are involved in overt intentional actions. This pair of suppositions does not contradict philosophical epiphenomenalism. But it does contradict a scientific epiphenomenalism according to which neither proximal intentions nor their neural correlates cause bodily motions. From a physicalistic neuroscientific point of view, proof that the neural correlates of proximal intentions cause physical events constitutes proof that proximal intentions cause physical events. It is philosophers who would worry about the metaphysical intricacies of the mind–body problem despite accepting the imagined proof about neural correlates, and the relevant argumentation would be distinctly philosophical. The scientific epiphenomenalism at issue in this section extends to the neural correlates of proximal intentions: the specific claim at issue is that neither proximal intentions nor their neural correlates cause physical events that intentions are thought to cause – those involved in the intended overt actions.

Findings

Some actions that people do not realize they are performing are detectable with sensitive devices. In one study, a person asked to think of an object to the left slowly moved a hand in that direction. A person asked to hide an object in a room slowly moved a hand in the direction of the object he hid when asked to think about it. And a person instructed to count a metronome's clicks made tiny hand movements that matched the rhythm.

The practice of facilitated communication was designed to help people with a disorder that hampers speech (e.g., autism or cerebral palsy) express themselves. A trained facilitator holds the hand of a client who is seated in front of a keyboard. Facilitators are supposed to help their clients communicate without influencing which keys the clients press, and there is considerable evidence that this is what many of the facilitators intended to do and believed they were doing. Often, people who had been uncommunicative apparently typed out sentences, paragraphs, or more. However, it was found that the facilitators were actually in control of what was being typed – without realizing that they were.

People suffering from a certain kind of damage to the frontal lobes display utilization behavior. For example, if an experimenter touches their hands

with a glass and a jug of milk, they may pour milk into the glass, even if they do not like milk. A person whose hands were touched with several pairs of eyeglasses puts them all on and wore several pairs at once.

In some experimental situations, people are caused to believe that they intentionally did things that they did not in fact do. In a well-known study by Daniel Wegner and Thalia Wheatley, a confederate and a subject, both of whom are wearing headphones, jointly operate a computer mouse. About fifty tiny objects are displayed on a computer monitor, and the mouse controls the movement of a cursor over the display. Subjects are asked how much they 'intended' to make a stop of the cursor on an image. When the subjects hear the name of an image in the display (e.g., 'duck') very shortly before the cursor stops on that image, they give, on average, a higher 'intended' rating to the stop than they do under other conditions, even though, in fact, the confederate is stopping the cursor on that image.

Scientific Epiphenomenalism

Studies and findings such as the ones described here are sometimes taken to support the claim that actions are never caused by conscious intentions or their neural correlates and that they are instead caused by processes of which the agents are not conscious. This is the thesis of scientific epiphenomenalism about conscious intentions. Now, it is true that the studies and findings indicate that people sometimes perform actions of which they are not conscious, sometimes do things for no good reason, and sometimes believe they intentionally did things they did not actually do. But how are these truths supposed to lead to scientific epiphenomenalism about conscious intentions?

The route that has been mapped in the empirical literature features the assumption that all actions are caused in basically the same way. If some actions are performed in the absence of conscious intentions to perform them and all actions are caused in basically the same way, that basic way includes neither conscious intentions to perform the actions at issue nor the neural correlates of such intentions. (Only existing conscious intentions have existing neural correlates.) Why then do we even have conscious intentions? Why did we evolve in such a way as to have them? Scientific epiphenomenalists about conscious intentions have replied that we have conscious intentions because they give us a sense of which of the things we do we are responsible for.

Are all actions in fact caused in basically the same way? That depends on how the expression 'basically the same way' is to be read. For example, if what is meant is simply that all actions are caused by brain events, the claim is true. But, of course, this leaves it open that some of the brain events that cause some actions are neural correlates of conscious intentions to perform actions of those kinds. What seems to be meant is something much more specific – that just as people who unknowingly move a hand slowly in the direction of an object they are thinking about are caused to do so by automatic processes of which they are unaware, all actions are caused by, and only by, such processes.

Some people may wish to reply that they can think of many occasions on which they themselves did something because they consciously decided to do it. They may say, for example, that they organized a surprise birthday party for a friend because they consciously decided to do that or that they drove to a particular restaurant because that is where they consciously decided to eat. Scientific epiphenomenalists about conscious intentions are committed to replying that these people are mistaken. And they may ask for scientific evidence that our conscious decisions to do things (or the neural correlates of those decisions) sometimes are among the causes of our doing them.

What might count as such evidence? Return to Libet's studies. Imagine a study of this kind in which subjects are explicitly instructed to make a conscious decision about when to flex a wrist and then to flex in response to that decision. Can subjects comply with this instruction? If they actually do comply, then it would seem that their conscious decisions (or their neural correlates) are among the causes of their flexing actions. A scientific epiphenomenalist about conscious decisions may reply that these subjects would have flexed even if they had unconsciously decided (or intended) to flex and that the conscious decisions (and their neural correlates) therefore played no causal role in producing the flexing actions.

There is a serious problem with this reply. The reply implicitly appeals to the following principle: If y would have happened even if x had not happened, then x is not among the causes of y. And this principle is false. Sally's mother drove her to school, and Sally arrived at school 5 min before the first bell rang. What Sally's mother did was a cause of Sally's arriving at school when she did. This is true, even though, if Sally's mother had not driven her to school, Sally's father would have done so and delivered her there at the same time.

Here, a scientific epiphenomenalist about conscious intentions may wish to appeal to an analogy of a kind familiar to philosophers of mind. The fact that Max struck a log with his red ax, thereby causing the log to split, certainly does not entail that the redness of the ax did any work in causing the log to split. And a scientific epiphenomenalist may observe that to say that a conscious decision to flex was a cause of a flexing action is not necessarily to say that the consciousness aspect of that decision (or the neural correlate of that aspect) did any work in causing the flexing action. Of course, scientific epiphenomenalists about conscious decisions are committed to claiming more than this: they must claim that the consciousness aspect of the conscious decision (and the neural correlate of that aspect) played no causal role in producing the flexing action.

If someone had painted Max's red ax green, it would have worked just as well. And perhaps an unconscious decision or intention to flex would have worked just as well as a conscious decision or intention. But imagine that Max is under strict instructions to chop wood only with red axes and that he is committed to following his instructions. Then, if his ax had been painted green, he would not have used it; and, in fact, he would have looked for a red ax. In this scenario, the redness of his ax is causally relevant to his splitting the log when he does. (If his ax had not been red, he would have looked for a red ax and split the log later, after he found one.) Similarly, in the imagined experiment, the consciousness aspect of the subjects' decisions to flex (or its neural correlates) would seem to be causally relevant to their flexing when they do. After all, their instructions call for them to flex only in response to conscious decisions, just as Max's instructions call for him to split wood only with red axes.

A scientific epiphenomenalist about conscious decisions may claim that, in the imagined experiment, the subjects' conscious decisions were not among the causes of their flexing actions, because the decisions themselves were caused by unconscious processes. However, a reader of this article who is tempted to accept this claim has failed to absorb the moral of the firecracker analogy in an earlier section. The fact that x has a cause does not entail that x is not among the causes of y.

There is evidence that some conscious distal decisions are among the causes of corresponding actions. In some experiments, subjects who have significant motivation to undertake a particular task later – for example, to exercise for at least a half hour, 1 day next week – are divided into two groups. Half are instructed to decide during the experiment – consciously, of course – on a specific place and time to undertake the task, and half are not given this instruction. In a wide range of scenarios across numerous experiments, subjects given this instruction are significantly more likely to undertake the task.

Someone might claim that although these conscious distal decisions do make a difference, corresponding unconscious distal decisions or intentions would be just as effective and therefore the consciousness aspect of the conscious distal decisions does no work at all. How plausible is this? How would the imagined unconscious distal decisions or intentions help to generate corresponding actions days or weeks later? Seemingly, not as a consequence of agents' consciously remembering them when the time for their execution is near. Proponents of the claim at issue should specify a process that links distal decisions or intentions of which agents are never conscious to corresponding intentional actions and produce evidence that the specified process is not a fiction. Once that is done, they can turn their attention to supporting their assertion of equal effectiveness.

In light of what you have just read, it might have occurred to you that, in some situations, deciding in advance on specific places and times to undertake certain tasks would improve your success rate at meeting due dates for assignments at school or work, for example. So what should you do? Should you sit back and hope that you will unconsciously make such decisions when they would be useful?

Would it be better consciously to settle on a policy of making relevant distal decisions when doing so is likely to help you meet deadlines; consciously to think about where and when to perform the desired tasks when you believe such thoughts would be productive; and consciously to settle on particular places and times to execute relevant tasks? The answer seems obvious.

Conclusion

Whether anyone has free will depends on what 'free will' means. Some people who are willing to grant that our intentions sometimes make a causal contribution to our actions may think that if all of our decisions and intentions have causes, then we lack free will and never perform free actions. Such people should try to explain why compatibilist and event-causal libertarian views about the meaning of 'free action' and 'free will' are false. According to views of both kinds, all free actions are caused, as are the causes of free actions.

Theoretical work on various implicit or explicit conceptions of free will that guide scientific studies may show that some of the conceptions are self-contradictory, that others are hopelessly magical or mysterious, and that yet others suggest potentially fruitful research programs. One would expect most scientists with an experimental interest in free will to be attracted to conceptions of the third kind. Empirical research guided by such conceptions of free will may help illuminate the place of consciousness in free action. The scientific work reviewed here sheds light on the production of actions, and what it reveals about action–production is compatible with its being true that we have free will and sometimes perform free actions.

See also: Automaticity and Consciousness; Brain Basis of Voluntary Control; History of Philosophical Theories of Consciousness; Intentionality and Consciousness; Neuroscience of Volition and Action.

Suggested Readings

Baer J, Kaufman J, and Baumeister R (eds.) (2008) Are We Free? Psychology and Free Will. New York: Oxford University Press.

Bargh J and Ferguson M (2000) Beyond behaviorism: On the automaticity of higher mental processes. Psychological Bulletin 126: 925–945.

Clarke R (2003) Libertarian Accounts of Free Will. Oxford: Oxford University Press.

Fischer J (1994) The Metaphysics of Free Will. Cambridge, MA: Blackwell.

Gollwitzer P and Sheeran P (2006) Implementation intentions and goal achievement: A meta-analysis of effects and processes. Advances in Experimental Social Psychology 38: 249–268.

Haggard P and Magno E (1999) Localising awareness of action with transcranial magnetic stimulation. Experimental Brain Research 127: 102–107.

Kane R (1996) The Significance of Free Will. New York: Oxford University Press.

Libet B (2004) Mind Time. Cambridge, MA: Harvard University Press.

Mele A (2006) Free Will and Luck. New York: Oxford University Press.

Mele A (2009) Effective Intentions: The Power of Conscious Will. New York: Oxford University Press.

Pereboom D (2001) Living Without Free Will. Cambridge, MA: Cambridge University Press.

Pockett S, Banks W, and Gallagher S (eds.) (2006) Does Consciousness Cause Behavior? An Investigation of the Nature of Volition. Cambridge, MA: MIT Press.

Spence S and Frith C (1999) Towards a functional anatomy of volition. Journal of Consciousness Studies 6: 11–29.

Wegner D (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

Wegner D and Wheatley T (1999) Apparent mental causation: Sources of the experience of will. American Psychologist 54: 480–491.

Biographical Sketch

Alfred R. Mele is the William H. and Lucyle T. Werkmeister professor of philosophy at Florida State University. He is the author of *Irrationality* (Oxford 1987), *Springs of Action* (Oxford 1992), *Autonomous Agents* (Oxford 1995), *Self-Deception Unmasked* (Princeton 2001), *Motivation and Agency* (Oxford 2003), *Free Will and Luck* (Oxford 2006), and *Effective Intentions* (Oxford 2009). He also is the editor of *The Philosophy of Action* (Oxford 1997), a coeditor (with John Heil) of *Mental Causation* (Oxford 1993), a coeditor (with Piers Rawling) of *The Oxford Handbook of Rationality* (Oxford 2004), and a coeditor (with Mark Timmons and John Greco) of *Rationality and the Good* (Oxford 2007).

Functions of Consciousness

A K Seth, University of Sussex, Brighton, UK

Glossary

Conscious content – The continuously changing phenomenal content (e.g., 'qualia' such as redness and warmth) and intentional content (e.g., conscious knowledge) present to varying degrees at nonzero conscious levels.

Conscious inessentialism – The view that all behaviors and cognitive processes can in principle be carried out in the absence of consciousness.

Conscious level – Applies to a whole organism and refers to a scale ranging from total unconsciousness to vivid wakefulness. A 'conscious organism' is one that is capable of having nonzero conscious levels.

Epiphenomenalism – The view that consciousness does not play any causal role in neural or cognitive operations.

Higher-order consciousness – Consciousness of consciousness, also referred to as 'reflective' or 'meta' consciousness and related to 'access' consciousness.

Integration consensus – The recurring idea that consciousness serves to integrate otherwise independent neural and cognitive processes.

Primary consciousness – The presence of a multimodal scene composed of perceptual and motor events, also referred to as 'sensory' or 'phenomenal' consciousness.

Teleofunction – The causal effect(s) of a property that accounts for its plausible origin via natural selection; the teleofunction of X is the effect that explains why it exists. Teleofunctions are distinct from 'Cummins functions,' which refer to explanatorily salient causal effects where salience is determined in the context of a larger system.

Zombie – A creature that is behaviorally indistinguishable from a normal (conscious) human, but is entirely without consciousness. Zombies embody the thesis of conscious inessentialism.

Introduction

A major challenge for the successful naturalization of consciousness lies in locating its biological function, or functions. Although common sense suggests that conscious experience has many important functional roles in our lives, experiments and theoretical arguments challenge these everyday intuitions. Many human behaviors can occur in the absence of consciousness, and the natural world contains many creatures capable of engaging in complex behaviors, at least some of which may be doing so entirely without consciousness (e.g., mollusks, microorganisms). Consciousness is a real phenomenon whether functional or not; however, without any defensible function its scientific study is rendered even more difficult than already supposed.

On some theories, consciousness has no function at all. On others, the function of consciousness is associated with capacities such as rationality and volition. Experimental evidence, however, challenges many of these intuitive ideas. Greater consistency with evidence is found for theories that invoke an integrative function for consciousness, yielding flexible behavior in the face of novelty, though these theories, too, face difficulties. Finally, some theories posit that consciousness, as a constellation concept, may have multiple functions.

Does Consciousness Have a Function?

Asking about the function of consciousness requires first considering the possibility that consciousness has no function. There are two sorts of arguments along these lines: conscious inessentialism (CI) and epiphenomenalism (EP).

Conscious Inessentialism

This is the view that Owen Flanagan describes thus: for any intelligent activity i, performed in cognitive domain d, even if WE do i with conscious accompaniments, i can in principle be done without these conscious accompaniments (see Suggested readings). CI is a radical proposition and one which certainly challenges common sense. However, since many biological functions can be carried out by a variety of mechanisms, CI is not obviously false.

Some behaviors, for example introspecting, seem obviously to require consciousness. If introspection is defined as an explicit verbal report of the content of conscious experiences, then introspection requires consciousness. However, examples like these are not strong challenges to CI because they represent cases in which consciousness is constitutively essential rather than causally essential. To the extent that introspection is defined in terms of consciousness, the notion of unconscious introspection is simply incoherent.

If CI is true then zombies are possible. Zombies in philosophy of mind are creatures that are behaviorally indistinguishable from normal (conscious) humans, but who are entirely without consciousness; there is no what it is like to be a zombie. It has been argued that if zombies are conceivable in worlds that share our physical laws, then we can conclude that even in our world, consciousness does not causally influence the physical events responsible for human behavior. But this inference is false. From the statement that consciousness is inessential for a particular behavior, it does not follow that consciousness is causally ineffectual for that behavior in those cases when it is present.

A more stringent form of the zombie argument requires that zombies not only display identical behaviors to normal humans but that they also have complete neurophysiological equivalence. However, such brainy zombies may be conceivable only if one allows nonnaturalistic explanations of consciousness. By all naturalistic accounts it seems impossible for there to exist two brains in exactly the same physical state but only one of which is conscious.

Epiphenomenalism

This is the view famously expressed by Thomas Huxley more than a hundred years ago. According to the epiphenomenalist (EP) suspicion, consciousness exists but does not play any causal role in neural or cognitive operations.

Even if CI is false, this does not by itself imply that EP is also false. It is conceivable that a behavior/activity requires the sort of brain activity that inevitably gives rise to consciousness, without consciousness itself playing a causal role in generating that behavior. Conversely, even if CI is true it does not necessarily follow that EP is true. Even if consciousness is inessential for a particular behavior it does not follow that consciousness has no causal role in generating the behavior in those instances in which consciousness is present. Just like CI, EP may be false, but it is not obviously false.

There are two varieties of EP. Metaphysical EP is the view that consciousness is entirely without causal powers. Since having causal powers is central to at least one concept of what it is to be real, metaphysical EP may imply that consciousness is not real. Biological EP is the view that consciousness may have causal effects in the world, but that these causal effects are not and never were involved in the reproductive success of conscious organisms. A good example is the thudding noise generated by the heart; this is indeed a causal effect of hearts but one that probably had nothing to do with selective advantage.

Empirical Evidence for CI and EP

Because neither EP nor CI can easily be dismissed nor proven on logical grounds alone, it is useful to consider the relevant experimental evidence. Experimental evidence in favor of CI would come from cases where behaviors for which

consciousness has been assumed necessary are shown to be exhibited in the absence of consciousness. Experimental evidence in favor of EP would come from cases in which the causal link between consciousness and behavior is challenged.

One well-known experiment that seems at first blush to satisfy both the above criteria is the famous readiness potential study of Benjamin Libet, replicated and extended in various ways by Patrick Haggard and colleagues. In the original Libet study, subjects were instructed to flex their right hand whenever they felt like it, as well as pay close introspective attention to the instant of onset of the decision to perform each such act and to the correlated position of a revolving spot on a clock face (indicating clock time). Throughout the experiment electroencephalographic (EEG) potentials were recorded from above the surface of the region of motor cortex implicated in hand movement. Strikingly, Libet found that the onset of easily recognizable motifs in the EEG traces (the readiness potentials, see Figure 1) reliably preceded awareness of the intention to act spontaneously, by around 350 ms.

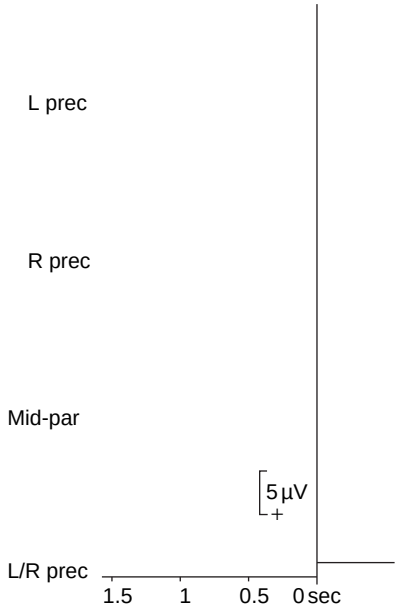


Figure 1 A typical recording of a readiness potential, measured from four different regions of the cortical surface. Reproduced from Kornhuber HH and Deecke L (1965) Changes in the brain potential in voluntary movements and passive movements in man: Readiness potential and reafferent potentials. Pflugers Arch Gesamte Physiol Menschen Tiere 284: 1–17.

Libet s findings have often been used to cast doubt on consciousness having a causal role in generating voluntary activity, thus providing evidence in favor of EP. As Libet asks, if the brain can initiate a voluntary act before the appearance of a conscious intention . . . is there any role for conscious function? Moreover, if apparently voluntary activity is determined prior to awareness of the intention to act, it seems plausible that this sort of action could in principle be carried out in the absence of consciousness, thus providing support for CI. But, even in the worst case that both of the above implications are justified, EP and CI will have been demonstrated only with respect to the specific sort of voluntary action investigated by Libet. From the statement that consciousness is inessential and/or epiphenomenal with regard to a certain sort of voluntary action it does not follow that consciousness is inessential and/or epiphenomenal in general.

Furthermore, as Libet notes, conscious causal powers are arguably preserved if consciousness is assumed to have a veto power over unconsciously initiated actions (free won t rather than free will). However, in this case there now arises the counterproblem of the possible existence of unconscious precursors to the conscious intention to veto an act. Finally, there is an important methodological issue with the Libet experiment and other similar studies: While the precise timing of the subjects hand flexions may be unpredictable, subjects are asked at the outset to make a conscious effort to let flexion occur spontaneously. Without having consciously assimilated the instructions it is unlikely that apparently spontaneous hand flexions would occur at all.

The ongoing brouhaha surrounding the Libet studies is troubling for many theorists. Why is it surprising that conscious processes are causally preceded by unconscious events? It seems that this could only be surprising given a Cartesian intuition that conscious intentions are prime movers, which are themselves without cause, or perhaps capable of being caused only by other conscious intentions. Flanagan offers a neat summary (see Suggested readings):

> Libet s results, far from offering solace to the suspicious epiphenomenalist, are precisely the sort of results one

would expect if one believes that conscious processes are subserved by nonconscious processes, and that conscious processes play variable but significant causal roles at various points in different cognitive domains.

There is plenty of other evidence relevant to EP and CI. Studies in diverse fields including cognitive psychology, social psychology, and neuropsychology have shown that, contrary to our everyday intuitions, many complex human behaviors can occur in the absence of consciousness. In fact, some activities for which consciousness is commonly assumed to be useful can in fact be performed better unconsciously (see Rational action section). While it is beyond the present scope to critically analyze all of this evidence, a good selection of it will appear in what follows.

Summary

Although both CI and EP are likely to be false it is not easy to establish their falsity, whether by experiment or by theoretical argument. And even if CI and/or EP turn out to be true, there remain deep questions of why conscious experiences appear to be uniquely associated with neural activity and not (as far as we know) with other physical processes (e.g., fermentation, photosynthesis), and why conscious contents almost always reflect functional roles (e.g., pains feel bad, sex feels good). Nevertheless, both CI and EP serve as useful null hypotheses to keep in mind in the light of more positive proposals, for the function(s) of consciousness.

Functional Analysis

To ask about the function of consciousness is to make at least two related inquiries: (1) Why did consciousness evolve? (2) What does consciousness do?

Why Did Consciousness Evolve?

For many people an adequate functional explanation of a biological feature is one that plausibly accounts for its origin by natural selection. In this view, the function(s) of consciousness are the same as those features of consciousness that explain why it came to be present and maintained in certain organisms: The function of X is the effect that X has, which explains why it is there. This interpretation of function can be called a proper function, or equivalently, a teleofunction.

Coming up with a plausible teleofunction for consciousness is challenging for several reasons. First, in complex highly interactive systems such as the brain, it can be very difficult to make a direct connection from any effect of a part to a selective advantage enjoyed by the whole. Second, the function a biological feature has in the present is not necessarily the function (if any) for which it was selected during evolutionary history. Our brains presumably were not selected for reading ability, yet reading is an important brain-dependent function for contemporary humans. Third, not all present-day biological features exist as a result of natural selection. For example, it is unlikely that the color of blood or the structure of the human chin were driven by natural selection (i.e., these features are not traits).

A further problem lies in coming up with the right sort of evidence that can turn a how possibly account into a how actually account. That this is difficult for adaptationist explanations in general has led to the famous criticism by Stephen Jay Gould and Richard Lewontin that such explanations are often little more than just so stories. For consciousness in particular there is the additional hurdle of widespread skepticism that empirical evidence has anything to do with consciousness. This skepticism derives from the idea that consciousness, as a subjective phenomenon, does not directly engage with objective evidence. However, as John Searle has emphasized, a science of consciousness requires only that we be epistemologically objective, a position which is entirely consistent with the correct characterization of consciousness as ontologically subjective.

According to Robert Brandon, the sorts of evidence that are required for establishing teleofunctionality are (1) evidence that selection has occurred (i.e., fossil evidence or other experimental evidence); (2) an ecological explanation of relative adaptedness; (3) evidence that the traits in question are heritable; (4) information about population structure; and (5) phylogenetic information about trait polarity (i.e., evidence that conscious organisms

evolved from nonconscious organisms and not vice versa). These criteria are not easy to satisfy. For example, fossil evidence for consciousness is difficult to imagine and the relevant experiments are hard to design and likely to be unethical; direct evidence for heritability is also hard to come by, and population structures in proposed adaptive environments for consciousness are mostly left unspecified. In short, coming up with a solid adaptationist account of the evolution of consciousness is difficult and requires going well beyond establishing what consciousness does for an organism.

What Does Consciousness Do?

Instead of asking why consciousness evolved, one can ask instead what causal effects consciousness has with regard to present-day brains, bodies, and behaviors. This approach attempts to isolate salient causal effects from among a multiplicity of effects that a given biological feature might have. A useful way to think about this is to consider the role played by the functionally characterized thing in how some larger system, of which the functionally characterized thing is a part, is able to exhibit a more complex capacity or behavior. For example, hearts have the function of pumping blood because this effect helps explain the capacity of the body to achieve circulation of oxygen. This concept of function can be called a Cummins-function after its originator, Robert Cummins (another equivalent term is causal role function).

Of course in many cases, including the example just given, a teleological interpretation may be granted to the larger capacity (i.e., achieving oxygen circulation is likely to have been strongly selected for), but this interpretation does not necessarily transfer to the functionally characterized thing itself. In other words, characterizing a salient effect as a Cummins-function can avoid the mereological (part whole relation) difficulties that sometimes attend ascriptions of teleofunctionality.

Both Cummins-functions and teleofunctions pick out functions from mere physical effects in virtue of explanatory salience; the thudding of the heart is neither teleofunctional nor Cummins-functional because it lacks explanatory salience. Also, proposing a Cummins-function for consciousness does not exclude a teleological role, it simply allows for the possibility that a readily identifiable teleological role may not exist. Therefore, on a Cummins-function view, one can evaluate the explanatory salience of consciousness with regard to the behavior of the organism as a whole, without necessarily making the claim that consciousness was picked out by natural selection precisely in virtue of these explanatorily salient effects.

Functions and Functionalism

It is important to distinguish the question of the function(s) of consciousness from the philosophical position of functionalism, one of the most important developments in twentieth century analytic philosophy. The core thesis of functionalism, as pursued by Jerry Fodor, Hilary Putnam, and others, is that mental states are second-order properties constituted by their causal relations to one another and to sensory inputs and motor outputs. For example, the mental state of being in pain is fully characterized by dispositions to say ouch, to wonder whether one is unwell, to take an aspirin, and so on. Functionalism remains a controversial position, most obviously because it implies that conscious states can be implemented in arbitrary physical systems and thus is seen by some to be too liberal.

Within a functionalist framework the question of the function of consciousness is obviously important. However, accepting functionalism does not by itself negate CI or EP because causal relations need not always have explanatory salience (in the sense of being Cummins functions). Nor is it necessary to subscribe to functionalism in order to inquire about the functions of consciousness. It is conceivable that consciousness could have explanatorily salient causal effects in virtue of intrinsic properties of its material substrates, a physicalist position that is contrary to functionalism.

Summary

The function(s) of consciousness can be considered in terms of teleological significance or in terms of nonteleological explanatory salience. While it is tempting to look for adaptationist explanations for the function of consciousness,

such explanations may be difficult to justify, as are adaptationist explanations in general.

Consciousness

It is often pointed out that consciousness is not a unitary phenomenon. Asking about the function of consciousness requires clarifying the concept of consciousness itself.

Conscious Level versus Conscious Content

A first important distinction is between an organism being a conscious organism and mental content being conscious mental content. A conscious organism is one which is capable of having conscious mental content. That is, a conscious organism has at any given time a particular level of consciousness. In humans, these levels range from complete unconsciousness (death, coma, general anesthesia) to full, awake, and alert consciousness. Conscious content describes the continually changing phenomenal content (e.g., qualia such as redness and warmth) and intentional content (e.g., explicitly held beliefs) that is present for conscious organisms at nonzero conscious levels. (There are many other ways to carve these distinctions. For example, David Rosenthal distinguishes being conscious of something which he calls transitive consciousness, from both creature consciousness (similar to conscious level), and state consciousness, which is the property of a mental state that makes it conscious.)

Many types of mental content can be either conscious or unconscious. For example, during normal waking consciousness we can either hold the implicit (unconscious) belief that the sun will rise in the morning, or we can hold this belief explicitly (consciously). There is good evidence that the same is true for many other types of mental content. For example, masked priming experiments and the phenomenon of blindsight suggest that we can have unconscious perceptual content. It is also widely accepted that desires, emotions, and even intentions can occur unconsciously, although whether linguistic thoughts can exist without being conscious is less clear.

As Rosenthal has pointed out, a consequence of the distinction between conscious level and conscious content is that the function that mental content has in virtue of being conscious cannot be inferred from the function of the organism s being conscious. In other words, the ability to have conscious content may serve a different (although probably overlapping) set of functions for an organism, than the fact that a particular sort of conscious content is present.

Where possible it is also useful to distinguish the function that mental content has in virtue of being conscious from the function that content would have even when unconscious. This can be difficult to do because a satisfactory distinction would require selectively carving off those mechanisms relevant to mental content being conscious mental content from those mechanisms underlying all other causal effects of that content. In a complex biological system it is not likely that such selective dismemberment will be possible. For example, phenomena such as blindsight may suggest that perceptual content can have function even when unconscious, however, the visually guided behavior of a blindsight patient is usually worse than that of a healthy control. Moreover, it can be argued that certain types of mental content, such as inner speech, cannot exist without being conscious.

Primary Consciousness versus Higher-Order Consciousness

A second key distinction is between primary consciousness and higher-order consciousness. Primary consciousness reflects the presence of a multimodal scene composed of perceptual and motor events. At its core, primary consciousness refers to the presence of a world; there is something like it is to be a primary conscious organism. Higher-order consciousness (HOC) reflects the observation that we (humans) are not only conscious, we are also conscious of being conscious. HOC involves the referral of the contents of primary consciousness to interpretative processes including a sense of self and, in more advanced forms, the ability to explicitly construct past and future scenes. While in humans these two forms of consciousness almost always go together (with possible exceptions in certain dreamlike or meditative

states), it is conceivable that primary consciousness could exist independently of HOC. Indeed, this may be case in many animals and perhaps as well in newborn infants.

The distinction between primary and higher-order consciousness is similar, but not identical to that made by Ned Block between phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness). Both HOC and A-consciousness involve metacognitive access, but HOC requires in addition that conscious contents are themselves higher order whereas A-consciousness does not. Importantly, it is this shared feature of metacognitive access that allows us to verbally report conscious content, and which therefore underwrites most experimental methods for studying consciousness, since it is only through explicit behavioral report that we (presently) are able to be sure whether a subject is conscious and if so know what she is conscious of. This also means that reported absence of conscious content could be due either to the absence of this content, or alternatively, to disruption of conscious access to this content that allows its report.

An influential philosophical framework that requires subtle interpretation in this context is Rosenthal s higher-order thought (HOT) theory. According to HOT theory a mental state is a conscious mental state in virtue of the existence of a HOT, distinct from that state, to the effect that one is in that state. That is, HOTs are not only essential for allowing report of conscious mental content but they are also constitutive of that content, and any mental state can in principle occur without being conscious. HOTs themselves are rarely experienced consciously (i.e., having a HOT does not imply the existence of corresponding HOC content): the theory holds that to do so would require a corresponding third-order thought, to the effect that one is having the second-order thought.

According to the HOT theory, mental content has the same causal effects whether it is conscious or not. This is because the difference between consciousness and unconsciousness of mental content is due not to any property intrinsic to that content, but instead is due to the presence or absence of the corresponding HOT. The function of consciousness therefore pertains to the content of these HOTs, which may well be distinct from that of their first-order target(s). As Rosenthal argues,

> [the] additional function that is due specifically to a first-order state s being conscious is simply the function of the higher-order state in virtue of which that first-order state is conscious. And it is natural to expect that the function of the higher-order state would be minimal relative to the function of the first-order state it is about. (See Suggested readings.)

By asserting that any mental state can occur without being conscious, HOT theory avoids an apparent obstacle for other theories in which some states cannot occur without being conscious. This obstacle exists to the extent that it is difficult to distinguish between the function a state has in virtue of being conscious from the function(s) it may have in virtue of other of its psychological or neurophysiological properties. However, if a theory can explain why a particular state entails that it is conscious (i.e., why it could not occur without being conscious), then an explanation of its function is also an explanation of its function in virtue of being conscious.

Summary

There is a difference between an organism s being conscious, and the conscious contents that such an organism has. One set of functions cannot be inferred from the other. Much mental content can be either conscious or unconscious, and the function of conscious content ought, where possible, to be distinguished from the function that content would have independently of being conscious. Finally, consciousness is not a unitary phenomenon, differentiating most importantly into primary (sensory) and higher-order (metacognitive) varieties.

Volition and Rationality

We are now in a position to outline a number of possibilities for the functional roles consciousness may play in humans, and possibly in other species as well. We begin with some intuitively appealing ideas.

Volition

The notion that the function of consciousness is to initiate and control voluntary action has enormous appeal: We consciously think about doing X and then we do X. James ideomotor theory closely follows this intuition by suggesting that actions are generated by having a thought about the action. For example, getting out of bed on a chilly morning is caused by a conscious image of being out of bed accompanied by representations of the day s intended activities. Together, this mental content displaces any thoughts or images of staying-in-bed, and the getting-out-of-bed mental content is then translated into the appropriate motor commands.

A function for consciousness in volition is plausibly teleological and may correspond with both primary consciousness and HOC. However, the added value that volitional mental content has in virtue of being conscious is less clear. An alternative explanation of what happens on a chilly morning is that both the actual getting-out-of-bed and the volitional conscious experience of the intention-to-get-out-of-bed are caused by a common set of unconscious processes, perhaps including unconscious intentions. These unconscious processes can be attributed with volitional content precisely because they give rise to volitional conscious content as well as to actions that appear from an external perspective, to be voluntary. On this view, the experience of volition is a conscious experience like any other and does not have any additional causal powers in virtue of its volitional content.

This interpretation has gathered support on both theoretical and empirical grounds. James himself left room for the possibility that voluntary actions may have unconscious causal precedents by saying that action-related thoughts have a tendency to cause the corresponding action; that is, a thought is not always guaranteed to produce the corresponding action. James, like Libet, also considered that conscious vetos (acts of express fiat) could intervene to stop a given action from occurring. More recently, Daniel Wegner has stated the strong position that conscious will is an illusion and that we experience volition only when mental content is inferred, rightly or wrongly, to have produced the corresponding physical action. According to Wegner s theory of apparent mental causation we tend to make inferences resulting in an experience of volition only when the corresponding mental content satisfies the constraints of (1) primacy (the content immediately preceded the action), (2) consistency (the content corresponds to the action), and (3) exclusivity (there is no other plausible causal factor) (see Figure 2).

If conscious will is illusory it should be possible for voluntary actions to occur in the absence of experiences of volition. Several neurological and psychological disorders show these features. For example, patients with alien hand syndrome report that one of their hands seems to be outside of their voluntary control, performing apparently voluntary complex action sequences without any accompanying experience of voluntary control. Schizophrenic auditory hallucinations also produce anomalous experiences of will in which the patients own volitional thoughts are attributed to others, or even to television sets.

Experiments in normal subjects have also addressed the issue of volition. On one hand the Libet studies attempt to show that voluntary

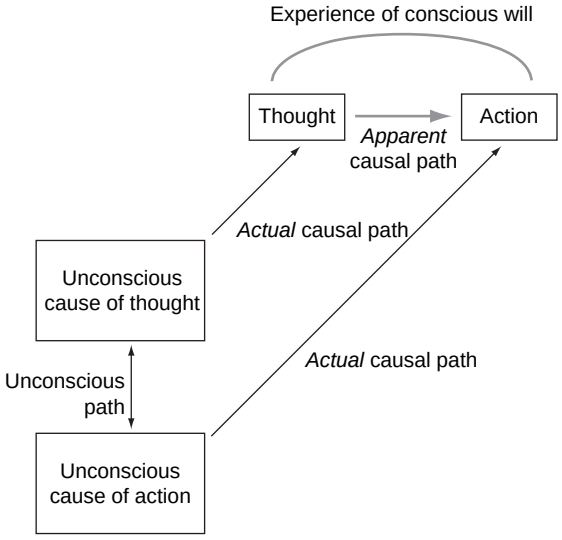


Figure 2 Wegner distinguishes two causal pathways, one leading from unconscious causes of actions to actions, and another leading from unconscious action-related mental content to conscious experience of that mental content. That conscious mental content becomes an experience of volitional mental content when an inference is made of an apparent causal pathway from the mental content to the action, which will happen when the constraints of primacy, consistency, and exclusivity are satisfied. Adapted from Wegner D (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

actions do not involve consciousness. On the other, several lines of evidence suggest that consciousness is important for controlling voluntary actions, but as with the Libet studies, their interpretation is rarely straightforward. In Larry Jacoby s exclusion task subjects are presented with a target word (e.g., reason) and then with the stem of a word (e.g., rea). The task is to complete the stem into any word except the target (e.g., realize). When subjects report conscious perception of the target they are usually able to perform this task successfully, but when they do not report conscious perception of the target (e.g., if it is presented very rapidly and/or followed by a masking stimulus) then subjects tend to fail, erroneously completing the word-stem with the target. These results are often taken to show a link between voluntary action and consciousness, since subjects are only successful at excluding words that they were conscious of having seen. But subjects who have not consciously seen the word are conscious of having not seen it; therefore, completing the word according to the not-consciously-seen (but still perceptually processed) word is still consistent with the task instructions. Also, studies with hypnotic subjects amply demonstrate that many types of exclusions can occur in the absence of a reportable conscious experience of the excluded target.

Perhaps more compelling are experiments showing that blindsight subjects are unable to initiate voluntary actions with respect to blind-field stimuli or to suppress corrective movements during rapid pointing. Conversely, normally unconscious actions can be brought under voluntary control by biofeedback training, but only when subjects are given conscious feedback of the process. In both cases, the essential link seems to be that consciousness of a stimulus is necessary for both voluntary actions, and for the conscious experience of volition with respect to that stimulus. But in neither case does it follow that the conscious experience of volition causes the voluntary action.

Rational Action

The association of consciousness with rationality is equally intuitive as the association with voluntary action. Introspection suggests strongly that conscious deliberation can supply rational responses in situations in which unconscious, automatic reactions may fail, and philosophers and scientists from Descartes and Locke onward have emphasized the benefits of conscious deliberation in decision making. Block, for instance, holds that mental content is access conscious (A-conscious) if it is poised for use as a premise in reasoning,. . . [and] for [the] rational control of action and speech. Cognitive and neural theories of consciousness that stress global access and integration of conscious content are also aligned with rationality as a function, since on this view conscious contents will be accessible to a wide range of potentially rational cognitive processes (see Flexible action and the integration consensus section). Along similar lines Christof Koch has argued that the function of consciousness is to provide an executive summary to those parts of the brain involved in planning and deliberation.

A rational action function for consciousness seems to associate more with HOC than with primary consciousness, and as with volition is plausibly teleological. But again, empirical evidence weighs against there being a direct connection: Not all conscious thinking is rational, and not all rational behavior is conscious. The dissociation between rational thinking and rational acting is strikingly illustrated by episodes in which subjects provide false (confabulatory) rationalizations for the cause of their actions. The classic experiments of Richard Nisbett and Timothy Wilson in the 1970s showed that people s descriptions of their own reasoning processes are often inaccurate and seem to be modeled after logical-sounding, idealized reasoning processes rather than the process actually used.

The choice blindness paradigm, recently developed by Petter Johansson and colleagues, extends Nisbett and Wilson s results in a novel direction. In this paradigm, subjects are shown pairs of pictures of female faces, asked to select the most attractive and then to describe why they made each selection. Unknown to the subjects, on some trials the pictures are covertly switched, with the switch carried out *after* a choice was made, but *before* introspective feedback is sought. Remarkably, on some of these switch trials the subjects fail to detect the switch, but nevertheless offer a plausible account why they chose a particular face, even though they had actually selected the other face. As with Nisbett and Wilson,

these results show a clear divergence between conscious thinking and rational action.

The association between consciousness and rationality is further challenged by results showing that certain decisions can be more rational when subjects are not consciously aware of the decision-making process. For example, the psychologist Ap Dijksterhuis hypothesized based on the limited capacity of consciousness that unconscious deciding could have an advantage over conscious deciding in complex situations involving trade-offs among multiple factors. In support of this hypothesis, he found that subjects deciding among complex alternatives (e.g., different makes of car) performed worse when they were encouraged to mull consciously over the different options, than when they were prevented from engaging in conscious deliberation. This performance difference was apparent both by objective criteria (i.e., when there really was a best car) and by subjective criteria (i.e., post-decision satisfaction), and it disappeared when the decision problem was simple (e.g., different types of towel).

These results resonate with neurological evidence that an excess of rationality can be maladaptive. Antonio Damasio has described numerous examples of patients with damage to prefrontal cortex who are hampered in their everyday life by an inability to make decisions effectively and efficiently. In one example, a patient endlessly enumerated the pros and cons of two suggested appointment dates, apparently paying as much attention to this trivial task as might be appropriate for deciding between two mortgages. Damasio s suggestion is that normal consciousness aids rational decision making, not by facilitating rational thinking *per se*, but by biasing rational deciding in certain ways, in order to reduce the space of possible options and the time and effort required to decide among them. This bias and channeling occurs as a result of the integration of emotional valence into conscious content related to the decision options.

Summary

Consciousness should not be excluded from functional roles in volition and rationality. Consciousness may be necessary for many aspects of volition; conscious decision making and conscious actions in general are often rational, and in many cases conscious deliberation may serve rational ends. But it is not the experience of volition that causes the voluntary action and a wealth of experiments show that consciousness is neither necessary nor sufficient for rational action.

Flexible Action and the Integration Consensus

A recurring idea in theories of consciousness is that consciousness serves to integrate otherwise independent neural and cognitive processes. This integration consensus, which can be traced back at least to Charles Sherrington in the early 1900s, has been expressed most forcefully in the cognitive context by Bernard Baars global workspace (GW) theory, and in the neurophysiological context by Gerald Edelman and Giulio Tononi. (The term integration consensus is due to Ezequiel Morsella.) Most participants within the integration consensus see consciousness as a highly functional adaptation, particularly with respect to enabling flexible, context-dependent behavior. Functional proposals based on the integration consensus are explicitly teleological and associate with both primary consciousness and with HOC.

Global Workspace Theory

The cornerstone of the GW theory is that consciousness involves a central resource (the GW), which enables distribution of signals among numerous otherwise informationally encapsulated and functionally independent specialized processors. GW theory states that mental content becomes conscious mental content when it gains access to the GW such that it can influence a large part of the brain, and a correspondingly wide range of behaviors (in this sense GW theory embodies Block s notion of A-consciousness). This principle invites a televisual or theatrical metaphor: A process becomes conscious when it is broadcast widely, or when it comes on stage in the presence of a large audience, but not when it remains private. A key aspect of GW theory is that conscious contents unfold in an integrated, serial manner but are the product of massively parallel activity

among specialized processors. The integrated states of the GW follow each other in a meaningful but complex progression, which depends on multiple separate processes, each of which might have something of value to add to the ongoing constitution of the GW.

According to Baars the premier function of conscious integration is to provide behavioral flexibility. There is good evidence that unconscious processes can be extremely rapid and free from the sorts of capacity limits that attend conscious processing, but that they are also more fixed and predetermined. GW theory accounts for this evidence by suggesting that in familiar situations automatic unconscious processors reel off canned responses with high efficiency, but in novel situations the broadcast of multimodal signals within the GW can mediate the production of novel responses.

The functional role of consciousness within GW theory has been refined in several directions. Baars notes that the interplay of serial and parallel processes in a GW architecture can subserve analogical reasoning because the integrated content of the GW can stimulate a wide range of unconscious processes to locate analogical content. Baars together with Murray Shanahan have further argued that this process can provide a solution to the notorious frame problem, that is, the problem of dealing effectively with signals from potentially any domain without having to explicitly sift the relevant from the irrelevant. This argument is based on the reasoning that any solution to the frame problem must involve an informationally unencapsulated process, that is, one that can draw on information from potentially anywhere, and that analogy formation exemplifies information unencapsulation. By the same token, an informationally unencapsulated architecture can underwrite behavioral flexibility by allowing the effective and efficient integration of multiple cognitive processes to produce something new.

A different and explicitly teleological take on the functional utility of a GW architecture is that it allows an organism to rely more on mental simulation and internal evaluation to select actions, reducing both energy expenditure and risk. Stanislas Dehaene and colleagues argue that the GW, by allowing a wide range of cognitive processes to bear on action selection, participates in an evolutionary trend toward increasing internalization of environmental representations, whose main advantage is the freeing of the organism from its immediate environment. This theme has been pursued further by Germund Hesslow and Antti Revonsuo, who argue that internal simulation of behavior and perception can explain the appearance in consciousness of an inner world.

As with most other participants in the integration consensus, the GW theory is vulnerable to both EP and CI as emphasized by the computational models that embody the principles and functionality of the GW architecture, but for which the attribution of consciousness seems implausible; similarly, in the HOT theory it may be that the function provided by access to the GW may be independent of the state being a conscious state. Also, integration for flexibility is arguably a very basic function of nervous system activity that is substantially present even in inverterbrates, although whether a GW architecture is present in these cases remains an open question.

Skill Acquisition and Learning

Flexible control is needed especially during acquisition of new skills. Many behavioral observations have indicated that acquiring a new skill requires conscious attention during the initial phases, but as learning progresses the execution of the skill becomes increasingly automatic. Consistent with these observations, recent brain imaging studies have shown a shift from widespread cortical involvement during early learning to predominantly subcortical activation during later learning phases. However, even expert behavior is still accompanied by conscious experience. Therefore, any functional role for consciousness in skill acquisition is likely to correspond more to HOC than to primary consciousness.

Victor Lamme has argued more generally in favor of a learning function for consciousness. According to Lamme, consciousness arises from the interaction of a feedforward sweep of stimulus-evoked neural activity and a recurrent or reentrant sweep originating in frontoparietal areas, an idea that fits neatly within the integration consensus. Lamme argues that the recurrent sweep promotes synaptic plasticity (and hence learning)

by allowing pre- and postsynaptic neurons to be active simultaneously. Challenge for this view includes ample evidence for unconscious learning, both in human subjects and in machines, as well as theoretical difficulties in establishing direct connections between synaptic plasticity and behavioral learning.

The Boundaries of Integration

An important issue within the integration consensus is, which kinds of signals are capable of being integrated into conscious scenes and which are not? For example, neural activity related to vegetative functions and to low-level perceptual processes does not evoke conscious contents. A useful approach to this question, advocated by Baars, is to look for common features when contrasting conscious and unconscious processes.

According to Morsella s supramodular interaction theory conscious contents mediate interactions among supramodular response systems. These are systems which have different high-level concerns (e.g., the food intake system, the instrumental learning system) and which can come into conflict at the level of the skeletomotor system. Morsella points out that consciously impenetrable processes such as pupillary reflexes, peristalsis, and bronchial dilation do not involve control of the skeletal muscle. By contrast, consciously penetrable processes like inhaling, coughing, swallowing, and defecating all do. On this view, the function of consciousness is to mediate interactive processing across subsystems to allow the organism at any given time to produce a single, adaptive, skeleomotor action.

A different contrastive approach is taken by Bjorn Merker, whose starting point is the stability of the consciously perceived world. Merker emphasizes the remarkable nature of this stability, given the confounding influence of the self-produced motion of sensor arrays mounted on multijointed and swivelling body parts. Conscious contents successfully exclude both the multiple sensory and sensorimotor transformations, which are needed to extract a stable world image, and the complex coordinations of muscle movements needed to produce actions. What is left is a stable arena for decision making and for planning our actions, and the function of consciousness in this view is precisely to provide and maintain this stable arena. An interesting implication of this proposal is that consciousness will be present in all organisms that face similar problems of coordination and neuronal logistics.

Discrimination and Complexity

In a series of influential articles Edelman and Tononi observe that conscious scenes are not only integrated, they are at the same time also differentiated from each other. Not only is every conscious scene experienced all of a piece (integration), but every conscious scene is also unique (differentiation). Thus, the occurrence of any conscious scene simultaneously rules out the occurrence of a vast number of alternative conscious scenes. In the strict sense that information corresponds to a reduction in uncertainty, every conscious scene, primary or higher-order, is therefore enormously informative. This is functional for the organism because each differentiated conscious scene can be linked to a different behavioral response. On this view, the function of consciousness is adaptive and flexible discrimination.

The foregoing is at the heart of Edelman and Tononi s dynamic core hypothesis (DCH), which is part of the more general theoretical framework provided by Edelman s theory of neuronal group selection (TNGS). According to the DCH, the balance between differentiation and integration in every conscious scene is underpinned by a corresponding balance in the neural dynamics responsible for consciousness, which in turn implies that consciousness is generated by interactions among widely distributed groups of neurons. A key feature of the DCH is the proposal of a quantitative measure for this balance: neural complexity, which uses information theory to express the extent to which large subsets of a system tend to behave coherently and small subsets tend to behave independently. (Seth has recently proposed an alternative quantitative measure, based on multivariate autoregression modeling, called causal density .) Computational modeling has shown that the highly reentrant neuroanatomy of the thalamocortical system is particularly well suited to producing dynamics of high neural complexity, whereas other neural systems such as the cerebellum and basal ganglia are not.

Accordingly, the DCH proposes that the neural mechanisms underlying conscious experience consist of a functional cluster in the thalamocortical system, this being the dynamic core.

A recent variant of the DCH, Tononi's information integration theory (IITC), proposes a different quantitative measure, Φ, which is based on identifying the informational weakest link within a system. Whereas the DCH proposes that high values of neural complexity may be necessary, but not sufficient for consciousness, the IITC proposes that Φ is by itself an adequate measure of the quantity of consciousness generated by a system. Therefore, according to the IITC, the function of consciousness is to integrate information for the simple reason that conscious experience is defined as information integration.

The DCH and the IITC sharpen and quantify the integration consensus and relate it directly to neurophysiological processes. They attempt to escape both CI and EP by explicitly identifying consciousness with discrimination (DCH) or information integration (IITC). However, this move can be criticized as defining away the problem and in addition it may be possible to find examples of unconscious systems with arbitrarily high neural complexity and/or Φ. Other neurophysiological perspectives that participate in the integration consensus include the so-called field theories of consciousness, but these theories make similar claims regarding conscious function and will not be discussed further here.

Neuronal Group Selection

While the DCH derives from the TNGS, the latter offers additional perspectives on the function of consciousness that extend beyond information integration and discrimination among conscious scenes.

Edelman's TNGS (also known as neural Darwinism) is a large-scale biological theory of brain function that has roots in evolutionary biology and immunology in viewing brain operations as selectionist in nature, rather than instructionist, like a computer. According to the TNGS, primary consciousness reflects an adaptive linkage of current perceptual categorization to past learning responses and to future needs. This linkage is value-dependent, where value reflects salience to the organism as mediated by neuromodulatory systems (e.g., the dopaminergic, cholingergic, and noradrenergic systems). The result of these interactions is the generation of a remembered present, a description that evokes James' specious present by emphasizing the historically contextualized nature of ongoing primary conscious experience. According to the TNGS, organisms in possession of a remembered present will enjoy increased discriminatory selectivity, flexibility, and planning capacity when responding to complex environments, as compared to their preconscious and unconscious competitors and ancestors. The added value of HOC in the TNGS is that the dependence of consciousness on present inputs is no longer limiting. The ability to explicitly construct past and future scenes extends the integrative capacity of consciousness, allowing the development and deployment of more sophisticated, flexible, and adaptive actions and action plans.

Summary

According to the integration consensus, consciousness functions to bring together diverse signals in the service of enhanced behavioral flexibility and discriminatory capacity. Theoretical proposals within this consensus are among the most highly developed and are increasingly open to experimental testing. However, integration theories must explain why consciousness is necessary, since many integrative functions seem plausibly executable by unconscious devices. The DCH and the IITC address this issue by relating phenomenology and complexity, but for these theories it remains unclear whether high values of neural complexity (or Φ, or causal density) are sufficient for consciousness.

Beyond Integration: Alternative Functions

This final section describes several alternative ideas, which both compete with and complement integrative functions. Because these proposals tend to associate consciousness with one or more existing cognitive functions, they are, as usual, vulnerable to both EP and CI.

Error Correction

Several researchers have argued that the most general of all brain functions is prediction or reduction of prediction error. Predictions help an animal anticipate appetitive and aversive events and facilitate the formulation and execution of the appropriate motor responses. Much of this prediction happens unconsciously. For example, when driving a familiar route we often arrive without remembering much about the journey itself. However, if during the journey another car unexpectedly swerves in front of us our conscious contents suddenly become dominated by the experience of driving. This example illustrates the idea that consciousness functions to detect and allow correction of prediction errors during behavior. According to John Gray s comparator hypothesis, consciousness is part of the brain s monitoring of whether its expectations have occurred or have failed to occur. Because error correction involves metacognitive monitoring of ongoing mental content, any functional role for consciousness in detecting and correcting prediction errors would apply more to HOC than to primary consciousness. After all, when driving a familiar route we are not wholly unconscious, rather, we may be unaware that we are having the experience of driving.

Social Interactions

Day-to-day human existence involves negotiating a maze of social interactions, which in turn leads to conceiving of other people as beings with minds. According to Nicholas Humphrey, humans must be excellent natural psychologists, quickly and effortlessly attributing to other people mental content such as beliefs, desires, moods, sensations, and the like. Humphrey s proposal is that consciousness, in particular HOC, fulfills this functional role: By being conscious of our own mental content we acquire an enhanced ability to infer the mental content of others, especially those that belong to our own social group. This view challenges the attribution of consciousness to creatures that lack highly developed mechanisms of metacognitive access and/or intricate social lives; such creatures include many nonhuman animals and possibly also infants and autistic people.

Dreaming

Most discussions of the function of consciousness focus on conscious contents during wakefulness, but much of our lifetime conscious experience occurs during sleep (i.e., when dreaming), and dream content is substantially different from waking content. For example, dreams have the property of naïve realism and certain qualia, such as odors, are rarely present.

Dreams pose an apparent challenge for ascriptions of function because the dreaming brain is actively inhibited from generating behavior. But the fact that behavior is restricted or absent during dreaming does not mean that dream content is causally impotent. It is easy to imagine that dream content can have causal power by affecting behavior in subsequent waking states. Revonsuo has proposed that consciousness in general has the function of providing a virtual reality arena in which the consequences of actions can be tried out without taking the action itself, and that this function is nowhere more evident than in dreaming. Dream content is often disturbing, threatening, and is associated with anxiety much more than waking consciousness, even in subjects without anxiety disorders or depression. According to Revonsuo, dreaming was originally a biological defense mechanism for simulating threat perception and rehearsing threat avoidance responses.

Conclusions

While there may always remain suspicious epiphenomenalists and die-hard conscious inessentialists, there is abundant and increasing evidence that consciousness is functional. This evidence pertains both to the functional utility of being a conscious organism, and to having particular conscious content. According to the integration consensus, being a conscious organism allows for the adaptive integration of many input and output signals in the service of behavioral flexibility, and the particular conscious content that is integrated functions to elicit a particular adaptive response. But because consciousness is a constellation concept covering a range of possible distinguishable processes, future experiments and theoretical developments will

doubtless refine and differentiate the range of conscious functions beyond those discussed here.

See also: Cognitive Theories of Consciousness; Free Will; Intentionality and Consciousness; Sleep: Dreaming Data and Theories.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. New York, NY: Cambridge University Press.
Block N (1995) On a confusion about a function of consciousness. Behavioral and Brain Sciences 18: 227–247.
Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory. New York, NY: Oxford University Press.
Damasio A (1994) Descartes' Error. London: MacMillan.
Dennett D (1991) Consciousness Explained. Boston, MA: Little, Brown, and London.
Dretske F (1997) What good is consciousness? Canadian Journal of Philosophy 27: 1–17.
Edelman GM and Tononi G (2000) A Universe of Consciousness: How Matter Becomes Imagination. New York, NY: Basic Books.
Flanagan O (1992) Consciousness Reconsidered. Cambridge, MA: MIT Press.
Gould SJ and Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B: Biological Sciences 205: 581–598.
Haggard P (2008) Human volition: towards a neuroscience of will. Nature Reviews Neuroscience 9(12): 934–946.
Humphrey N (1982) Consciousness Regained. Oxford: Oxford University Press.
Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. Behavioral and Brain Sciences 8: 529–566.
Rosenthal D (2008) Consciousness and its function. Neuropsychologia 46: 829–840.
Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5: 42.
Wegner D (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

Biographical Sketch

Anil Seth was born in Oxford, England. After receiving his first-class degree in natural sciences from the University of Cambridge, he gained his MSc in computer science from the University of Sussex. He also received his DPhil in artificial evolution and ecological modeling from Sussex. From 2001 to 2006, he was a postdoctoral fellow and then an associate fellow in theoretical neurobiology at The Neuroscience Institute, San Diego, where he worked closely with the Nobel laureate Gerald Edelman. He returned to Sussex in 2006, where he is currently a senior lecturer in informatics. Seth has published more than 50 peer-reviewed articles, he is an associate editor of the journal *Adaptive Behavior*, and he has edited special issues of *Neuroinformatics* and *Philosophical Transactions of the Royal Society*. His research interests cover various areas of theoretical and experimental neuroscience, with a focus on the science of consciousness. In 2008, he was awarded an EPSRC leadership fellowship to develop new methods in computational neuroscience, in order to advance the scientific study of consciousness. For further information, visit www.anilseth.com.

General Anesthesia

M T Alkire, University of California, Irvine, CA, USA

Glossary

Amnesia – A loss of memory function in which old memories cannot be remembered or new memories cannot be formed.
A defining characteristic of anesthetic molecules is that when they are given in a sufficient amount they are all capable of inducing a temporary and reversible state of amnesia, where new memories cannot be formed.
Arousal – The brain and body's state of alertness. It varies and fluctuates on a continuum from low to high. Very low levels of arousal are seen with sleep, coma, and anesthesia. Very high levels of arousal are found with anxiety disorders, mania, and seizures. Optimal arousal is found in a middle zone between too little and too much arousal. Optimal arousal is associated with the best performance on behavioral and mental tasks. This relationship can be graphed as an inverted-U and is known as the arousal–performance curve.
Gaseous anesthesia – Anesthetics that are delivered by breathing them into the lungs. These molecules are gases at room temperature.
Immobility – When a large enough dose of anesthesia is given, a person will not move during an operation even when they are being cut open. This immobility feature is considered one of the defining characteristics of anesthetic molecules. This dose-dependent effect of anesthesia occurs at different concentrations for various anesthetics. A basis for comparing the potency of anesthetic substances has been related to their ability to cause immobility. The minimum amount of anesthesia that needs to be dissolved in the lungs (i.e., the minimum alveolar concentration (MAC) of an agent) that will produce immobility in 50% of a population is known as the agent's MAC value. Thus, 1 MAC of isoflurane is 1.2%, whereas 1 MAC of desflurane is 6%; however, both agents, at these doses, have an equivalent effect on behavior.
Inhalational anesthesia – These are also anesthetics that are delivered by breathing them into the lungs. However, these molecules are liquids at room temperature. The vapor of the liquid is mixed with a carrier gas (such as oxygen) and the mixture is then inhaled into the lungs.
Intravenous anesthesia – Anesthetics that are delivered by injecting them into the blood stream through a vein.
Unconsciousness – During anesthesia a person is considered unconscious when they fail to move or respond to a rousing shake or a loud voice.
Wakefulness – The state of being awake. It implies a brain physiological state that differs from that found with sleep or anesthesia, but it does not by itself signify the presence of consciousness, as demonstrated by those in a persistent vegetative state.

Introduction – Why Study Consciousness with Anesthesia?

The neurobiology of consciousness is now a legitimate topic for scientific inquiry. Not long ago (and still to a large extent) scientists curious about the nature of human consciousness could only study its neurobiology through the use of a good cover story. Anesthesia is a unique tool for the modern scientific study of consciousness because it offers direct experimental control over the variable of most interest (i.e., the presence of consciousness

itself). It allows brain functioning to be examined in the same person when they are either in a conscious or an unconscious state. Anesthetic manipulation of consciousness can be used with any number of different modern-day research technologies ranging from studies of cellular electrophysiology to whole-brain neuroimaging techniques.

Anesthetic Molecules

There are a number of molecules that can induce a temporary state of unconsciousness when inhaled or injected into the body. All early anesthetics, developed around the mid-1800s, were inhaled agents. At that time, inhalation and swallowing were considered the only effective means for introducing a medication into the body. The idea that medications could be effectively delivered directly through the skin and into the bloodstream had to await the development of the hypodermic syringe in 1853. Both the French physician Charles Pravaz and the English physician Alexander Wood are credited with independently pioneering the hypodermic syringe. Anesthetics that are inhaled are either gases or they are the vapor of liquids at room temperature. Nitrous oxide, xenon, ethylene, acetylene, and cyclopropane are examples of anesthetic gases. Ether (i.e., diethyl ether), chloroform, halothane, enflurane, isoflurane, sevoflurane, and desflurane are a few examples of volatile anesthetics. Anesthetics that are injected are commonly referred to as intravenous agents. They are usually given intravenously (i.e., into a vein), though some would also be effective if given intramuscularly (i.e., into a muscle) or even orally. Sodium thiopental or 'Pentothal,' propofol, etomidate, and ketamine are examples of intravenous anesthetics. The use of intravenous anesthetics only became popular following the development of the thiobarbituric acid derivative 'Pentothal.' This was developed by Abbott Laboratories researchers Ernest H. Volwiler and Donalee L. Tabern, in the 1930s. They were looking for a drug that could induce anesthesia when injected directly into the bloodstream. Ralph M. Waters, MD was the first anesthesiologist to use thiopental on a patient in 1934.

Only a handful of inhaled agents comprise the selection of drugs currently available for modern anesthesia care. Most of the historical agents have long been abandoned for routine clinical use because many of them were explosive and often were associated with toxic or even potentially fatal side effects. The modern selection of inhaled agents now available for routine clinical use consists only of nitrous oxide, isoflurane, sevoflurane, and desflurane.

How Anesthesia Works – Cellular and Molecular Mechanisms

How anesthesia works has been a mystery since the discovery of anesthesia, itself. There are really two issues imbedded within the concept of determining how anesthesia works: one is to understand the molecular sites of anesthetic action and the other is to understand the behavioral effects caused by anesthesia. The first issue of interest is to understand the mechanism by which an anesthetic molecule interacts with its molecular targets to cause a functional change in the nervous system that ultimately leads to inducing the anesthetic state. Many important questions are contained within this query. What are the potential molecular sites of anesthetic action? Which molecular effects are relevant to the ultimate behavioral effects of anesthesia? By what process does an interaction between an anesthetic molecule and its molecular site of action lead to the behavioral effects of anesthesia? Most research into 'mechanisms of anesthesia' has focused on finding answers to these important questions. Indeed, much progress has been made in providing answers to these questions. As a number of excellent reviews on the subject already exist (see 'Suggested readings'), this article will only briefly touch upon some of the relevant findings.

The second issue regarding the mechanisms of anesthesia is behavioral. It is more abstract and yet perhaps more relevant to issues of consciousness. It can be illustrated with a simple thought experiment. Assume for a moment that we know with 100% certainty that the molecular site of anesthetic action is say, for example, the

gamma-aminobutyric acid type A ($GABA_A$) protein channel (i.e., the $GABA_A$ receptor). GABA is the primary inhibitory neurotransmitter in the mammalian brain. Activation of this channel by GABA opens a hole through the membrane of neurons that selectively allows negatively charged chloride ions to flow into the cell. This takes the electrical potential of the neuron farther away from that which allows it to fire an action potential. When the membrane potential of a neuron moves toward the action potential firing threshold value of a cell, it is called depolarization. When the potential moves away from the threshold it is called hyperpolarization. Thus, GABA and anything that might enhance the actions of GABA, or possibly open the $GABA_A$ channel directly, will hyperpolarize neurons that contain $GABA_A$ channels and serve as an inhibitory influence on those neurons to keep them from firing action potentials. Consequently, if anesthesia activates $GABA_A$ channels, then anesthesia will enhance inhibition in the brain and essentially shut down the neural activity throughout the brain upon which consciousness depends.

However, even if GABA is the molecular mechanism by which anesthetics act, the bigger behavioral questions of 'why?' still remain. Why does consciousness go away when activity in the brain is suppressed with anesthesia? Why doesn't anesthesia simply turn off pain processing and leave consciousness alone? What fails to function in the brain at the moment that consciousness is lost? If we could watch consciousness turning off in the brain during the induction of anesthesia, would the whole brain turn off as a unit, or would some set of 'consciousness neurons' be identified as the last regions in the brain to turn off when consciousness fades? Does consciousness go away because anesthesia suppresses neural activity, or does it interfere with the ability of neurons to talk with each other?

Anesthetics can be used to study consciousness at doses lower than those needed to cause unconsciousness. By using these lower doses, we can begin to tease apart which brain regions and which neural circuitry starts to fail first as anesthesia begins to suppress consciousness. To clarify things further, we can get at the molecular mechanisms that may be responsible for the effects that anesthetics have on various behavioral endpoints (see the section titled 'The big three dose-related effects of anesthesia: amnesia, unconsciousness, and immobility'). To do this, we can take advantage of another intriguing property of anesthetic agents. Even though all anesthetics seem to ultimately produce the same behavioral effects (i.e., what makes them anesthetic molecules), they seem to do it through a number of different molecular mechanisms. By examining a number of different agents that have a number of different proposed mechanisms of action, the common effects that contribute to a loss of consciousness (LOC) with anesthesia can thus be focused upon and identified. Essentially, only that process turned off by all anesthetic agents, regardless of their particular molecular mechanisms of action, might be the most relevant for understanding consciousness.

Many believe that the most relevant targets for anesthetic actions are ligand-gated ion channels. These include numerous types of receptors including GABAergic, Glycinergic, nicotinic, histaminergic, and serotonergic. Of those, most anesthetics (particularly the intravenously delivered ones) along with most of the inhaled agents seem to exert their effects through actions on the $GABA_A$ channel. GABAergic agents (GABA agonists) are thought to include the intravenously delivered barbiturates, etomidate, and propofol, as well as the inhalational agents: halothane, enflurane, isoflurane, sevoflurane, and desflurane. However, a few anesthetics, namely, nitrous oxide, xenon, cyclopropane, and ketamine, do not seem to have much effect on $GABA_A$ channels. It has been proposed that these agents might exert their anesthetic effects through actions on the N-methyl-D-aspartate (NMDA) channels. Interestingly, the NMDA channel was long thought to be fairly insensitive to the inhalational agents. However, recent work suggests that the inhalationals may have important effects upon this channel by acting as a competitive blocker of the glycine-binding site.

A host of other potential targets are still considered potentially relevant. Recently, anesthetic actions on two-pore domain background potassium leak channels have been identified and proposed as an important mechanism for anesthetic action. These channels are affected by both the '$GABA_A$' and the 'NMDA' selective groups of

agents. Some of the other potential targets may be very selective for a specific class of anesthetic agent, such as the alpha-2-adrenergic agonist dexmedetomidine. Other targets may be less selective for a particular agent, but may be affected by a broader spectrum of various agents, like the background potassium channels just mentioned.

Still other effects of anesthetics may be important. Perhaps voltage-gated ion channels fit in this category. Recent work suggests that voltage-gated potassium channels play a more important role in mediating anesthetic effects on consciousness than previously thought. Anesthetics also have effects on cell-to-cell electrical transmission that is mediated through protein channels in gap junctions. These channels help synchronize local electrical activity among groups of neurons. Anesthetic effects on these channels will tend to disrupt coordinated network activity.

The molecular mechanisms of anesthetic action are sufficiently complex that their full discussion could take up most of this book. We shall return to these issues as needed; yet now we turn our attention toward understanding what the behavioral effects of anesthesia are. Knowing these will greatly improve our framework for understanding consciousness.

The Big Three Dose-Related Effects of Anesthesia: Amnesia, Unconsciousness, and Immobility

The first public demonstration of successful ether anesthesia occurred at the Massachusetts General Hospital in Boston on 16 October 1846. Dentist and anesthetist William Thomas Green Morton anesthetized a patient for the surgical removal of a congenital vascular malformation from the patient's neck. Newspaper articles of the time proclaimed, 'We have conquered pain.' This referred to the fact that up until the discovery of anesthesia a surgical operation was one of the most brutal and painful incidents a person could ever experience. Imagine having your leg amputated without anesthesia. In those days, you would be forcibly held down as you screamed in agony and you would likely pray that your surgeon was both speedy and accurate with his hacksaw.

The surgeon Henry Jacob Bigelow was present at the first public demonstration of ether anesthesia and he was captivated by the potential of the discovery. He subsequently observed Dr. Morton in his dental practice as he used ether on his patients. He wrote about his observations and published them shortly thereafter. His publication, 'Bigelow HJ. Insensibility during surgical operations produced by inhalation. Boston Med Surg J 35:309–317, 1846' represents one of the first published accounts regarding the use of ether anesthesia from a medical perspective. His keen observations and eloquent descriptions of what he saw captures the most essential elements of how anesthesia affects the brain and behavior. These observations led us toward understanding some of the important behavioral endpoints of anesthesia and the issues related to anesthetic-induced unconsciousness. Many quotes are repeated here to allow the reader an unbiased account of this fascinating process of medical discovery.

First, Dr. Bigelow describes the apparatus used to give the anesthetic.

> It remains briefly to describe the process of inhalation by the new method, and to state some of its effects. A small two-necked glass globe contains the prepared vapor, together with sponges to enlarge the evaporating surface. One aperture admits the air to the interior of the globe, whence, charged with vapor, it is drawn through the second into the lungs. The inspired air thus passes through the bottle, but the expiration is diverted by a valve in the mouth piece, and escaping into the apartment is thus prevented from vitiating the medicated vapor.

The device is shown in Figure 1. Modern anesthesia machines today still use the same concept of flow control in the breathing circuit. The anesthesia is mixed with a carrier gas and delivered to the patient through one tube, while exhalation is diverted with control valves to another tube. Thus, with each inspiration the patient gets the same proportion of anesthetic gas as intended and as controlled by the anesthetist. This is how the level or dose of anesthesia is controlled and this is what allows patients, or experimental subjects, to be held in any state of suppressed consciousness, ranging from slightly intoxicated to completely anesthetized, with no spontaneous electrical activity occurring within the brain.

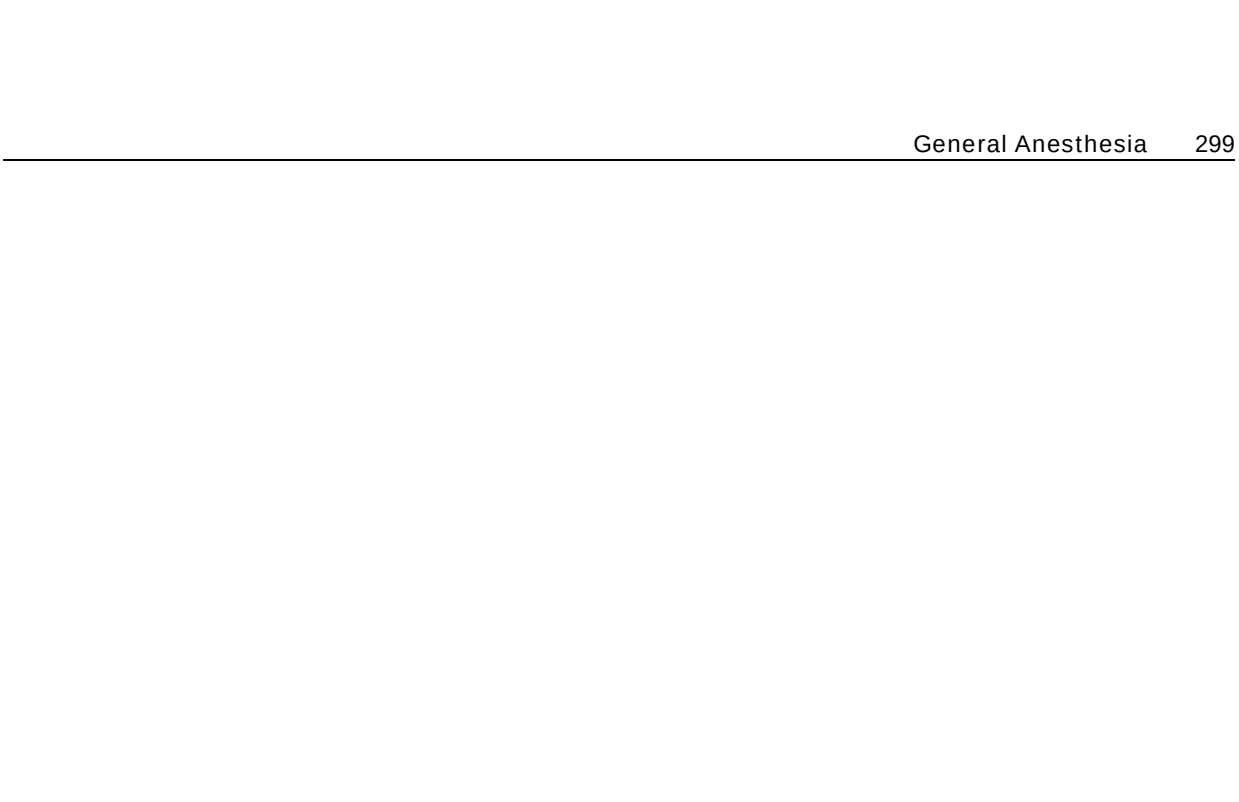


Figure 1 On the left is the original 'anesthesia machine' (c. 1846). Reproduced from Desbarax (2002) Morton's design of the early ether vaporizers. Anaesthesia 57: 463–469, with permission from Blackwell Publishing. The device is a sponge soaked with ether inside of a glass ball that has an inlet and an outlet port. On the right is the modern day counterpart (c. 2008). Despite the apparent differences, both machines work through essentially similar mechanisms, whereby anesthetic vapor is mixed with a carrier gas and drawn into the lungs through controlled flow of the carrier gas.

Dr. Bigelow continues and describes his first patient observation.

> A few of the operations in dentistry, in which the preparation has as yet been chiefly applied, have come under my observation. The remarks of the patients will convey an idea of their sensations. A boy of sixteen, of medium stature and strength, was seated in the chair. The first few inhalations occasioned a quick cough, which afterwards subsided; at the end of eight minutes the head fell back, and the arms dropped but owing to some resistance in opening the mouth, the tooth could not be reached before he awoke. He again inhaled for two minutes, and slept three minutes, during which time the tooth, an inferior molar, was extracted. At the moment of extraction the features assumed an expression of pain, and the hand was raised. Upon coming to himself he said he had had a 'first rate dream – very quiet,' he said, 'and had dreamed of Napoleon – had not the slightest consciousness of pain – the time had seemed long;' and he left the chair, feeling no uneasiness of any kind, and evidently in a high state of admiration.

This observation captures a number of important points about the effects of anesthesia. First, unconsciousness is not instantaneous. It takes several minutes of breathing ether to get an effect. Nonetheless, the first primary effect of anesthesia is clearly evident; at some point the patient appeared to become unconscious.

Second, though apparently unconscious, the patient seemed to experience pain, as suggested by his facial reaction and his reaching toward the site of injury. This demonstrates an important clinical point of anesthesia practice. Pain is arousing. It wakes people up from the anesthesia. In modern anesthesia practice we anticipate this and adjust the amount of anesthesia a person is getting to keep them from moving when the surgical stimulation is particularly painful.

Third, the completeness of this patient's unconsciousness is questionable, not only because he moved but also because he reports having had a dream. In modern anesthesia practice this might be considered a case of intraoperative awareness, a condition in which the patient is expected to be fully unconscious during the operation, yet still retains some (or perhaps most) mental functions.

Fourth, he demonstrates the second primary effect of anesthesia (i.e., amnesia) and claims no memory for

the experience of pain. This amnesic effect occurs at doses of anesthesia that are much lower than those needed to produce unconsciousness. Given that amnesia always comes before unconsciousness, as the dose of anesthesia is increased, one is always left with the question: How do we know people under anesthesia are really unconscious, if the drug also blocks their memory of the experience?

Fifth, he demonstrates an important psychophysical effect of anesthesia that is often reported by patients upon waking up, yet remains an area that has not been fully investigated in the scientific literature. He reports a change in perception of the passage of time. One of the first questions people often ask upon recovering their senses after anesthesia is "What time is it?" Also, they often ask, "Where am I?" Both questions point to the suggestion that low doses of anesthesia interfere with one's perception of space and time. Studies do support this notion, but the real meaning of these effects remains to be fully explored. Interestingly, upon waking up patients never ask, "Who am I?" This suggests that some fundamental sense-of-self is a required prerequisite in order for a person to generate questions about their own place in space and time.

Dr. Bigelow continues with another observation.

> On Saturday, the 7th November, at the Mass. General Hospital, the right leg of a young girl was amputated above the knee. . .
>
> The last circumstance she was able to recall was the adjustment of the mouth piece of the apparatus, after which she was unconscious until she heard some remark at the time of securing the vessels – one of the last steps of the operation. Of the incision she knew nothing, and was unable to say, upon my asking her, whether or not the limb had been removed. She refused to answer several questions during the operation, and was evidently completely insensible to pain or other external influences.

This observation illustrates the third primary effect of anesthesia in that all anesthetics, when given in a sufficient dose, will cause a patient to remain immobile in response to surgical pain. Many modern day researchers use this immobility endpoint as the clinical definition for surgical anesthesia (i.e., the MAC concept).

Dr. Bigelow summarizes his observations as follows.

> The character of the lethargic state, which follows this inhalation, is peculiar. The patient loses his individuality and awakes after a certain period, either entirely unconscious of what has taken place, or retaining only a faint recollection of it. Severe pain is sometimes remembered as being of a dull character; sometimes the operation is supposed by the patient to be performed upon somebody else. Certain patients, whose teeth have been extracted, remember the application of the extracting instruments; yet none have been conscious of any real pain. As before remarked, the phenomena of the lethargic state are not such as to lead the observer to infer this insensibility. Almost all patients under the dentist's hands scowl or frown; some raise the hand. The patient whose leg was amputated, uttered a cry when the sciatic nerve was divided. Many patients open the mouth, or raise themselves in the chair, upon being directed to do so. Others manifest the activity of certain intellectual faculties. . .
>
> In none of these cases had the patients any knowledge of what had been done during their sleep.

With these statements he captures the fact that anesthesia often leads one to feel as if they were having an out-of-body experience, again suggesting a disconnection between one's sense-of-self and one's awareness of time and space. Yet, to the outside observer the patients often seem to respond dramatically to the pain, in the moment. This raises a question about the completeness of the unconsciousness, but clearly and impressively it also establishes the solidity of the amnesia, as none of the patients remembered too much about the pain of their operations.

Thanks to Bigelow, within a month of its first public demonstration, three of the most important clinical behavioral endpoints caused by anesthesia were described and published in the medical literature: amnesia, unconsciousness, and immobility. What was not entirely clear to Bigelow, though he suspected it, was that many of the effects he witnessed were dose-related. Those patients that did better and seemed to have a more complete effect appeared to get a larger dose. At the other extreme, he clearly describes one episode of a near fatal overdose, which he attributed to an incident of prolonged breathing on the apparatus. It was not until 1847 and the brilliant work of the British Physician and anesthetist John Snow that the dose-related aspects of ether anesthesia would be carefully described.

Snow classified the intoxication of ether into five dose-related degrees of etherization, or stages (following the format of Sir Humphy Davy's work with nitrous oxide in 1800). In stage 1 (a very low dose), subjects felt the effects of the agent, but remained aware of where they were and what was going on and they remained in control of voluntary motor functions. In stage 2 (low dose), subjects remained conscious and could still make voluntary movements, but now they occurred in a 'drunk-like' uncoordinated manner. In stage 3 (moderate dose), subjects were unconscious. They would not make voluntary movements, but they could still move in response to stimulation. In stage 4 (much greater surgical dose), subjects were unconscious and made no responses to intense physical stimulation. Yet they remained able to breathe. In stage 5 (slightly greater lethal dose – not examined in humans), subjects made only feeble attempts at breathing, with the apparent paralysis of respiratory muscles.

It is the dose-dependent nature of anesthetic effects on human cognitive functioning that makes anesthesia such a valuable tool in the study of consciousness. With most drugs, the pharmacological effect might vary greatly from person to person. Large numbers of subjects might need to be studied in order to find out if the drug is doing what it is expected to do. With anesthetics, the drug is given at a dose that produces the desired result. Moreover, the drug effect occurs quickly and dissipates rapidly when the drug is removed. This means that to study consciousness, one can experimentally manipulate it with anesthesia in any given individual, essentially at will.

Many of the responses observed by Bigelow would not be seen with modern anesthesia practice. A number of the responses occurred because the dose of ether was simply too little and the drug was wearing off when the procedure was attempted. Today, anesthesia gas is delivered in a continuous manner and the amount of gas being exhaled is monitored. This allows one to verify that the drug is reaching the brain at the intended dose, because the amount being exhaled is a correlate for how much is coming out of the brain and passing through the body. Also, in modern practice, patients do not experience most of the potentially interesting cognitive effects of anesthesia. Patients are almost always given a rapid intravenous induction of anesthesia, rather than a slow controlled inhalational induction. This is done so that patients pass quickly through the low-dose excitement phase of anesthesia, which is often associated with coughing and possibly uncontrolled motor movements. Additionally, modern anesthetic agents like sevoflurane represent a great improvement over ether. They work faster and are less irritating to the airway. This allows one to do an inhalational induction, if so desired. Indeed, inhalational induction with sevoflurane is tolerated well and is often the method of choice for putting children to sleep. With a few quick breaths of a high-dose of sevoflurane a child will rapidly fall asleep.

'Turning Off' Arousal

From a neurophysiology perspective, what is known about the process of anesthesia in the brain is that most anesthetics cause a generalized dose-dependent slowing of the brain's electrical activity, as measured with an electroencephalograph (EEG). The electrical activity of the awake brain is described on EEG tracings as being of a low-voltage amplitude with many high-frequency (i.e., fast) components. This means that if one were to record the electrical activity coming from the brain over some period of time, the tracing seen would appear to bounce back and forth very quickly and never deflect too far from the baseline. In contrast, the EEG during deep anesthesia becomes one of higher voltage with slowly oscillating activity. This means that the tracing would deviate far from the baseline in long slow rolling waves that would follow one another in a synchronized pattern. Awake EEG patterns are thus often described as 'fast' or 'desynchronized,' whereas anesthetized EEG patterns are often described as 'slow' or 'synchronized.' These features are shown in Figure 2.

Importantly, the transition from an awake (i.e., conscious) desynchronized EEG pattern to an anesthetized (i.e., unconscious) synchronized EEG pattern occurs in a dose-dependent manner that is fairly consistent between various anesthetic drugs. In general, when more anesthesia is given, the EEG will show more slow-wave activity. This

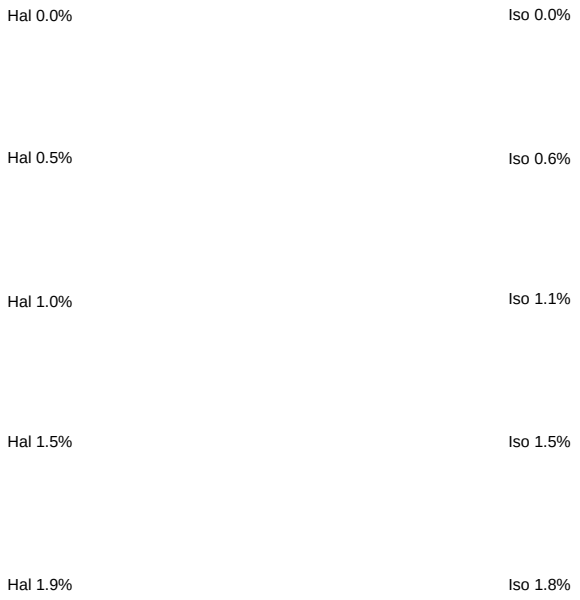


Figure 2 Dose-dependent changes in EEG activity seen with halothane (hal) or isoflurane (iso) anesthesia in the rat. Note how the waveforms progress from low-voltage fast activity to high-voltage slow activity with increasing anesthetic dose. Note also that isoflurane is capable of causing a burst-suppression pattern of activity, in which the EEG is silent except for an occasional brief burst to be followed again by no activity. The pattern of changes is similar to that found in humans during the transition from wakefulness to slow-wave sleep. Reproduced from Hudetz A (2002) Effect of volatile anesthetics on interhemispheric EEG cross-approximate entropy in the rat. Brain Research 954: 123–131, with permission from Elsevier.

change in EEG activity from desynchronized to synchronized during the induction of anesthesia is remarkably similar to the same types of changes in the pattern of the EEG found naturally, during the transition from normal wakefulness (i.e., conscious) to nonrapid-eye-movement (NREM) sleep (i.e., unconscious).

The similarity in how the EEG responds to both sleep and anesthesia is one reason why many have proposed that sleep and anesthesia likely share common neurophysiological mechanisms. Indeed, it was a long held belief that the neurophysiological basis of the anesthetic state was a direct result of anesthetics interfering with the normal brainstem mechanisms that mediate arousal and wakefulness. To understand current work in the field, some fundamental historical observations about the arousal system must be incorporated to provide a proper theoretical framework.

A Brief History of Arousal

In 1949, Giuseppe Morruzzi and Horace W. Magoun discovered that electrical stimulation applied to the brainstem of alpha-chloralose anesthetized animals would change cortical EEG activity from a pattern associated with deep sleep or anesthesia (i.e., high voltage slow activity) to one associated with wakefulness (i.e., low voltage fast activity). This work, coupled with contemporary findings of its day, established the existence of an ascending reticular activating system (ARAS). Neural activity in the ARAS was identified as being intimately involved with regulating the state of cortical arousal (or wakefulness) and the effect was thought to be mediated in part through actions from the ARAS passing through two pathways on its way to the cerebral cortex. One pathway seemed to involve the diffuse thalamic

projecting system and the other pathway seemed to involve the basal forebrain. Cortical arousal (and hence wakefulness) was seen as being an active process that was dependent upon some level of neural activity occurring within the ARAS.

Shortly after its discovery, the ARAS became a theoretical focus for mediating the mechanism of anesthesia on consciousness. It was known that anesthetics suppressed neural activity and if wakefulness was dependent upon activity in the brainstem then perhaps anesthetics worked by simply shutting down the arousal influences of the ARAS. This seemed quite logical and other data strongly supported the idea of needing an active brainstem in order to have wakefulness. Indeed, it was known since the 1930s that the cerveau isolé preparation of Frédéric Bremer, where the brainstem of a cat was transected at the midbrain midcollicular level, was associated with the behavioral and EEG manifestations of sleep. Thus, it was thought that arousal and wakefulness of the cortex depended upon the activity of the ARAS. So if this arousing influence was removed by anesthetic actions on the ARAS, then perhaps anesthetics produced a sleeplike state by causing a functional temporary 'chemical lesion' of the ARAS.

In 1953, John D. French and colleagues, working at the Long Beach Veterans Hospital in California, demonstrated that anesthetics did appear to have a selective depressant effect on the ARAS and they proposed this as the neurophysiologic basis of the anesthetic state. This conceptualization dominated thinking about the mechanism of anesthetic-induced unconsciousness for many years and was readily incorporated into most textbooks of the day. The conceptualization that anesthetics might have localized actions on specific parts of brain neuroanatomy was revolutionary in its day. At that time a unified mechanism of anesthetic action was still being sought.

However, the site-specific idea of anesthetic action on the ARAS was problematic for a number of reasons. For instance, anesthetics are capable of anesthetizing primitive creatures that do not have an ARAS. Anesthetics have effects on cells in culture that are devoid of neuroanatomy influences. Not all anesthetics were found to cause suppression of activity in the ARAS. The suppression effect of anesthesia depended upon exactly where in the ARAS one was trying to record from. In some portions of the ARAS, anesthetics actually increased neuronal activity. Despite these issues, the ARAS suppression theory of anesthetic action dominated thinking for a long time. Even today, it stands as a testament to the controversy that exists between those who believe anesthetic action is a global phenomenon versus those who believe anesthetic action occurs through specific interactions with specific neuronal pathways.

Yet, there were at least two key studies that ultimately served to extinguish enthusiasm for the ARAS theory. First, in 1966, Alemã and colleagues found that injection of barbiturate selectively into the vertebral arteries of humans did not result in a LOC or even in the loss of response to auditory or visual evoked responses. Such a localized infusion of drug into the brain vasculature supplying the area of the brainstem ARAS should have caused a LOC if anesthetics worked by 'turning off' or causing a 'chemical lesion' of the ARAS. The authors concluded, ''In man the most important subcortical structures ultimately responsible for maintenance of the level of consciousness are located rostral to the brainstem, perhaps in the diencephalon.'' The diencephalon includes the thalamus, the hypothalamus, and portions of the basal forebrain.

Second, in 1973, Roy F. Cucchiara and John D. Michenfelder performed a series of experiments on dogs. These studies took advantage of the fact that most anesthetics greatly reduce cerebral metabolism throughout the brain in a dose-dependent manner that closely parallels their effects on the EEG. They reasoned that if the ARAS was essentially an anesthetic-mediated consciousness 'off' switch then (1) the magnitude of the cerebral metabolic suppression caused by anesthesia should be greatly diminished in their dogs if they were made decerebrate. This would occur if the ARAS theory was right, because the animals should essentially already be unconscious. (2) If the ARAS was acutely disconnected from the forebrain during deep anesthesia, the forebrain should awaken because the active off switch would be removed with the decerebration. This removal of the off switch should be reflected as a measurable increase in forebrain metabolism, at the moment the decerebration occurs. They found

that decerebration did not change the amount of metabolic suppression caused by barbiturate anesthesia in the cerebral hemispheres and that acute decerebration did not change forebrain metabolism during steady anesthesia. These findings were taken as strong evidence that anesthesia had a more widespread generalized effect on brain matter and that its actions for suppressing consciousness are not simply limited to suppressing an active consciousness off switch located in the ARAS.

A Current Understanding of Arousal

With time, the details of the ARAS have been clarified and the specific brainstem nuclei involved in generating cortical arousal, as it relates to regulating states of sleep and wakefulness, have now been identified and well reviewed (see 'Suggested readings'). In essence, at least four brain sites are considered important in regulating cortical arousal and controlling states of vigilance; these include the (1) mesencephalic reticular formation (MRF), (2) the basal forebrain, (3) the hypothalamus, and (4) the thalamus. The brainstem arousal nuclei of the MRF are now known to cluster in specific sites and are associated with the efferent release throughout the cortex of particular monoaminergic neurotransmitters.

The primary sites and transmitters thought to be involved in mediating cortical arousal are as follows: cholinergic from the dorsal tegmental area and the mesopontine tegmentum involving the pedunculo-pontine tegmentum (PPT) and the lateral dorsal tegmental (LDT) nuclei, with additional influence arising from the nucleus basalis of Meynert (NBM) in the basal forebrain and the septal nuclei; noradrenergic from the locus coeruleus; serotonergic from the dorsal raphe; and dopaminergic from the substantia nigra and ventral tegmental area. Numerous other amino acid neurotransmitters and neuromodulators also play a role in regulating arousal.

What is now known is that sleep is not just simply the passive withdrawal of arousal influences on the cortex that emanate from a reduction of activity in the ARAS. Indeed, the pathways and neurotransmitter systems that play a role in regulating sleep and causing the induction of sleep, with its various stages, are numerous and complex. A number of these brain sites and their associated neurotransmitters are shown in Figure 3.

The ARAS is now thought to involve both the reticular thalamic nucleus (NRT), an area surrounding the outer part of the thalamus, and the intralaminar nucleus of the thalamus medially. It has been proposed that with the inclusion of these diencephalic extensions of the reticular formation into the thalamic areas, the ARAS system should be renamed as the extended reticulo-thalamic activating system (ERTAS). This shift would help to account for mechanisms related to shifting attention at the thalamic level and how information in widely dispersed cortical areas might be bound together into singular conscious perceptions. It has also been proposed that the nucleus accumbens is a subcortical focal point in the basal forebrain that is involved with linking frontal planning systems and hippocampal memory systems with thalamic attentional mechanisms. It is suggested by Newman and Grace (1999, see suggested readings) that the nucleus accumbens acts as a master gatekeeper to the thalamic gates of attention, in order for one 'to select and ''stream'' conscious episodes across time (hundreds of milliseconds to several seconds).'

Is Consciousness a Local or a Global Phenomenon?

One debate in consciousness research involves trying to understand the extent to which consciousness can be localized within the brain. Are there particular consciousness neurons? Or, can consciousness arise only through the complex interplay of widely separated groups of neurons? This controversy can be seen as separating proponents into two camps: the locationists versus the globalists. As an example of a 'locationist,' neurosurgeon Joseph Bogen proposed conscious awareness is localized within the thalamic intralaminar nuclei (ILN). He came to this conclusion based on studying the anatomy of the ILN. The ILN receives widespread afferent connectivity from sensory cortical and brainstem arousal areas, and sends efferent projections to the striatum. Together, these properties place the ILN at an intermediate processing stage in a sensory–motor loop, where it can contribute to the control of motor output after being influenced by sensory information

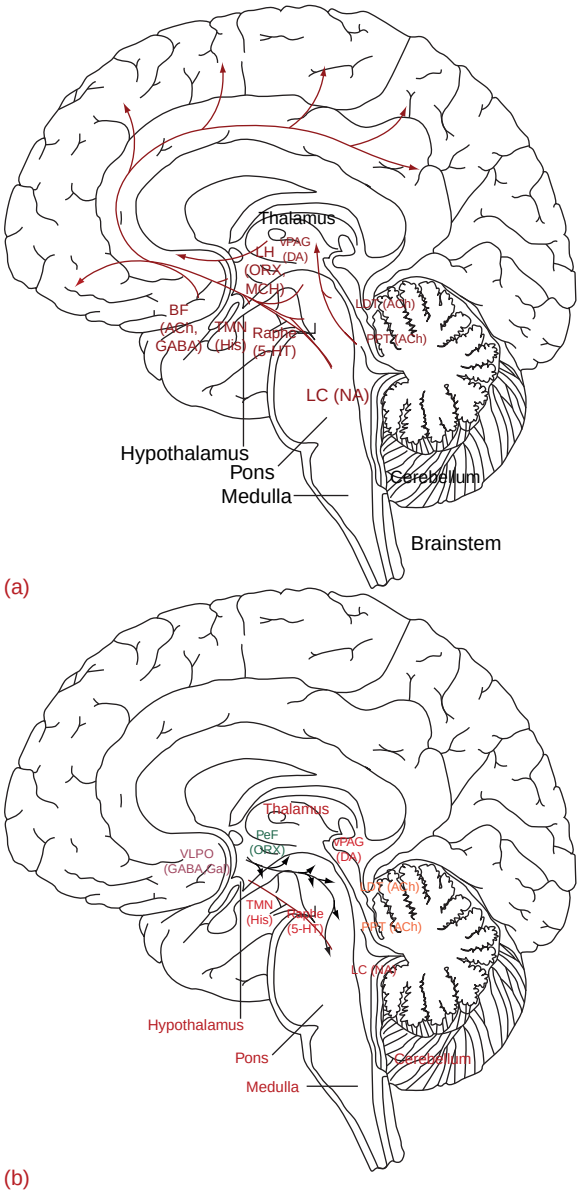


Figure 3 (a) Major brain centers participating in the ascending pathways of cortical arousal and sleep regulation are shown. (b) Descending influences from the VLPO and their interactions with ascending control points are shown. Abbreviations: Structures: LC, locus coeruleus; LH, lateral hypothalamus; PPT, pedunculopontine tegmentum; LDT, laterodorsal tegmentum; vPAG, ventral periaqueductal grey; TMN, tuberomammillary nucleus; PeF, perifornical nucleus; VLPO, ventral lateral preoptic nucleus; and BF, basal forebrain. Neurotransmitters: NA, noradrenaline (norepinephrine); Ach, acetylcholine; 5-HT, serotonin; His, histamine; GABA, gamma amino butyric acid; ORX, orexin; DA, dopamine; and MCH, melanin-concentrating hormone. Reproduced from Saper C, et al. (2005) Nature 437(7063): 1257–1263, with permission from Nature Publishing Group.

processing. Furthermore, he argued that there were only two sites in the brain where small bilateral lesions would cause an abrupt LOC: the MRF and the ILN. In essence then, turning off ILN activity should turn off consciousness. Perhaps anesthetics exert their effects on consciousness not through actions on the ARAS directly, but through localized actions on the ERTAS, or on the ILN specifically.

As an example of a 'globalist,' neurosurgeon Wilder Penfield in the 1950s hypothesized that there is no single site in the brain that mediates consciousness. He felt that it emerges from the complex interactions of multiple brain regions. These regions interact through a centralized subcortical system that integrates the information associated with conscious voluntary movement. His work on epileptic patients led him to this conclusion. He found that patients with large excisions of cortical tissue immediately in front of or behind the precentral motor gyrus remained capable of performing complex intricate motor tasks. This implied that the synaptic activity associated with carrying out such intricate motor behaviors did not arrive at the motor cortex through a corticocortical pathway, rather it must have come from some subcortical structures with equal access to both hemispheres. Penfield termed this empirically deduced system the 'centrencephalic integration system,' and considered it was composed of structures in the higher brainstem, which includes the thalamus. An early conceptual schematic of the centrencephalic integrating system is shown in Figure 4.

For Penfield, the centrencephalic system represented the highest level of neuronal integration associated with consciousness. In arguing the case for a centrencephalic integration system Penfield stated,

> Consciousness exists only in association with the passage of impulses through ever changing circuits of the brainstem and cortex. One cannot say that consciousness is here or there. But certainly without centrencephalic integration it is nonexistent.

Thus, Penfield's hypothesis could be considered a global integration hypothesis of consciousness. Bogen's hypothesis is considered a location-specific hypothesis of consciousness in which consciousness is placed directly in the ILN. To understand which

Figure 4 The centrencephalic integrating system as conceptualized by Penfield in 1954. Reproduced from Penfield and Jasper (1954) Epilepsy and the Functional Anatomy of the Human Brain. Boston, MA: Little, Brown and Company, with permission from Little, Brown & Company.

might be more correct, we can image brain activity in subjects rendered unconscious with anesthesia and see whether any brain areas specifically 'turned off' with the LOC caused by anesthesia.

Insights from Neuroimaging

In line with the expectations of Penfield, Alemã, Cucchiara, and Michenfelder, neuroimaging the effects of anesthetics on human cerebral metabolism reveals that most agents cause a rather dramatic dose-dependent reduction in overall cerebral metabolism. The magnitude of the reduction at the LOC endpoint is in the range of a 30%–60% decrease from baseline. This is consistent with animal studies conducted over the years and confirms that the brain of man (presumably an animal with consciousness) is no different in response to anesthesia than that of other mammals. Anesthesia seems to work through some generalized global mechanism.

Yet, in line with the expectation of Bogen, French, and Magoun, neuroimaging does reveal some interesting regional effects that accompany the global suppression of brain activity. Figure 5 shows data from several human studies where anesthetics were given at a dose that resulted in the LOC. These studies reveal that unconsciousness is almost always associated with a suppression of relative thalamic activity (the centralized cluster within each image) that occurs above and beyond that found with just the global suppression effect. The studies also show a regional effect on posterior parietal brain regions and an effect on frontal brain regions. The localized thalamic finding, integrated with the history of regional sites of anesthetic action, led to the development of the 'thalamocortical consciousness switch' hypothesis (see 'Suggested readings).

The thalamocortical switch hypothesis proposed that unconsciousness during anesthesia occurs because the thalamus and cortex become functionally disconnected. This happens with a number of different anesthetic agents (which all might have differing cellular mechanisms of action) because they all ultimately cause neuronal hyperpolarization within the thalamocortical system (see 'Suggested readings'). Electrophysiology studies reveal that levels of arousal or states of vigilance are associated with different patterns of firing in thalamocortical network neurons (i.e., thalamocortical, corticothalamic, corticocortical, and reticulothalamic neurons). Tonic firing patterns are associated with wakefulness. It allows incoming sensory information to be faithfully transmitted through the thalamus on its way to the cortex. Burst firing predominates during unconsciousness. The membrane potential of the network cells determines the pattern of firing. For instance, hyperpolarization causes bursting activity, whereas depolarization produces tonic firing

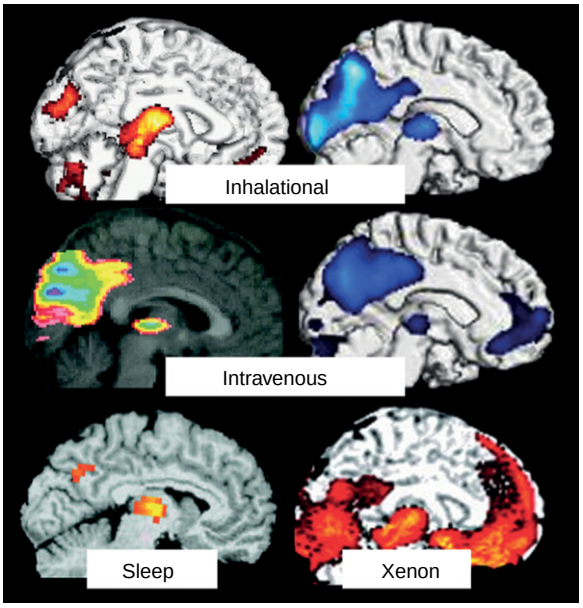


Figure 5 Neuroimaging results investigating the regional relative effects of anesthetics and NREM sleep on consciousness. (a) Regions commonly suppressed during LOC caused either by halothane or isoflurane inhalational anesthesia. The conjunction of the effect between the two anesthetics is shown. Reproduced from Alkire MT, Haier RJ, and Fallon JH (2000) Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. Consciousness and Cognition 9: 370–386, with permission from Elsevier. (b) Regions suppressed during sevoflurane inhalational anesthesia. Reproduced from Kaisti, et al. (2002) Anesthesiology 96: 1358–1370, with permission from Lippincott Williams & Wilkins (Philadelphia). (c) Correlation between increasing doses of propofol anesthesia and reductions in regional cerebral blood flow (CBF) during unconsciousness. Reproduced from Fiset, et al. (1999) Journal of Neuroscience 19: 5506–5513, with permission from Society for Neuroscience. (d) Regions suppressed by propofol during unconsciousness. Reproduced from Kaisti, et al. (2002) Anesthesiology 96: 1358–1370, with permission from Lippincott Williams & Wilkins (Philadelphia). (e) Suppression during NREM sleep. Reproduced from Nofzinger, et al. (2002) Brain 125: 1105–1115, with permission from Oxford University Press. (f) Suppression during xenon anesthesia. Reproduced from Rex, et al. (2006) Anesthesiology 105: 936–943, with permission from Lippincott Williams & Wilkins (Philadelphia). Despite different color scales, all agents show suppressive effects. A link between a regional localized effect on the thalamus and suppression of consciousness seems evident across these different studies and appears to be somewhat independent of the mechanism producing the unconsciousness.

activity. With burst firing, virtually all cortical neurons show a slow oscillation (<1 Hz) of their membrane potential and this reduces the amount of information brain neurons can process. These large coherent fluctuations in membrane potentials are reflected in the cortical EEG as slow waves of high-voltage activity. It is this change in the amount of cortical EEG slow-wave activity that is seen with increasing doses of anesthesia (see Figure 2). Anesthetics hyperpolarize neurons in the central nervous system (CNS) in proportion to their potency as anesthetic agents.

Which Regional Effect Is Important?

The specific regional interactions between various arousal substances or anesthetic-like substances and changes in levels of consciousness can be conceptualized as identifying brain areas that act as consciousness 'on' or consciousness 'off' switch sites. Most of these sites involve interactions with known components of the arousal system that mediate sleep and arousal (as seen in Figure 3). Figure 3 shows a number of relevant sites, but a full clarification of how all of these sites interact (and any additional sites) remains to be forthcoming.

'OFF' switch sites: Sites where localized microinjections enhance the sedative effects of anesthesia.

Sleep research has long implicated the hypothalamus as a site intimately involved in mediating mechanisms of arousal and sleep induction. The hypothalamic nuclei involved in sleep regulation may have relevance for anesthetic-induced unconsciousness as the vast majority of them utilize anesthetic sensitive receptors for signaling. The sedative nature of GABAergic anesthetic agents is thought to overlap with endogenous sleep mechanisms by causing GABAergic inhibition, in part, within the hypothalamic tuberomammillary nucleus (TMN). This nucleus sends arousing histaminergic projections to the cerebral cortex and to the ventral lateral preoptic area (VLPO). The VLPO is thought to act as a sleep onset switch, and during wakefulness it is under tonic inhibitory control from the brainstem via the locus coeruleus (LC). When activated the VLPO sends GABAergic inhibition to numerous sites involved with modulating cortical arousal including the TMN, LC,

dorsal raphe, and perifornical area, which is an area that is involved with orexinergic signaling. Enhanced GABAergic inhibition in these areas (as caused by many anesthetics) will lead to decreased levels of cortical arousal.

The hypothalamic orexin system has been implicated in mediating arousal by helping to stabilize the functional state of various sleep-related regulatory switches, thus acting essentially as a 'flip-flop' switch between states of sleep and wakefulness. Narcolepsy is associated with a mutation of the orexin receptor. Anesthetic interactions with the orexin system are hypothesized to contribute to the overall suppression of arousal and may contribute to rare events of delayed awakening from general anesthesia.

A role for GABAergic and cholinergic mechanisms in the central medial (CM) thalamus, a part of the so-called intralaminar thalamus, in mediating arousal was determined from work on seizure mechanisms. The CM is a primary recipient of efferent arousal signals from the ARAS and the cortex. It also interacts heavily with arousal influences from the hypothalamus, along with influences from the nucleus reticularus of the thalamus (NRT). The CM and the NRT are part of the rostral extension of the ARAS into the diencephalon and together they make up the so-called ERTAS. The CM is also considered to be part of the diffuse projecting system made up of thalamic matrix cells. Matrix cells show immunocytochemical staining for the calcium-binding protein, calbindin. Various agonists and antagonists microinfused directly into the CM will change behavioral and EEG signs of arousal. GABAergic agonists such as pentobarbital and muscimol will cause a rapid behavioral LOC when microinfused directly into the CM. Anesthetic interactions with the basal forebrain should theoretically effect cortical arousal, as acetylcholine (ACh) levels in the cortex are controlled in large part through the activity of the NBM. Cholinergic arousal influences on the cortex arise from two pathways ascending from the ARAS: a dorsal one innervates the thalamus and a ventral one passes through the subthalamus and posterior hypothalamus before reaching the basal forebrain. Anesthetic-induced suppression of cortical ACh levels has long been suspected of contributing to the anesthetic state. Recent in vivo microdialysis work has confirmed that isoflurane reduces cortical ACh levels in a dose-dependent manner.

Most recently, an area in the midbrain has been identified and named the mesopontine tegmental anesthesia area (MPTA) because microinjections of barbiturates in this area cause a rapid apparent LOC, slow-wave EEG and atonia (loss of muscle tone). The MPTA has multiple sites of interaction with the more rostral components of the ARAS including (1) the intralaminar thalamus, (2) the basal forebrain, (3) the hypothalamus, zona incerta, septal area, and striatopallidal system, and (4) the frontal cortex. Many of these areas have already been shown to interact with the sedative component of anesthesia. Additional recent work shows that these and other portions of the limbic system participate in regulating the arousal suppressing aspects of anesthesia.

'ON' switch sites: Sites where localized microinjections block the sedative effects of anesthesia and awaken an anesthetized animal.

The ARAS is the prototypical 'on' switch site, as discovered by Moruzzi and Magoun in 1949. The original report described that the cortical arousal response to electrical stimulation in the ARAS was difficult to obtain in animals that were more deeply anesthetized with alpha-chloralose and it could not be obtained in any animals anesthetized with barbiturates. This revealed that suppression of arousal by anesthesia is both a function of the dose of anesthesia delivered and the specific anesthetic agent being used. Since then, several studies have demonstrated signs of arousal following manipulations of the orexinergic, cholinergic, and glutamatergic systems involving the basal forebrain and the pedunculopontine tegmentum. Importantly, infusion of a-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid (AMPA) into the basal forebrain reverses the reduction in cortical ACh levels seen during isoflurane anesthesia and is associated with EEG signs of arousal that are suggestive of a possible return to consciousness.

There is also a long history showing that cholinergic agonists have an ability to arouse the sleeping animal and lighten the depth of anesthesia. Recently, it was established that the CM thalamus

plays a role in mediating nicotinic mechanisms of arousal such that the microinfusion of nicotine into the CM was found to rapidly restore consciousness to an anesthetized animal (in the continued presence of the anesthetic). Figure 6 shows an animal returning to consciousness following a microinfusion of nicotine specifically into its CM thalamus while inside a chamber filled with sevoflurane. Independent of this anesthesia work on arousal, which focused on the CM thalamus, recent work in humans has also focused on this region in attempts to help patients recover from the lowered arousal secondary to brain injury. Deep brain stimulation of the CM thalamus in one patient in the minimally conscious state helped the patient regain some functional recovery and enhanced awareness.

Neuroimaging and the site-specific work primarily focuses our attention for elucidating mechanisms of consciousness on understanding the dynamic interactions occurring between the thalamus and cortex, as they relate to consciousness. Perhaps the thalamus, as a target of anesthetic action, changes its functional state first and disconnects from the cortex to cause unconsciousness. Or, perhaps the cortex is directly suppressed by anesthesia first and the thalamic suppression effect is only a correlate for the removal of an excitatory corticothalamic feedback tone. Conversely, perhaps the cortex is really the only target of anesthetic action and the changes seen with neuroimaging are simply epiphenomenal to the LOC caused by anesthesia. So, which is affected first when consciousness goes away, and which is the primary site of anesthetic action on consciousness – thalamus or cortex? Neuroimaging lacks the temporal resolution to answer this question.

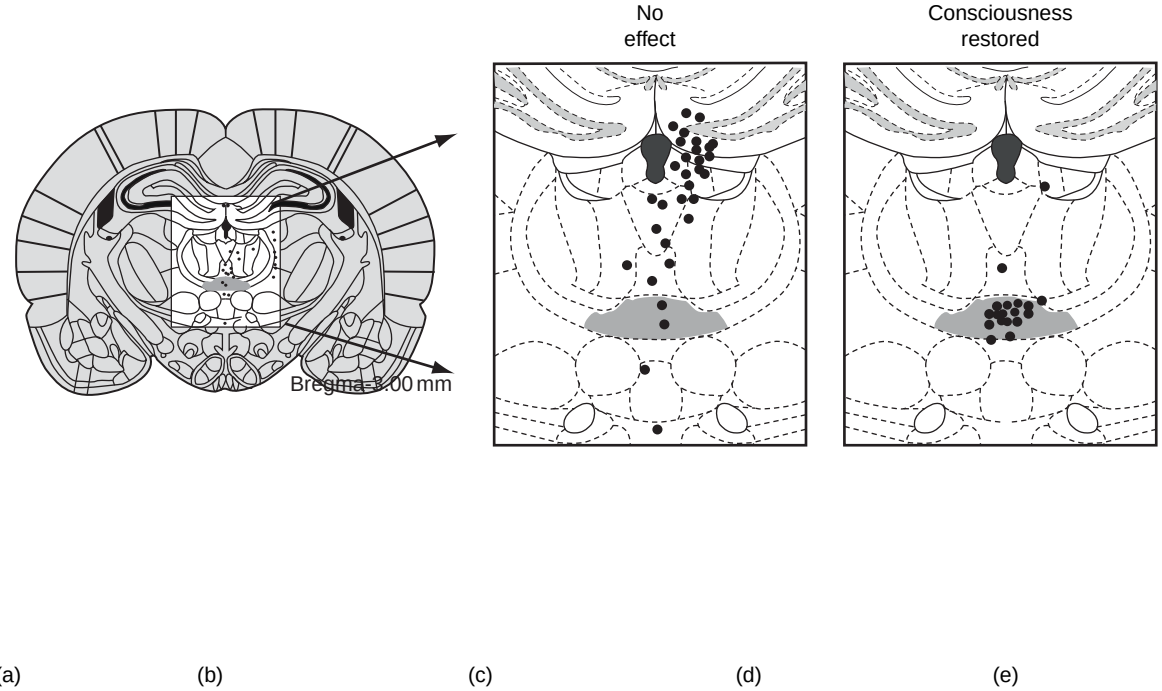


Figure 6 The relative location of the CM thalamus of the intralaminar thalamic nuclei is shown on a coronal rat brain section (gray area in center of white rectangle). Infusion sites of nicotine that either had 'no effect' or 'restored consciousness' to rats under sevoflurane anesthesia are shown as larger inserts. The typical arousal response is shown for one rat. The rat is seen unconscious, lying on its back. Following the site-specific infusion of nicotine into the CM thalamus, the rat awakens and begins to move around. The site-specific nature of this finding suggests the CM thalamus plays an important role in regulating levels of consciousness. Reproduced from Alkire, et al. (2007) Thalamic microinjection of nicotine reverses sevoflurane-induced loss of righting reflex in the rat. Anesthesiology 107: 264–272.

Fortunately, a recent electrophysiology study suggests an answer. Velley and colleagues reported on a series of Parkinson's patients undergoing a two-stage procedure for implantation of a deep-brain stimulator system. After a multichannel electrode was placed into the subthalamic region during one operation, the patients returned for a second operation for the implantation of the electronic stimulator component. During induction of anesthesia for the second operation, EEG activity could thus be obtained simultaneously from both the surface EEG, representing the cortex, and from the subthalamic electrode, using contact points that passed through the thalamus. They found that cortical EEG activity was suppressed at the instant when the patients became unconscious. This cortical change occurred well before similar changes in thalamic EEG-like activity occurred. This result strongly suggests that the cortex is the primary target of anesthetic action and immediately suggests that consciousness is likely to be primarily dependent upon the cortex.

To summarize, site-related interactions between arousal system components and anesthetic-induced suppression of consciousness (or the reversal of the suppression of consciousness) offers overwhelming evidence that the anesthetic state is indeed mediated to some extent by a 'chemical lesion' of arousal as proposed by French and Magoun. Evidence further suggests that the suppression of consciousness likely occurs at the level of cortical interactions, or possibly through some corticothalamic interactions. Much further work will be required to elucidate how consciousness might arise in the cerebral cortex and which parts of the cortex might be most important for consciousness (see 'Suggested readings').

Theoretical Perspectives to Anesthesia and Consciousness

A number of theories have been proposed to explain why anesthesia causes a LOC. A few are briefly summarized here for the interested reader.

Hans Flohr – An information processing theory of anesthesia (1991 and 2006). This theory consists of four basic hypotheses. (1) It is proposed that the occurrence of states of consciousness depends on the formation of higher-order representations in the brain, which represent the internal state of the brain itself. (2) It is proposed that the higher-order representations occur through the spatiotemporal activity patterns occurring in large-scale neuronal assemblies, which can bind together activity from smaller scale assemblies. (3) It is considered that the N-methyl-D-aspartate (NMDA) synapse plays a crucial role in generating conscious states because it underlies the binding mechanism that helps produce the large-scale assemblies. (4) The rate at which representational structures can be built up depends upon the activation state of the NMDA receptor. Unconsciousness occurs if, and only if, NMDA-dependent binding processes are inhibited. All anesthetics are thought to inhibit this process either directly (NMDA antagonists) or indirectly through actions on other receptor systems such as GABA.

Michael Alkire, Richard Haier, and James Fallon – A thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness (2000). As a dose of anesthesia is increased in the brain, at some specific point patients will stop responding the instant they become unconscious. This theory places emphasis on the change in global neuronal firing patterns that switch their state of activity from one of a tonic firing pattern to one of a burst–pause pattern, with the accompanying changes in the state of vigilance. This change in firing patterns occurs through anesthetic-induced neuronal hyperpolarization and is likely mediated primarily by a change in a potassium conductance (in a manner similar to natural sleep), as well as due to changes in chloride conductance caused by GABAergic effects. The hyperpolarization likely occurs through both direct cellular effects of anesthetics and indirect mechanisms mediated by neuroanatomy. Thus, suppressing cortical functioning can result in unconsciousness, and can suppress brainstem mechanisms of arousal. Unconsciousness results when a sufficient proportion of thalamocortical, corticocortical, corticothalamic, and reticulothalamic networks switch their patterns of activity from tonic to burst, and block the brain's ability to integrate information. The switch in firing states also causes a change in the magnitude of functional and effective connectivity that can occur between and among various brain regions.

George Mashour – Anesthetic-induced unconsciousness is cognitive unbinding (2004). The issue of 'binding' refers to the process by which the neural activity that occurs in widely separated and functionally specialized regions of the brain is bound together in a unified singular conscious perception. For instance, when we see a newscaster reporting on the television we see and hear what he says as his lips move. Yet, vision and hearing are processed in different parts of the brain. How do these two information streams come together and form a unified whole? Binding might occur through convergence. This suggests that everything is funneled in toward a set of specialized consciousness neurons. Binding might occur through assembly. This suggests that those neurons which respond to an object tend to fire together and become functionally linked together over time. Binding might occur through synchrony. This suggests that neurons that fire together in the temporal domain may form the basis of neural assemblies. This binding process may be reflected at gamma frequencies in the EEG. As anesthetics slow down the EEG, it is proposed that at some dose of anesthesia, binding that is dependent upon higher frequencies must surely stop.

Giulio Tononi – An information integration theory of consciousness (2004). Although not specifically a theory regarding how anesthesia causes unconsciousness, this theory forms the basis for understanding such effects by providing a principled neuroscientific theory of consciousness. Put simply, the theory claims that consciousness corresponds to the capacity of a system to integrate information. The theory does not require a solution to the binding problem, as no binding problem exists, because the unity of each conscious experience arises when causally effective information is integrated within a system of sufficient complexity.

E. Roy John and Leslie Prichep – The Anesthetic Cascade (2005). This theory represents a neurophysiologic approach to the issue of anesthetic-induced suppression of arousal and consciousness. It incorporates much of the EEG changes caused by anesthesia and relates these changes to presumed neuroanatomic influences. A six-step suppression is proposed to account for the amnesic and unconsciousness-producing effects of anesthesia.

Step 1: It is proposed that the influences of the ARAS on the thalamus and cortex are first reduced.

Step 2: It is suggested that the blockade of memory storage then occurs through the depression of mesolimbic–dorsolateral prefrontal cortex interactions.

Step 3: The nucleus reticularis of the thalamus is then proposed to be released from inhibition through further depression of the ARAS. This is suggested to lead to a closing of thalamic gates through a hyperpolarizing action of GABAA-mediated inhibition of the nucleus reticularis. This then blocks. . .

Step 4: Thalamocorticothalamocortical reverberations that are proposed to mediate perception, which then results in. . .

Step 5: The uncoupling of parietal–frontal transactions which are proposed to block cognition, and. . .

Step 6: The prefrontal cortex becomes depressed causing a proposed reduction in awareness.

Stuart Hammeroff – Entwined mysteries of anesthesia and consciousness (2006). It is proposed that consciousness correlates with gamma synchronized conformational activities of neuronal dendritic proteins in the cortex and other brain regions. It is considered that within each protein, its conformational states are regulated by the endogenous London forces at work in the hydrophobic pockets. The existence of zero-phase lag gamma synchrony suggests that consciousness might involve collective fields that are mediated by the long-range dipole correlations occurring among these endogenous London forces. By forming new exogenous London forces with anesthetic exposure, it is thought that anesthetic gases prevent consciousness by impairing the endogenous London forces in the hydrophobic pockets of the dendritic brain proteins.

Michael Alkire, Anthony Hudetz, and Giulio Tononi – Anesthesia as information disintegration. Consciousness and Anesthesia. Science, Vol. 322 (No. 5903): 876–880 (2008). If consciousness is information integration, then perhaps anesthetic-induced unconsciousness occurs through a lack of information integration. The process of information integration may be stopped by numerous

mechanisms that can cause unconsciousness such as sleep, seizures, and anesthesia.

Conclusions

The search for neural correlates of consciousness can be guided by the use of general anesthetic agents. Anesthesia can interact with 'consciousness neurons' in a temporary and reversible manner. By experimentally inactivating such neurons or preventing them from properly interacting with each other throughout the brain, their functional behaviors can be illuminated and their locations pinpointed with modern neuroimaging techniques.

See also: Coma, Persistent Vegetative States, and Diminished Consciousness; The Neurochemistry of Consciousness.

Suggested Readings

Alkire MT, Haier RJ, and Fallon JH (2000) Toward a unified theory of narcosis: Brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness. Consciousness and Cognition 9: 370–386.

Alkire MT and Miller J (2005) General anesthesia and the neural correlates of consciousness. Progress in Brain Research 150: 229–244.

Arhem P, Klement G, and Nilsson J (2003) Mechanisms of anesthesia: Towards integrating network, cellular, and molecular level modeling. Neuropsychopharmacology 28(supplement 1): S40–S47.

Campagna JA, Miller KW, and Forman SA (2003) Mechanisms of actions of inhaled anesthetics. New England Journal of Medicine 348: 2110–2124.

Flohr H (2006) Unconsciousness. Best Practice & Research Clinical Anaesthesiology 20: 11–22.

Franks NP (2006) Molecular targets underlying general anaesthesia. British Journal of Pharmacology 147(supplement 1): S72–S81.

Hameroff SR (2006) The entwined mysteries of anesthesia and consciousness: Is there a common underlying mechanism? Anesthesiology 105: 400–412.

Hudetz AG (2006) Suppressing consciousness: Mechanisms of general anesthesia. Seminars in Anesthesia, Perioperative Medicine and Pain 25: 196–204.

John ER and Prichep LS (2005) The anesthetic cascade: A theory of how anesthesia suppresses consciousness. Anesthesiology 102: 447–471.

Mashour GA (2004) Consciousness unbound: Toward a paradigm of general anesthesia. Anesthesiology 100: 428–433.

Newman J and Grace AA (1999) Binding across time: The selective gating of frontal and hippocampal systems modulating working memory and attentional states. Consciousness and Cognition 8: 196–212.

Penfield W (1958) Centrencephalic integrating system. Brain 81: 231–240.

Saper CB, Scammell TE, and Lu J (2005) Hypothalamic regulation of sleep and circadian rhythms. Nature 437: 1257–1263.

Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5: 42.

Velly LJ, Rey MF, Bruder NJ, et al. (2007) Differential dynamic of action on cortical and subcortical structures of anesthetic agents during induction of anesthesia. Anesthesiology 107: 202–212.

Biographical Sketch

Dr Michael Alkire graduated summa cum laude from the University of Oregon and received his medical degree from the University of California, Los Angeles. He is currently an associate professor of anesthesiology and a fellow member of the Center for the Neurobiology of Learning and Memory at the University of California, Irvine. He uses anesthesia as a tool for investigating memory, consciousness, and pain processing in both animal and human models. In 1995, he published the first positron emission

tomography (PET) brain imaging study on the cerebral metabolic effects of propofol anesthesia in volunteers. Subsequently, he completed a series of neuroimaging studies on the effects of various anesthetic agents. This work led to the thalamocortical consciousness switch hypothesis of anesthetic-induced unconsciousness (2000). Current work is focusing on the role played by the central medial thalamus in regulating consciousness (2007), and the role of the amygdala in mediating emotional memory and anesthesia (2008). His work has received numerous editorial comments and highlights, and most recently won the 'Best of Meeting' abstract award competition at the International Anesthesia Research Society Conference. Dr. Alkire's pioneering and award winning anesthesia research continues to reveal knowledge crucial for understanding consciousness and the human brain.

Habit, Action, and Consciousness

H Aarts and R Custers, Utrecht University, Utrecht, The Netherlands

Glossary

Action-chaining – Learning of action sequences by associating each response with the next.
Attention – Selectively focusing on one aspect of the task or environment while ignoring other things.
Closed-loop process – Process that uses information about its outcomes as input.
Goal – An internal representation of a desired state, such as a behavior or an outcome.
Intention – An explicitly formulated plan to perform a specific behavior or attain a specific outcome.
Open-loop process – Process that does not use information about its outcomes as input.
Reinforcement – Strengthening of an S–R link by rewards.
Sensemaking – The process of creating situational awareness and understanding in ambiguous situations in order to make decisions and to gain self-insight.
Skill – Overlearned behavioral routine resulting from practice.
S–R habit – A learned habit to react to a particular stimulus with a particular response.

Introduction

People are creatures of habit. A major part of our behavioral repertoire is frequently exhibited in the same physical and social environment and has taken on a stable character. The concept of habit has a long-established history and rich tradition in the study of human conduct. Given this history and tradition it is not easy to define the concept of habits in one specific way. For instance, early sociologists conceived of habits as well-structured patterns of behavior that are supposed to result from the interaction between society and individuals and thus serve to control people and to produce civilization and culture. They used a broad definition of habits to account for the stability of social institutions that is known by terms such as customs, traditions, social norms, or values. The term habit was also used by theorists who published about evolutionary processes who used the concept to denote the elementary behaviors of lower species. The evolution theorists' work was related to the instinct literature of that time revealing an interest in the inherent dispositions of living organisms toward a particular behavior that are not based upon prior experience, such as reproduction, animal fighting and courtship behavior, building a nest and feeding. It was in this sense that evolution theorists spoke of such things as the 'feeding habits of British insects.' Later, the term habit was used for reflex actions, which were conceived of as motor responses activated by nerve cells excited by stimuli external to the organism. The term habit, then, refers to stable behavioral patterns that evolve from biological and social processes.

In psychology, much theoretical and empirical research on habits focuses on the individual level of analysis and examines the mental processes underlying the formation and performance of habits and the involvement and disengagement of consciousness in these processes. In general, two issues can be distinguished in examining the role of consciousness in habit and action: the dysfunctionality of habits that has been demonstrated by the limited role of the conscious will in the counter-regulation of habits; and the functionality of habits tested in studies showing decreased reliance on conscious attention and intention as a result of learning and practice. Accordingly, the study on habit, action, and consciousness aims to understand how people manage to act against one's habits, such as overcoming the well-practiced routine of taking the elevator rather than the stairs to

reach the second floor or resisting the habitually driven temptation to eat junk food, and how such habits are learned and can be executed outside conscious awareness in the first place.

In this article, we first briefly address the role of consciousness in overcoming habits by presenting research in which habits are opposed to the will. Next, we discuss theories and research on the development and performance of habits, and examine evidence showing how habits can be performed without conscious attention and intention. Finally, we focus on the relationship between habits and sensemaking by discussing how people can use their habits to gain insights into their conscious goals and other personal dispositions.

Habit versus the Will

Early theorists proposed that people's behavior is governed by the will. The will, it was argued, plays a causal role in goal setting, striving, and attainment, and should especially be important in overcoming habits (the tendency to respond in a specific way to a specific stimulus as a result of strong associations between the two). Thus, habits were conceived of as reflexive processes that, once activated, follow a ballistic route to completion and, as such, are uncontrollable unless an inner force could take a hold of them. This inner force pertaining to the will has also been labeled in several other ways, such as volition, self-determination, and commitment, and forms the core aspect of modern views on the role of consciousness in self-control and the regulation of behavior. Examining the human ability to counteract habits by the will thus promotes a better understanding of when and how habits may intrude and produce errors and action slips that go against the will.

In the study of human control to overcome habits, so-called 'combined method' experiments are commonly used in which habit and the will operate in opposition. In this method, participants are first taught specific stimulus–response associations to establish a habit (or the habit is assumed to preexist before they engage in the experimental task). After some practice, the same stimulus is presented but a different response is required. Practice leads to direct associations between stimuli and responses, so that presenting the stimulus later on automatically activates the habitual response. If this response is not the correct one, it is up to the person (or the will) to counteract the now dysfunctional habit and to make sure that the intended response is produced. If the learned habit is strong enough, then this extra demand suffers from proactive interference of the habit, that is, the tendency to produce the previously associated but now incorrect response.

Although such studies show that people exhibit control over habits to some degree, occasionally they make mistakes or slow down, and these accuracy and speed effects vary as a function of task (e.g., cognitive load, amount of practice, or automation) and personal variables (e.g., frontal lobe damage, subjective importance of negating the habit). One reason for these failures of the will has to do with the notion that the S–R links established by practice can be very strong. Accordingly, once the response is triggered by the stimulus it may be hard to control it by selecting and producing another counter-habitual response, especially when the response time window is short and attention is located elsewhere. Another reason for the limited control over habits by the will concerns the fact that specific aspects of the habit are deeply encapsulated in the information processing system and operate independently from higher level cognitive processes such as acting on the conscious will. This dissociation between habit and conscious will is reflected in dual process models of cognition and behavior that make a distinction between two separate systems (e.g., implicit or nondeclarative memory vs. explicit or declarative memory) that independently contribute to human performance. Thus, even when we consciously try to put new good intentions into place, those previously learned habits remain stronger in more automatic, unconscious forms of memory.

Habit, Learning, and Performance: From Stimulus-Driven to Goal-Directed Action

In order to gain insight into the role of the conscious will in overcoming habits, researchers have focused on how habits are formed and established

as a function of practice. In this area of research, habits are often conceived of as skills that are extremely useful, as they enable us to perform our actions in a mindless, automatic fashion. That is, the more actions we can delegate to the unconscious, the more room there is to do things that necessarily require conscious processing. Writing an article would be a more difficult affair if typing (and driving, and taking a shower, and even brushing our teeth) required conscious planning and awareness of the processes involved in producing the relevant actions. The scientific study of the formation and establishment of habits thus seems relevant to understand the role of consciousness in human behavior.

The understanding and examination of the contribution of conscious and unconscious processes in habitual behavior depends on how one conceptualizes habits, and particularly the underlying structure and mechanism that give rise to the automation of behavior as a result of practice. In general, there are two different approaches to this issue that can be characterized as representing either low level stimulus–response learning and performance or a higher cognitive level of goal-directed learning and performance.

Habits as Stimulus–Response Links

At the lowest level of analysis, habits can be regarded as mere stimulus–response links. According to behaviorist S–R theories, in essence all learning involves associations between stimuli and response, and such links can be established and reinforced by rewards that follow responses to a stimulus. If a child, for instance, picks up the phone after it rings and enjoys the conversation that follows, the response of picking up the phone becomes more likely to occur upon the ringing sound. If these rewards consistently follow a particular response to a particular stimulus an S–R link develops that can be considered a basic habit.

Rewards that play a crucial role in the development of S–R links may arise from different sources. They may, for instance, be administered by other people to promote the development of basic skills. Using operant conditioning, the frequency of a voluntary performed response can be increased by rewarding it. Through classical or Pavlovian conditioning techniques, more complex relations can be learned between non-rewarding and rewarding stimuli. As a result, a dog – for instance – may be trained to sit at a particular command by rewarding the proper response with cookies or strokes. These conditioning techniques reflect the basic learning mechanisms that are responsible for the formation of S–R habits.

These mechanisms also operate unasked for in everyday life. The rewards that drive it are assumed to mainly result from basic biological and social needs. In the light of these needs, certain objects or behaviors that have been learned to reduce the need may acquire incentive value (i.e., become associated with potential rewards) and motivate actions that as a consequence may satisfy the need. Drinking a glass of water, for example, may prove rewarding when one is thirsty and hence the sight of such a glass may evoke the action.

Although rewards play a crucial role in the development of S–R habits, they may at some point no longer be needed for the execution of a response to a stimulus once the habit is formed and stored in memory. As a result of the repeated execution of an action in response to the presence of a stimulus, cognitive associations develop that tie the two events together. When the association is sufficiently strong, perception of the stimulus activates the memory code or mental representation of the response that, in turn, triggers the corresponding bodily response because these representations match with the sensory-motor cortex that controls the motor programs. At this point, the incentive value that at first motivated the behavior may drop out of the equation as behavior follows the cognitive pathways that were worn out by its motivating power in the past.

Even though habits may rely on such preformed cognitive structures, research has revealed that these structures are not as rigid as one may think. Most notably, some S–R links appear to be conditional on a particular goal or context, and as such promote the translation of goals in behavior. Upon hearing the sound of the alarm clock, someone may stumble to the shower on a workday when she has to get to the office, but may without much thought stumble downstairs to pick up the Saturday paper on the weekend. Depending on the person's goal (work or leisure), the same stimulus

thus may set off a different response that promotes the completion of the goal at hand. This flexibility in switching between different S–R relations is reflected in work demonstrating that people are able to quite easily switch between different well-learned S–R rules according to task instructions and execute them with the efficiency that characterized habitual behavior. In this way, many habits can be regarded as goal-dependent.

Habits as Skills Organized and Directed by Goals

Whereas considering habits as single responses to stimuli may work well for basic actions such as walking to the door when the bell rings, most actions in daily life – such as making coffee or driving to work – are far more complicated. Nonetheless, these actions can be executed in a habitual manner without much conscious thought. How do these skills develop and what do their underlying structures look like?

One way to consider these skills is to regard them as a chain of responses instigated by a particular goal. The habit or skill of making coffee after dinner may then be triggered by activation of that goal and set off a chain-reaction in which each response triggers the next. Putting in the filter may trigger fetching the coffee bin, which in its turn triggers the action of filling the water reservoir. The execution of such response chains, however, has a ballistic character in the sense that previous responses and not the actual behavioral outcomes determine the next action. Relying on such an open-loop mechanism, which does not take into account the result of the performed responses, may be the only way to execute complex behavioral patterns when there is no time to process such feedback information (e.g., when playing a fast sequence of notes on a piano). However, this mechanism only works when the exact same sequence of responses is required. Any small change in the environment or execution of previous actions will lead the mechanism astray and cause the chain to break.

As such changes occur more often that not, researchers have proposed that complex actions are guided by internal models in which top-down and bottom-up processes interact in producing behavior. These models are assumed to be hierarchical and rely on closed-loop processes in which lower order actions are directed by higher order goals. Because of these internal models, perceived results can be compared to their anticipated consequences and subsequent actions can be selected and tuned to produce the desired effect. When driving a car to work for example, the required actions are largely the same (starting the car, turning right at the traffic light, etc.), but slightly different on subsequent occasions (the traffic light is red instead of green, or there is a nasty side wind). Because of closed-loop mechanisms that use perceptual feedback as input for the selection and fine-tuning of responses, people are able to obtain the same goals under different circumstances.

This pivotal role of goals in complex behaviors does not mean that their executions necessarily rely on conscious processes. When learning to drive a car, for instance, conscious selection of actions may be required at first, but conscious involvement may drop out of the equation when this skill becomes overlearned. Although the behavior still relies on a closed-loop mechanism, execution becomes much more efficient and much less dependent on conscious attention. This increased efficiency can be explained in several ways. First, execution of multiple steps can occur faster and more reliable because knowledge about this procedure is efficiently stored in memory and is therefore readily available. An example of such procedural knowledge are scripts that specify the fixed patterns of actions that are executed in much the same manner in recurrent situations. The script of brushing your teeth, for example, may describe the usual sequence of actions of unscrewing the cap of the tooth paste, putting the paste on the brush, putting the cap back on, brushing teeth, gargling, and checking your teeth in the mirror. Once retrieved, this knowledge can guide the execution of the different steps without deliberation on what is going to be the next step. Second, during repeated execution of actions different strategies can develop that produce the same effect by a different route. In order to shift to the proper gear in a car, for example, one first keeps an eye on the velocity and uses certain rules to shift to a particular gear at a certain speed. After practice however, one simply relies on the sound of the

engine and shifts to a higher or lower gear based on its pitch. This simplified rule or shortcut produces the same results through a far more efficient route. In different ways, these processes contribute to the habitualization of complex actions.

At still a more abstract level, realizing a particular goal or outcome may not just require the execution of a sequential behavioral pattern that has to be adjusted to situational changes, but the execution of totally different competing action patterns. Getting to the university, for example, may be accomplished by means of a bike or a bus, which each require a different complex pattern of actions (i.e., unlocking the bike, taking the right route, walking to the bus stop, buying a ticket, etc.). Thus, realizing this goal starts with the selection of one of those two means. After repeated and consistent selection of one particular means, not only the behavior itself, but also the choice for that particular means may become automated. That is, it becomes associated with the goal representation and is therefore triggered when the goal is activated. It has indeed been demonstrated that people respond faster to bike-related words upon presentation of the word university when the goal to go to the university is activated and that the magnitude of this effect increases with the frequency with which that action is used to reach that goal. As such, habitual processes may also involve the selection of habitual means in reaction to the activation of a goal. Hence, habits can not only be goal-dependent, but also goal-directed.

The idea of habits as a form of automatic goal-directed behaviors has been pushed even a bit further. Specifically, although most models on goal-directed behavior assume that goal setting is characterized by a conscious reflection process, and that goal striving is associated with conscious intent, it is suggested that goals are mentally represented and can be activated outside of conscious awareness themselves to then have their effect on behavior. Recurrent and consistent pursuit of a goal upon perception of a specific (social) situation is thought to strengthen the link between the representations of the situation and the goal. Consequently, the mere perception of the situation or environment causes the goal-directed behavior to be triggered directly. Importantly, theory and research on this type of nonconscious goal pursuit considers goals and intentions as distinct concepts that can operate independently from each other, served by different processes. Whereas intentions are the product of conscious deliberation to engage in a behavior or to attain a goal, goals as mental representations of desired states that have become linked to specific means or skills allowing for effective goal attainment without conscious intervention. Therefore, the mere priming of these goal representations causes the person to recruit the associated means or skills directly, and thus goal-directed behavior is launched and guided in the situation at hand without conscious intent and thought.

Habit and Nonconscious Processes

William James once aptly said ‘habits diminishes the conscious attention with which our acts are performed.’ This notion captures the essence of research that examines the relation between habits and nonconscious processes, that is, processes that do not require conscious attention and intention in order to occur. The role of conscious and nonconscious processes in habits have been investigated with several methods. Some research scrutinizes the ability to efficiently perform habits without attentional control in a dual task paradigm, measured by the reduction or lack of interference of habits and skills during performance of a task that demands conscious attention. Other studies have examined the way habit formation and practice modulate the activation in cortical areas involved in controlled (conscious) and automatic (nonconscious) processes. Apart from inferring nonconscious processes of habits from interference tasks and neuroimaging data, there is also research that considers the limited role of explicitly expressed intentions and other subjectively introspective insights of behavior in the prediction and control of habitual behavior. Whereas each method has its own merits and drawbacks, together they may offer a good picture of how our mental life shifts from a conscious to a nonconscious status during forming, establishing, and performing habitual behaviors.

Habits and the Lack of Interference in Dual Tasks

Many habits require several information processing steps in order to become efficient and automated. For instance, the seemingly simple task to push a designated key in response to a specific stimulus among an array of others presented on the computer screen involves the ability to keep the task goal in mind, to encode the proper stimulus, to select the right response and to monitor and process feedback to check whether it produced the desired effect. Things get even more complicated when the task calls for sequential actions (e.g., typing text, playing piano, or riding a car) that need to be orchestrated in the right order and requires the inhibition of a recently performed action in order to switch to the next one. The control of these information processing steps are supposed to rely on several operational components in a processing system that have been labeled in terms such as 'working memory' or 'executive control.' Given that these control processes are often considered to be inevitably conscious and hence, require conscious attention, there should be a connection with the observation that habit practice leads to greater skill at applying the information processing steps up to the level that attentional control is no longer required to perform the sub-steps involved in the habit.

A well-accepted way to examine this idea is to subject participants to a skill learning task in which they either repeat the task until it becomes habitual according to some behavioral criterion (e.g., no further improvement in terms of speed or accuracy). Next, participants are given a secondary additional task that requires attentional control (e.g., short-term-memory task), and performance on both tasks is assessed. In such a dual task setting, interference may result from a single-channel constraint that allows processes to run serially or capacity sharing of resources for different tasks. Thus, interference produces impairment (in speed or/and accuracy) on one of the two tasks when concurrent processes (e.g., monitoring or feedback processing) have to be used to perform both tasks or when processing resources are allocated to one task that leaves a little less for the other task.

By and large, results that show up in this kind of studies are that when one task is overlearned, the participants can perform the other task at the same time with little interference; but there is considerable interference between the additional task and the skill learning task when the learning task is new or not overlearned. Assuming that the amount of interference in a dual task setting represents a measure of conscious control, the findings that performance of a well-learned set of behavioral responses and schemas does not seriously affect the performance of the other task suggest that habits can run and interact with the environment without conscious attention to the processes producing the behavior.

Habits and Decreased Attentional Control in the Brain

Another area of research that may reveal the role of conscious and nonconscious processes in habits concerns studies that employ modern neuroimaging methods such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) to map the functional anatomy underlying practice and skill learning. Thus, several studies have started to explore the changes in brain activity that occur as a result of practice on a range of motor, visuomotor, perceptual, and cognitive tasks. A common framework proposed in this research is that the prefrontal cortex (PFC), anterior cingulated cortex (ACC), and posterior parietal cortex (PPC) are the main areas taking care of attentional and control processes, consistent with theories of PFC function and the involvement of these areas in the distributed working memory system. According to this framework, the cortical areas involved in attentional and control processes are recruited (and hence, activated) to cope with unskilled, nonhabitual actions. After the action has been sufficiently practiced and has become habitual, processes (e.g., feedback processing) and associations that are involved in the control of the habit are more efficiently stored and accessed in the brain and the attentional and control areas fall away, evinced by decreased activation in these areas. On the other hand, increased activation associated with highly practiced performance is primarily seen in task-specific areas such as the primary and secondary sensory and motor cortex. This change in patterns of cortical activations accompanied with practice is

called functional redistribution due to pruning of attentional and control areas.

In one study, for example, participants performed a sequential motor action (producing a specific sequence of finger movements, such as done when playing piano), while the neural correlates of task performance were monitored up to the point that the task was overlearned and thus comprised features of a strong habit. Results demonstrated the functional redistribution of activation pattern: activations in the prefrontal attentional and control areas decreased during skill learning, while increased activations were observed in task-specific motor areas. Interestingly, when participants were asked to consciously reattend to the performance of the overlearned motor task, there was a reactivation of prefrontal areas while the activation in the motor areas was unaffected. These findings thus indicate the association between conscious attention to task performance and the PFC, and the association between implicit or nondeclarative aspects of task performance and brain areas involved in the habit or skill itself.

It should be noted that redistribution of activation patterns is not the only cortical plasticity associated with skill learning and habits. The assumption underlying functional redistribution is that practice and habit formation causes people to develop greater skills in applying the cognitive process required to perform the action or task at issue, or in terms of cortical plasticity, practice facilitates neural efficiency. However, there are studies showing that practice may indeed decrease the activation in the prefrontal attentional and control areas, but sometimes it also leads to activation of cortical areas that were not involved in the initial stage of the task itself.

For example, in a study on effects of the practice of semantic processing on changes in cortical activation, PET was used to image brain activity while participants repeatedly generated verbs to a list of nouns until it had become overlearned. Subsequently, participants had to generate verbs to both the original list and a new list of nouns. The data of this study showed decreased activation of the PFC (in line with the idea that less conscious attention and control is needed to do the task), but increased activation of the Sylvian-insular cortex (an area aiding the automatic production of a verbal response during language processing) that was not involved during the initial stage of the task. These findings suggest that there was a switch away from conscious attentional, semantic processing and selection from an unlimited set of responses to episodic memory-based associations to nouns that minimized attentional semantic processing and the set of possible responses. This increased pattern of activations of brain areas initially not involved in the task has been called functional reorganization, and reflects the situation in which the person has learned a new cognitive skill or shortcut during practice and habit learning to deal with the task at hand.

Both functional redistribution and reorganization of cortical activity patterns facilitate the nonconscious operation of well-practiced skills or habits, but the cognitive processes by which the individual accomplishes this can reflect efficiency in applying a single strategy or learning a new strategy. Furthermore, it has been suggested that specific properties involved in habits have their own time-course of change with practice (e.g., some habits may evolve faster from conscious to nonconscious cognitive processes than from nonefficient to efficient skill performance of the perceptual-motor components of the task). While the study of the neurological underpinnings of habits is in its infancy, these new advances and insights open the possibility to examine how habit formation across different kinds of tasks and behaviors causes people to switch from a more conscious mode of thought and action (e.g., select a means from a set of options to attain a goal) to a more nonconscious mode (e.g., memory-based access to associated habitual means), and how this switch is neurologically implemented and consciously experienced by people.

The Role of Habit and Intention in the Prediction of Behavior

The idea that habits diminish the role of conscious processes has also been examined in research dealing with the prediction of behavior. The main question addressed in this research concerns the extent to which human behavior is under intentional or habitual control. From this perspective, a variety of different behaviors have been investigated that share

the characteristic of being repetitive in nature, such as students' class attendance, purchasing fast food, physical exercise, condom use, drug use, seat belt use, watching TV, commuting with the car, and recycling. In a typical study, people are asked to explicitly express their intentions to engage in a specified behavior and the strength of their existing habits (reflected in frequency of past performance in a given context) and future performance are assessed. Structural equation modeling is used to predict future performance from people's conscious intentions and their habits. The standard result is that habit strength and intentions are independent predictors of the extent to which people perform activities.

The independent contribution of intention and habit in the prediction of behavior suggests that some parts of the behavior involve conscious attention and other parts are subserved by nonconscious processes. For instance, a person may set the conscious plan to watch TV and subsequently execute the action in an automatic, ballistic way controlled by an open-loop mechanism (e.g., switching on the TV and watching soaps until falling asleep). Or, upon having the goal to go to office one automatically takes the car and drive to work in a skilled, routinized fashion, but from time to time one relies on a feedback control system to make sure that one reaches the desired travel goal. In other words, performance of relatively complex habitual behavior or skills is often contingent on a current goal. However, the question here is whether habitual goal-directed behaviors can also be instigated and guided by nonconscious processes or always require conscious attention and intention in order to occur.

A few studies using the structural equation modeling technique may offer an answer to this intriguing question. In approaching the issue of habitual control of behavior from a slightly different but important view, these studies hypothesized that as the same behavior is more frequently executed in the past and increases in habit strength, it is less guided by conscious intention and attention to perform that behavior. Habit strength thus moderates the relationship between intentions and subsequent goal-directed behavior: a hypothesis that requires a test showing that habit and intention interact in their prediction of later behavior – instead of merely showing that habit predicts behavior over and above a measure of intention. In a study exploring this possibility, inhabitants of a village nearby a larger city filled out a survey that required them to indicate their intentions and habit strength of using the car to commute to the city. Next, the respondents' travel behavior was monitored for a few weeks so that their car use could be predicted by their intentions and habit strength. Results clearly demonstrated that a measure of habit indeed interacted with intentions in the prediction of future travel behavior: when habit was strong intentions did not predicted car-commuting behavior, whereas the behavior was predicted by intentions when habit was weak.

This interactive pattern of habit and intention in predicting behavior has also been observed for other types of human conducts, such as social interactions in interpersonal relations, fast-food consumption in cafeterias, and drinking alcohol when going out, and especially shows up when the goal-directed behavior is repeatedly and consistently performed in the same context. The importance of context stability in triggering goal-directed behavior without conscious intent underscores the idea that habitually pursued goals can guide associated means and skills without much involvement of conscious attentional processes when entering and interacting with the context at hand.

Habits and Subjective Reports of Consciousness

A fourth area of research that investigates the conscious and nonconscious parts of habitually performed behavior resorts to operationalizations that typify consciousness, such as verbal reportability. For instance, an extensive literature on implicit learning shows that, while intentionally repeating actions in line with the experimenter's instructions, people learn associations between stimuli and responses and even rules of responding to complex sequences of stimuli without awareness of what is being learned. In a task often used to investigate implicit learning, the serial reaction time task, participants tap a key when a stimulus appears on the screen. The stimulus can appear in one of four locations, corresponding to four response keys. Unannounced to the participants,

the stimuli appear in a repeated sequence or not. In general, participants seem to learn the sequence of spatial locations (i.e., get better at the task with practice when there is a repeated sequence) even when they are not able to verbally describe it.

Implicit learning research suggests that people can acquire knowledge relevant to establish skills in the absence of conscious awareness. Many of our habits or skills performed in daily life involve closed-loop or feedback processes to have their desired effects. Some parts of our habits and skills are difficult to mentally access, as they are represented in nondeclarative memory. Other parts are represented in declarative memory, and thus can be more easily reflected on. For instance, if one writes a letter it may be impossible to report on how one controls the muscles of one's hand and fingers when moving the pen, but one may become aware (even with one's eyes closed) of the curves that one makes to shape certain letters (e.g., try it with the letter P). Similarly, car-driving may involve actions that vary in the ease with which they can be recalled from explicit memory. Based on this distinction, most researchers agree that some parts of adjustment processes underlying skill performance occur outside of awareness but other in the presence of awareness.

However, there is evidence suggesting that adjustments of which we can become aware of remain unconscious, hence questioning whether our conscious experiences tell us the true story about how we regulate parts of our skills and habits. In a study on hand movement monitoring, participants were given the goal of drawing a straight line on a computer screen (a well-practiced skill that most people already learn early in their life). Participants could not see their hand or arm, and received false visual feedback via a mirror presentation of the computer screen about the trajectory of their hand movement. Thus, participants had to make considerable deviations to achieve their goal of drawing a straight line. Whereas participants displaced their hand in the opposite direction for producing the desired goal state (a straight line), verbal reports showed that participants were unaware of making deviant manual movements in response to the false feedback – in fact, they claimed to have made straight movements. These findings indicate that people adjust their skilled actions in response to deviations but that this type of action control underlying the achievement of goals can occur without conscious awareness.

The findings that people can control their skilled actions and habits in the absence of awareness indicate that, when goals are pursued regularly, the need to pay conscious attention to details dwindles. In fact, they show that when specific well-practiced responses are bound to fail, conscious processes are not always called to the fore to complete the skill and to attain the goal. According to the idea that habits are a form of automatic goal-directed behaviors however, it may even be possible that the goals directing the skill themselves are activated nonconsciously, and hence, people are not even aware of controlling the habitual behavior as a result of the goal. Several studies have tested this possibility.

Capitalizing on the notion that people habitually recruit and execute different skills to attain their goals to achieve and perform well, in one study the relation between achievement priming and actual performance was studied. Participants were exposed to words such as 'strive' and 'succeed' as part of a word search task to prime the goal of achievement. Next, they were offered the opportunity to display their performance skills (finding as many words as possible in an anagram puzzle task). Results indicated that participants primed with the achievement goal outperformed those who were not primed with the goal. Of importance, after the experiment participants offered insights into their explicit thoughts about their commitment to perform well on the task, and these conscious ratings were unrelated to the priming effects. These results indicate that goals facilitate the utilization of skills and habitual procedures without conscious awareness of the activation and the operation of the goal, even though these skills have not been previously applied to the task at hand. Similar findings have been reported in other studies that identified the social triggers that repeatedly influence people to pursue goals (e.g., performing well, earning money), such as the observation of another person's goal pursuit or the mere perception of important others (e.g., partners, parents) who have frequently encouraged us to pursue specific goals in the past.

The observation that explicit thoughts about behavior do not mediate the environmental priming effects on actual goal-directed habit performance suggests that goals can guide action schemas nonconsciously. However, explicit thought may not mediate the effects for other reasons than the outsourcing of cognitive processes to the unconscious. First, as there is a time-lag (often more than one minute) between action performance and the verbalization of explicit thoughts, it may be the case that people have forgotten all about why they performed the behavior. Second, even though participants may be able to recall the goals causing their behavior, it may also be the case that they are not willing to report them accurately (due to social desirability or demand characteristics). In other words, the source that is responsible for the emergence of goal-directed habits is conscious, but people are not able or motivated to report this conscious (goal) source.

One way in which researchers have tried to circumvent this problem is to render the triggering source of goal-directed habits unconscious itself. For example, in a study on the utilization of social perception skills, participants were subliminally primed either with an impression formation goal or not by exposing them to words such as impression, evaluate and judgment presented outside the most sensitive part of the retina (parafoveally) for very short time intervals. Crucially, subliminality tests showed that participants could not consciously perceive the stimuli. Next, to explore whether the nonconscious goal encouraged participants to rely on their social perception skills, they read trait attributes of a fictitious person. It is known that explicit task goals to form impressions cause individuals to form evaluative judgments as soon as information is provided about a target person. When a subsequent judgment of the target is required, they rely on the available judgment that was formed online, or otherwise consult memory of the target to arrive at the judgment. The subliminal presentation study showed that goal priming indeed led to more online rather than memory-based judgments. Other studies using different subliminal priming tasks have documented similar results for other habitually pursued social goals, such as achievement, cooperation, and socializing. Whereas effects of subliminal stimulation on cognition and behavior are still open for debate, especially social behavior resulting from higher mental processes such as our goal pursuits, the findings alluded to above provide compelling evidence in support of the hypothesis that people can engage in habitual behavior instrumental in attaining their goals without awareness of the source of these effects, that is, whereas we may reflect on, and become aware of the behaviors that we habitually perform, it does not mean that we are conscious of the causes and the processes underlying them.

Habit and Sensemaking

Although habits may be controlled by the situation and emerge from nonconscious goal-dependent processes, people can use their habits to gain insights in their goals and other personal dispositions. In that case, the causal pathway between habits and consciousness is reversed such that people reflect on the products of higher mental processes in the form of their own past behavior and outcomes to infer personal attributes and arrive at self-insight. As a general rule, self-insights manifest themselves if personal beliefs and desires gain access to consciousness, and such access is gated by focus of attention that occurs when people are, for example, directly asked to indicate their current goals (e.g., What do you want to drink or eat?) or are otherwise forced to reflect on their experiences (e.g., when the situation is ambiguous and calls for a speedy decision). The notion that self-insights are affected by knowledge of habits is so obvious that it comes across as fairly trivial. After all, it would be a waste not to use our past experiences and the contents of memory. However, in the context of habits there are three mechanisms typical for the working of the mind that offer intriguing insight into the way of how reflection on habits promotes self-insight: self-perception, experienced ease of retrieval and authorship ascription.

Self-Perception

People often have limited introspective access to the causes and processes of their habits. When internal causes of behavior are weak or ambiguous, people may be forced to draw inferences about

them from their own behavior and the situation in which it occurs. Indeed, such self-perception processes are rather pervasive. When we are not sure how we feel about a specific behavior, or we do not know whether we want to engage in it, our own behaviors may offer an answer. For instance, when one is asked whether one likes spicy peanuts and does not know the answer, one could turn to observing one's own behavior (e.g., I frequently ate spicy peanuts in the past, so I must find eating them desirable).

Importantly, people judge whether their feelings really reflect their potential causes of habits or whether it was the situation that made them perform them. When an external cause can be identified (e.g., I only frequently eat spicy peanuts because they are served for free in the local pub), information about the frequency of behavior is discarded. Thus, only when no external cause can be brought up, self-perception of habits is likely to provide information about internal causes of ones own behavior. These self-perception effects have been demonstrated in many studies, including the observation that frequency of past behavior often correlates fairly with attitudes and intentions to perform the behavior. Interestingly, attitudes and intentions that follow from self-perception do not always predict future behavior. Instead, frequency of past behavior predicts future behavior directly, suggesting that actual behavior is under control of habitual, nonconscious processes.

Experienced Ease of Retrieval

Rather than assessing and relying on information about the frequency or amount of previous behaviors, habits may inform people about their goals and other personal dispositions by the experienced ease of retrieving instances of the behavior from memory. In this case, cognitive fluency or the experiential state accompanying the working of the mind, rather than the content of the recall itself may serve as a source of information to arrive at self-insights.

In a study demonstrating this effect, participants were asked to list either 6 or 12 instances in which they behaved assertively. Pretests indicated that recalling 6 examples was experienced as relatively easy, while recalling 12 examples was experienced as difficult. After retrieving the examples, participants rated how much difficulty they had experienced retrieving the examples and evaluated their assertiveness on a 10-point scale. If participants were merely to rely on the content of recall, they would report higher assertiveness after recalling 12 rather than 6 examples. However, this is not what happened. Self-ratings of assertiveness showed that participants perceived themselves as more assertive after recalling 6 rather than 12 examples of assertive events. Apparently, participants who were asked to retrieve 6 instances concluded that they were pretty assertive simply because retrieving the instances was experienced to be easy.

According to the principle of ease of retrieval, then, habits can produce self-insight in two ways: if previous instances of the occurrence of a habit are easy to retrieve, positive feelings about the habitual behavior may emerge, whereas experiencing difficulties in retrieving these instances may tell one that the behavior is not a future goal one wants to engage in. Whereas the former effect implies that people have access to the occasions or situations in which they performed the habit, the latter suggests the absence of such access. In that case, people may rely on the frequency or amount of previous behaviors to arrive at self-judgments.

Authorship Ascription

While self-perception and experienced ease of retrieval point to the human capacity (and willingness) to arrive at the content of conscious goals by reflecting on one's own habitual behaviors performed in the past, they are not the source of people's every day life experience of causing their actions and outcomes, that is, the experience of authorship. The experience of authorship is derived in part from interceptive sensations of the body's movement that occur both before and after action. Such sensations are supplemented by visual and auditory feedback, as we can often see and hear the consequences of our actions. Although the absence of experiences of agency and goal achievement is an essential part of habits (even of goal-directed ones), people frequently have these experiences. So, if we assume that our habits arise from nonconscious sources and operate outside our awareness how do we arrive at

experiences of agency and hence, believe that conscious will causes behavior?

Our belief in agency or willful causation is thought to originate from the human capacity to foresee events, and hence, to anticipate goal attainment. When the goal is attained that we intended to pursue, we are likely to infer that we caused it because it matches our previously activated goal state. In more conceptual terms, one experiences personal causation of an observed action effect (where action effect refers to any possible outcome that may arise from concrete skilled actions) because the representation of the effect is primed before one performs the given action. Whereas this matching process of predicted and actual goal attainment offers a key to understanding how people establish a sense of agency, it can guide these experiences independently of direct sensation and actual causation, resulting in illusory experiences.

In a study illustrating this possibility, participants practiced a computer task in which they themselves and the computer independently controlled a rapidly moving square on a display. At a certain point in time, participants had to stop the movement by a simple key-press. The stopped position of the square could either be caused by the computer or the participant. Accordingly, the stopped position could be conceived of as the desired effect that matches the goal controlling participants' action of pressing the key. The location of this position was subliminally primed just before participants pressed the key and saw the presented square. Results showed that priming of the position of the presented square produced a sense of agency associated with stopping the square that was independent of actual control.

It should be noted though, that in the studies discussed above, the prime occurred always briefly before participants performed the skilled action. People can hold an item in short-term memory for no longer than a few seconds without rehearsal. This brief time window suggests that a prime that appears far in advance may not yield the experience of agency. Indeed, it has been shown that the effects of priming on agency experiences only show up if effect information is primed 5 or 1 s in advance, but not with a time interval of 30 s. This suggests that the feeling of agency during action performance derives from a match between the prime and observed effect occurring close in time. Because individuals have limited or no direct conscious access to the operating procedures guiding their actual habits, the matching signal of primed and observed effect information is a key source for grasping a sense of agency, especially when the two events are close together in space and time and thus are more likely to be perceived as causally related.

Conclusions

Habits are the result of practice. They provide us with a well-learned set of skills and schemas that are often contingent on, and orchestrated by current goals and can run and interact with our environment without us paying conscious attention or forming an explicit intention or plan to perform the behavior. Accordingly, habits are commonly accompanied by decreased awareness of the environmental events and action components involved, which suggests that we delegate habits to the unconscious. Such habits may be difficult to overcome by an act of conscious will. Furthermore, although many of our habits can occur without conscious intervention, people sometimes reflect on their habits. Such reflection produces self-insights that offer information about our goals and enhance our feelings of agency. However, our sense of agency may be tricked, as we are prone to falsely infer that our goals cause the execution of our skills or habits when these goals are observed as outcomes after performing them. We can therefore never trust our insights as to whether our habits are goal-directed or not, especially because these goals can operate outside of conscious awareness themselves.

Acknowledgment

The preparation and writing of this article was financially supported by the Netherlands Organization for Scientific Research as part of the VENI-VIDI-VICI Scheme (grants 451-06-014, 452-02-047 and 453-06-002).

Suggested Readings

Aarts H (2007) Health behavior and the implicit motivation and regulation of goals. Health Psychology Review 1: 53–82.

Bargh JA and Ferguson MJ (2000) Beyond behaviorism: On the automaticity of higher mental processes. Psychological Bulletin 126: 925–945.

Berridge KC (2001) Reward learning: Reinforcement, incentives, and expectations. In: Medin DL (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory, vol. 40, pp. 223–278. San Diego, CA: Academic Press.

Cooper R and Shallice T (2000) Contention scheduling and the control of routine activities. Cognitive Neuropsychology 17: 297–338.

Hommel B, Muesseler J, Aschersleben G, and Prinz W (2001) The theory of event coding (tec): A framework for perception and action planning. Behavioral and Brain Sciences 24: 849–937.

James W (1890) The Principles of Psychology. London: Macmillan.

Jeannerod M (1997) The Cognitive Neuroscience of Action. Oxford, England: Blackwell.

Kelly AMC and Garavan H (2005) Human functional neuroimaging of brain changes associated with practice. Cerebral Cortex 15: 1089–1102.

Pashler H and Johnston JC (1998) Attentional limitations in dual-task performance. In: Pashler H (ed.) Attention, pp. 155–189. Philadelphia: Taylor & Francis Press.

Powers WT (1973) Feedback: Beyond behaviorism. Science 179: 351–356.

Wegner DM and Sparrow B (2004) Authorship processing. In: Gazzaniga M (ed.) The Cognitive Neurosciences, 3rd edn., pp. 1201–1209. Cambridge, MA: MIT Press.

Biographical Sketch

Henk Aarts is trained as an experimental social psychologist at Nijmegen University where he worked on habit and decision making and received his PhD in 1996. He worked at Eindhoven University of Technology and Leiden University. Since 2004 he has been a full professor in social psychology at Utrecht University. His work deals with several topics related to the role of goals in automatic processes of social cognition and behavior and is published in fundamental and applied journals. One recent discovery in his research program concerns the notion that in contrast with what often is assumed, conscious intentions do not play a strong causal role in behavior as well as that people infer goals from their own and others' behavior. This suggests that although goals play a pivotal role in human behavior, these goals may well operate outside of consciousness. In his research he tries to unravel core aspects of this intriguing and important topic.

Ruud Custers received an education in human–technology interaction at Eindhoven University of Technology, where he graduated Cum Laude on work investigating the role of memory in the formation of judgments about environments. Subsequently, he moved to Utrecht University to pursue a PhD in experimental social psychology. He received his PhD Cum Laude in 2006 for his dissertation on the underlying mechanisms on nonconscious goal pursuit, which mainly focused on the role of affective signals in this process. He published several papers in fundamental journals, of which one was regarded as the best paper of the year on social cognition by the International Social Cognition Network in 2006. As an assistant professor at Utrecht University, he continues to study the processes that allow people to pursue goals without conscious awareness.

History of Consciousness Science

B J Baars, The Neurosciences Institute, San Diego, CA, USA

Introduction

Scientific attitudes toward consciousness emerged in several distinct phases.

1. Beginnings
   The first phase is a very rich intellectual history, long before systematic scientific research began, starting in the early centuries of written thought in Asia and the West. Most current topics in consciousness science can be traced back to this remarkable intellectual tradition.
2. Early scientific findings
   A second phase begins with the Renaissance, when careful studies of brain anatomy, the optics of the eye, color perception, visual perspective and the like put those topics on a sound foundation. Visual perspective and color perception were routinely discussed in terms of the observer's conscious experiences.
3. A Golden Age of consciousness science
   The period of 1780–1910 was a kind of Golden Age, with systematic scientific studies of hypnosis, conversion disorders, dissociativity, linguistics, memory, sensory psychophysics, sensory physiology and neuroanatomy, Broca's and Wernicke's aphasia, and much more. This period is well summarized in William James' *Principles of Psychology* of 1890. Most of it fits James' slogan that 'psychology is the science of Mental Life,' meaning conscious mental life.
4. Behaviorist rejectionism
   For both scientific and professional reasons, the Golden Age was met with decades of behavioristic rejection of the very topic of consciousness, beginning around 1913 with John B. Watson's 'behavioristic manifesto.' Watson and later B.F. Skinner argued that almost all psychology before the twentieth century was unscientific because it dealt with human subjectivity. 'Consciousness is nothing but the soul of theology,' wrote Watson in 1913. It was viewed as inherently flawed, and therefore to be discarded from any true scientific approach.

   Similar movements arose in physiology and biology with I.P. Pavlov, T.H. Huxley, Jacques Loeb, and others. The 'unscientific' nature of nineteenth century studies became a standard cliché, which helped to rationalize the new academic professions of psychology, philosophy, and brain biology. Prior work in each of those fields was performed by amateurs of genius like Charles Darwin, William James, and Hermann Helmholtz. The new academic professions required new rationales, however, and behaviorism provided them.

   Nonetheless, historians have repeatedly questioned this *post-hoc* rejection of an extraordinary age of discovery before 1900.
5. The cognitive revolution
   Cognitive science began in the 1950s and rose to dominance by the 1970s. It retained the experimental rigor of behaviorism, but encouraged theoretical inferences about such topics as 'working memory,' 'syntactic rules,' and indeed 'consciousness.' While cognitive approaches slowly replaced behaviorism, it was not until the end of the twentieth century that scientific studies of conscious experiences were visibly recovering from decades of rejection. However, cognitive science only gradually began to broach the topic of consciousness by way of euphemisms like 'explicit cognition,' 'episodic memory,' 'strategic processes,' 'supraliminal perception,' and the like. The private, subjective aspect of these processes was still widely neglected.
6. Rediscovery
   A series of important discoveries, technical improvements and new insights triggered a period of rediscovery from the 1980s onwards. Brain imaging techniques made it possible to integrate the psychology and physiology of the living brain. A number of prominent scientists joined the quest to understand conscious functions. Today, a PubMed search using the terms 'conscious' and 'consciousness' brings up more

than 50 000 publications. In 2005, *Science* magazine listed 'the biological basis of consciousness' as one of the fundamental questions in contemporary science.

Beginnings

The history of human thought simply brims with ideas and observations about human consciousness, beginning with Aristotle and Plato in the fourth century BCE in the West, http://plato.stanford.edu/entries/aristotle-psychology/, http://plato.stanford.edu/entries/ancient-soul/#3.1, and with Vedanta, Buddhism, and related views in Asia around the same time. In-depth exploration of human consciousness therefore begins with the spread of writing about the human mind; whether there is an earlier, unwritten history of ideas on the subject is simply unknown, but it seems quite likely. Most topics of modern research were already named in the earliest languages we know, suggesting ancient origins for concepts like thinking, memory, visualization, meaning, sensory perception, meditative practices, goals and plans, imagination, emotion, and motivation. The very word 'cognition' (knowledge) comes from the same Indo-European root as the ancient Sanskrit 'jñana.' All modern Indo-European languages have analogues for those words.

More than half of the semantic categories in Roget's Thesaurus are psychological in nature, covering 'consciousness, will, intellect, and affections.' All those terms refer to conscious, reportable events. These terms are not just vague folk ideas; they were systematically analyzed and refined by philosophers from Aristotle to the Scholastics, Kant, Locke, and James. According to Chomsky, the first transformational grammar was devised by a ninth century Indic linguist named Bhartrhari.

Thus there is a great stream of interesting ideas about conscious experience to be found over many centuries.

Scientific Studies

While significant discoveries go back as far as Hippocrates in the fourth century BCE, systematic scientific studies begin with the Renaissance. Sensory consciousness was explored in the first accurate brain anatomy by Andreas Vesalius. The thalamus was proposed to be a place for the interaction between sensory inputs and the neocortex, as we still believe today. Newton's studies of sunlight falling on glass prisms revealed the color categories that the brain imposes upon the light-wave continuum. That finding inspired the first color theories, designed to account for the categorical nature of the spectral colors. Likewise, Descartes' experiments on sheep's eyes revealed the optics of the mammalian eye, including the startling fact that the optical projection on the back of the eye is both aberrant and upside-down.

The nineteenth century witnessed a wave of remarkable discoveries: The stunning phenomenon of hypnosis, for instance, or surgical anesthesia, mental imagery, the variety of human sexual feelings, unconscious emotions and much more. The nineteenth century was a golden age of consciousness science, including contributors like Fechner, Helmholtz, Charcot, Janet, Broca, Wernicke, Darwin, Galton, Mueller, Freud, Jung, Donders, Wundt, Cajal, and numerous others. But just after the death of William James in 1910, the central role of conscious thought was purged from the sciences. It became an academic taboo. Today, almost a century later, it is still making its way back.

William James stands out as the culmination of consciousness science. He summarized its first century in his masterpiece, *The Principles of Psychology* (1890), often considered the greatest work of psychology in the English language. The 1400 pages of the *Principles* are full of important facts and ideas about attention and memory, concept formation, perception, the brain, 'fringe consciousness,' disorders of the will, reaction time, the senses and the stream of consciousness. Essentially all the phenomena described in the *Principles* are accepted today; it is a compendium of facts about human consciousness.

James' abridged version, the *Briefer Psychology* (1893), became the most popular introductory text for the first generations of academic psychologists. Its influence can be clearly traced into the next century. With the 'cognitive revolution' in psychology, most of the empirical phenomena James discussed were rediscovered.

Unconscious Processes

Nevertheless, the nineteenth century encountered some difficult conceptual problems, as one might expect in a young science. For example, unconscious mental processes were often viewed as paradoxical.

From our current perspective, the experimental study of conscious events requires comparison conditions, such as comparisons between conscious and unconscious stimuli in dichotic listening or binocular rivalry. To study consciousness 'as such' we need to treat it as a variable of interest in its own right. Thus consciousness science goes hand-in-hand with the study of unconscious comparison conditions.

The nineteenth century found it very difficult to imagine unconscious information processing. A variety of unconscious brain events were known, such as habit formation after practicing a predictable mental task. But unconscious events were generally viewed as mechanical or 'unintelligent' physical phenomena, not as complex, symbolic processes involving language or reasoning. An intelligent, symbolic unconscious is very much a later twentieth century idea, one that only spread with the digital computer. Broadly speaking, the nineteenth century was still deeply committed to the Aristotelian notion of the conscious intellect as a rational faculty that served to control a lower, unconscious, instinctual, animal mechanism. For example, William James provides ten excellent empirical arguments for unconscious mental processes, but then maintains that all ten phenomena are not *really* unconscious. Seemingly unconscious mental events were either fleetingly conscious (but too fast to be recalled), or they had become 'physical' – they had somehow jumped the divide from the conscious mind to the physiological brain. Thus James, for all his remarkable insights, still thought that the mind must be entirely conscious, and that unconscious events must correspond to the physical brain. He thereby opened himself up to metaphysical dualism, a smoking stick of dynamite for later generations. The period of behaviourism can be seen as a resolute stand for a consistent physicalistic foundation for psychology and brain science. It was a reaction to the towering figure of William James.

Toward the end of the nineteenth century other scientific thinkers – notably Pierre Janet and Sigmund Freud – began to infer unconscious processes quite freely, based on observable events such as posthypnotic suggestion, conversion hysteria, multiple personality, slips of the tongue, motivated forgetting, and the like. Freud's insights have achieved extraordinary cultural influence. While Freud popularized the unconscious in the years after 1900, the Freudian unconscious was considered to be a 'cauldron of seething excitations,' a source of primitive, magical, dreamlike images and impulses. Scientific agreement on unconscious information processing arrived quite late in the twentieth century. This made it difficult to study conscious experiences 'as such,' since there were no plausible comparison conditions for conscious visual percepts, for example. That difficulty has only been overcome in the last several decades of research.

''Introspectionism'': A Presentist Myth about the Nineteenth Century

The widely taught history of ''nineteenth century introspectionism'' has come under intense criticism in recent years. A number of scholars believe that the very term 'introspectionism' involves a *post-hoc* 'myth of origins' that was popularized in the first half of the twentieth century by behaviorists like John B. Watson. Since the nineteenth century as 'introspectionism' is still widely taught, this profoundly misleading impression is still a serious problem for the scientific study of consciousness.

Ernest R. Hilgard has written that

> . . . despite the widely and wrongly accepted cliché that (Wilhelm) Wundt's experimental psychology was the epitome of introspection. . . examination of the nearly 180 experimental studies published in Philosophischen Studien between 1883 and 1903 (by Kurt Danziger). . . turned up only four that used qualitative introspective data, and Wundt himself reported shortcomings in two of these. Most of the studies were quite objective, as in reaction-time measures, which required no verbal reports based on inner perception.
>
> (Hilgard, 1987)

Methods involving self-report are universally employed today, with appropriate methodological care. Scientists have used such reports about conscious experiences for more than two centuries, ever since the beginnings of psychophysics,

with high reliability and empirical utility. Almost everything we know about the sensory brain depends upon that kind of evidence. Sensory reports about conscious events are accurate to the minimum physical limit of single photons activating single retinal receptors, and in audition, down to the Brownian motion of air molecules in the inner ear canal. We can also report an extraordinary range of other conscious events, like episodic memories, inner speech, short-term memory, explicit problem solving, mental images, short-term intentions and the like. For the subjects who make those reports they always refer to personal conscious experiences. Self reports must be obtained under optimal experimental conditions, of course, such as minimum distraction and a short time lag between the event and the report. Nevertheless, there is clearly an enormous range of experiences that humans routinely report with high accuracy. In that sense, 'introspection' is just part of the armamentarium of science. Introspective reports are routinely used in medicine, human–computer interfaces, optometry, audiology, music, and the arts.

Thus 'introspective reports' are not different in principle from other useful empirical indices in science and technology. They have their pitfalls, to be sure, but with some basic precautions they are reliable and useful. There are constant efforts afoot to improve self-reports methodologically, using signal-detection theory, process-dissociation, and novel brain correlates of conscious events.

There are also well-known domains where consciousness reports do not work well. For example, self-reports are not reliable in the case of vague intentions to act, for unconscious or quasiconscious events, for such topics as personality factors, practiced automatisms, and fleeting mental images. Introspective self-reports are therefore often useful, but not always, just like any other measure in science.

This point was clearly understood by most scientists in the nineteenth century. ''Introspectionism'' was the label given by some 20th century psychologists to the 19th century consciousness science. John B. Watson, B.F. Skinner and others claimed that 19th century psychology was tainted by the effort to see inside of the mind. Introspectionism therefore came to be a term of disparagement. In 1966 the psychophysicist S.S. Stevens was caustically criticized for believing that psychophysical reports were about the conscious experiences of his subjects, as they plainly are (Stevens, 1966). Some nineteenth century psychologists could accurately be called introspectionists, notably the Wuertzburg School, in its efforts to characterize experiences of fleeting thoughts, and Edward Titchener's efforts along the same lines. Some of Titchener's experiments were vitiated by a lack of reliability and by large individual differences. Wilhem Wundt's brief popular introduction to psychology also started with some experiential demonstrations, leading to the fundamental misunderstanding that Wundt was an 'introspectionist,' a notion he vigorously denounced. These examples of 'introspectionism' were always distinct minority positions, and in the case of Wilhelm Wundt, they reflect a basic misunderstanding of his vast body of objective experimental work.

The Imageless Thought Controversy

Experiential reports did not lead to a scientific consensus when psychologists tried to discover whether there could be thoughts without conscious images or feelings. Woodworth and Schlossberg described their efforts to answer that question:

> When O's (Observers) were asked what mental images they had (while solving a simple problem) their reports showed much disagreement, as we should expect from the great individual differences found in the study of imagery . . . Some reported visual images, some auditory, some kinesthetic, some verbal. Some reported vivid images, some mostly vague and scrappy ones. Some insisted that at the moment of a clear flash of thought they had no true images at all but only an awareness of some relationship or other ''object'' in (a) broad sense. Many psychologists would not accept testimony of this kind, which they said must be due to imperfect introspection. So arose the 'imageless – thought' controversy, which raged for some years and ended in a stalemate.

These problems seem to reflect an attempt to push experiential reports beyond their useful limits. But all measures have limits. Apparent failures like the 'imageless thought controversy' must be balanced against solid successes in fields like perception, psychophysics, imagery, and immediate memory. In those areas experiential reports are highly reliable.

To use the term 'introspectionism' as an invidious label for the entire nineteenth century gravely distorts an extraordinarily productive period of science. The myth of an unscientific nineteenth century continues to be taught. Students still learn that Watson and Pavlov became the founders of psychological science by purging the vitally important topic of human subjectivity. The myth of origins continues to shape attitudes to the study of conscious experiences.

Philosophical Barriers

For twenty-six centuries philosophers have debated the relationship between conscious experiences and the public or intersubjective world, the world of physical science. Plato and Aristotle analyzed mind–body questions in depth. Similar views were advanced in Asia at the same time. They still dominate contemporary philosophy.

The mind–body debate is commonly posed in terms of apparent paradoxes raised between the first person perspective of private experience (conscious, subjective) and the third-person perspective of the physical sciences (public, intersubjective, and physicalistic). Everyday thinking flips naturally between those viewpoints, easily taking in propositions like 'taking an aspirin helped my headache.' But a headache is a subjective experience, while an aspirin is a physical object. When one attempts to rigorously relate first-person to third-person events, a variety of paradoxes are believed to arise, sometimes called the 'Hard Problem.'

It is important to note that mind–body paradoxes are not empirical but conceptual. Conscious experiences such as color perception have been successfully studied since Newton, and no empirical paradoxes – like wave–particle duality or quantum entanglement – have been encountered so far. Thus experimental studies do not seem to be affected by philosophical puzzles. A traditional scientific approach is to sidestep such philosophical debates if they do not seem to be empirical relevant. Over time, testable questions usually lead to progress.

However, William James and his contemporaries were repeatedly drawn into mind–body questions. At various points in the *Principles* James tried to reduce all mind–brain phenomena to conscious experiences (mentalism), while at others he tried to reduce them to brain processes (physicalism); this dual reduction led him to the classical position of mind/body dualism, much against his will. Conflicting reductionistic commitments created endless paradoxes for James. In some of his last writings (1904) he even suggests that 'consciousness' should be dispensed with altogether, though momentary conscious experiences must be retained.

These different claims are so incompatible as to rule out a clear and simple reductive foundation for psychological science. Thus many psychologists found James to be a great source of confusion, for all his undoubted greatness. And James himself felt confused. By 1893 he was writing in despair, ''The real in psychics (conscious experiences) seems to ''correspond'' to the unreal in physics, and *vice versa*; and we are sorely perplexed'' (p. 460).

A pragmatic approach to consciousness science can simply sidestep mind–body debates, just as Isaac Newton sidestepped philosophical questions about gravity, and early biologists sidestepped vitalistic challenges from philosophers. Productive science can often proceed without solving purely philosophical questions.

Behaviorism and the Rejection of Conscious Experience

Behaviorism is an attempt to explain human nature in terms of physical stimuli and responses. Various physicalistic approaches arose in the late nineteenth and early twentieth century in physiology, biology, and psychology. Behavioristic views became dominant with the rise of I.P. Pavlov, John B. Watson, and later, B.F. Skinner. They largely erased the rich psychology of consciousness developed in the nineteenth century, as summarized in James' *Principles*.

As John B. Watson wrote in 1925

> . . . the time has come for psychology to discard all reference to consciousness. . . it is neither a definable nor a usable concept, it is merely another word for the 'soul' of more ancient times. . . Consciousness is just as unprovable, as unapproachable as the old concept of the soul . . . the Behaviorist must exclude from his scientific vocabulary all subjective terms such as sensation, perception, image, desire, purpose, and even thinking and emotion. . .

This program was carried out with astonishing success over the twentieth century, with the vigorous support of philosophers like the early Bertrand Russell, Gilbert Ryle, Ludwig Wittgenstein, and others. Its roots can be found as early as David Hume's *Enquiry Concerning Human Understanding* (1748), which literally suggested 'consigning to the flames' books of metaphysics that did not result in testable hypotheses or in mathematical propositions. Some of that book-burning passion inspired radical behaviorists as well.

While small in number, radical behaviorists were extraordinarily influential. Pavlov, Watson, and Skinner were popularly famous for decades, while moderate behaviorists were never known to a wider public. The radicals thereby put others on the defensive. They commonly accused traditional psychologists of being unscientific. More than a century of solid scientific progress was branded as 'introspectionism' or 'mentalism,' in contrast to 'respectable behaviorism.'

Watson's scientific purge of 'sensation, perception, image, desire, purpose, and even thinking and emotion' was carried out with great thoroughness in Britain, the United States, and the Soviet Union. Other psychologists believed that the complete rejection of consciousness was too extreme. But time and again the rejectionist camp rose to greater prominence. In the upshot, a small but vigorous minority purged psychology and brain physiology of its most central problems for most of the century. By keeping moderates on the defensive they made empirical progress nearly impossible.

In biology, the rejection of consciousness found its own champions. T.H. Huxley, I.P. Pavlov, and Jacques Loeb were among the pioneers, though others, like Sir Charles Sherrington, dismissed it as extreme and implausible. Because scientific psychology and philosophy arose as academic professions at the same time as behaviorism, the impact was by far the greatest in those fields.

Starting in 1913, with Watson's 'behavioristic manifesto,' 3 years after the death of William James, behaviorism began to spread rapidly among academic psychologists. In the following decades the lexicon of psychological common sense was purged from science. *Roget's Thesaurus* can help us to give us a rough estimate of the number of words and concepts that were ruled out. *Roget* divides the entire English vocabulary into six great semantic classes, from Space, Matter, and Abstract Relations to Intellect, Volition, and Affections. The last three categories are purely psychological – in Watson's words 'mentalistic' – and therefore unscientific to behaviorists. They cover 63% of the Thesaurus. In cleansing the vocabulary of psychology, Watson therefore proposed to discard almost two-thirds of the words we use to describe human experiences and actions. Consciousness went first, but it was quickly followed by volition, attention, self, imagery, planning, thinking, knowledge, inner speech, intentions, expectations, memory, and perception. Everything that could not be observed directly was to be tossed out.

Edna Heidbreder wrote in 1933 that

> Within American psychology the rise of behaviorism has been both conspicuous and important. . . . Indeed, one of the signs of the vigor of this extraordinarily vigorous movement is the way in which it has managed to get its attitudes recognized by those who oppose its fundamental doctrines.
>
> (Heidbreder, 1933)

Behaviorism lasted well into the 1980s, and left behind a lingering suspicion of conscious experience in the sciences even today. As George A. Miller said about the 1950s

> (Behaviorism) . . . was the point of origin for scientific psychology in the United States. The chairmen of all the important departments would tell you that they were behaviorists. Membership in the elite Society of Experimental Psychology was limited to people of behavioristic persuasion; the election to the National Academy of Sciences was limited either to behaviorists or physiological psychologists, who were respectable on other grounds. The power, the honors, the authority, the textbooks, the money, everything in psychology was own by the behavioristic school. . . . those of us who wanted to be scientific psychologists couldn't really oppose it. You just wouldn't get a job.
>
> (Baars, 1986)

Critics of Behaviorism

To be sure, there were critics of the behavioristic rejection of conscious experience. A.A. Roback wrote that

> . . . no more than a person who has lost consciousness for all time, can be said to remain a person, can psychology, without consciousness, exist as such.

And about 1960, Sir Cyril Burt wrote

> Nearly half a century has passed since Watson proclaimed his (behavioristic) manifesto. Today, . . . the vast majority of psychologists, both in this country (Great Britain) and America, still follow his lead. The result . . . is that psychology, having first bargained away its soul and then gone out of its mind, seems now, as it faces an untimely end, to have lost all consciousness.

When Watson was redefining the past in the 1920s, William James was still a well-known presence in American and British life. In literature, the stream-of-consciousness novel flourished. James Joyce and Virginia Woolf borrowed from James' *Principles*. His psychology textbook was still widely used. William's brother Henry was famous as a pioneer of the psychological novel. In Europe, phenomenologists like Huesserl were inspired by the *Principles of Psychology*. Yet in the sciences, human subjectivity was relentlessly chased out. The legacy of the golden age was gradually erased.

B.F. Skinner

B.F. Skinner rose to fame as the foremost radical behaviorist after John B. Watson left the academic world in the mid-twenties. B.F. Skinner's 50-year career marked the height of radical behaviorism. As he wrote in 1955, he was convinced that "'mind' and 'ideas' are nonexistent entities, invented for the sole purpose of providing spurious explanations." Skinner was not a physical stimulus–response reductionist. He defined stimuli and responses functionally, in terms of their effects on operant (rewarded) behavior. But Skinner always ruled out any attempt to make inferences from behavior, while current cognitive neuroscience is constantly using explanatory inferences that go beyond directly observable stimuli and responses. Consciousness is one inference that is easily made from self-reports.

Many Moderates Continued to Think About Consciousness

Radical behaviorists were a powerful minority. Moderates rarely disavowed consciousness completely. Many just thought that the topic was too difficult for the infant discipline to tackle. Thus Clark Hull and Edwin G. Boring took quite open-minded positions.

In 1937 Hull wrote that

> . . . to recognize the existence of a phenomenon (such as consciousness) is not the same as insisting upon its basic, i.e., logical, priority (as James did). Instead of furnishing a means for the solution of problems, consciousness appears to be itself a problem needing solution."

And Boring wrote

> . . . human consciousness is an inferred construct, a capacity as inferential as any of the other psychological realities, . . . literally immediate observation, the introspection that cannot lie, does not exist. All observation is a process that takes time and is subject to error in the course of its occurrence.

That is indeed how consciousness is treated in contemporary science: as an inferred entity based on reliable evidence, and as a basic question to try to resolve. Making inferences about mental processes is not unusual – it is what humans always do. No one can publicly observe someone else's wish, feeling of love or hate, or pain in the belly. These experiences are best treated as inferences, based on reliable public observations.

The Cognitive Revolution

Dissatisfaction with the narrow limitations imposed by strict behaviorism led to the 'cognitive revolution' in psychology and brain science. Scientists began to understand the mind–brain in terms of information processing of a great variety of representations. However, the cognitive shift was primarily driven by a number of discoveries and rediscoveries – of short-term memory, momentary sensory memories, the elements of speech and language, the varieties of long-term memory, the notion of stages of information processing, of the possibility of parallel processing in the massive society of brain networks, and of neural net models. A wealth of empirical phenomena provided the basis for a new perspective.

Current cognitive theories speak of information processing, representation, adaptation, transformation, storage, retrieval, activation, and the like, without assuming that these are necessarily conscious events. That is to say, modern theoretical languages are neutral with respect to conscious

experience. For example, it is common to speak of information processing and the concept of representation. A representation is a theoretical object that bears an abstract resemblance to something outside of itself. Human knowledge can be naturally viewed as a way of representing the world and ourselves. We can think of percepts, images, plans, intentions, and memories as representations. Everyday psychology can be translated into these terms in a natural way. Some scientists prefer the notion of 'adaptation' to 'representation'; both are useful theoretical constructs.

All this may seem obvious in an age of digital computers, but it is actually a painfully achieved historic insight. William James, as noted above, felt that all psychological events must be reducible to conscious experiences, while the behaviorists denied the relevance of either consciousness or unconsciousness. Either position makes it impossible to compare similar conscious and unconscious events, and to ask the question, ''Precisely what is the difference between them?'' Because it is neutral with respect to conscious experience, the language of information processing gives us the freedom to talk about inferred mental processes as either conscious or unconscious. This is a significant step toward clarity.

The ''Cognitive Revolution'' Did Not Address Consciousness as Such

Cognitive science is not merely a fallback to the nineteenth century. For one thing, it generally did not address the foundation issue of consciousness *as such*. Because it is not possible to study conscious human subjects without somehow addressing the reality of their subjective experiences, the issue of consciousness was handled by vague implications. Thus all subjects in standard experiments are asked to pay attention to a set of stimuli and to follow task instructions. But following those instruction always results in specific subjective experiences, as human subjects are happy to tell us. Such obvious facts were rarely acknowledged explicitly, because scientists still felt the topic of human subjectivity was tainted by behavioristic criticisms. For example, Ulric Neisser, a founder of cognitive psychology, continues to assert that ''I do not think psychology is ready for consciousness.''

Until recently scientists were so skeptical about studying personal experiences that many important questions were not even asked. As a result we have no idea even today whether people can give reliable descriptions about their feelings of food cravings, for example, something that might be important for obesity research. We know almost nothing about the experiences of sexual pleasure, about awareness of aggressive impulses, the subjective dimensions of drug effects, the struggle to control unwanted habits, and many more topics that could be explored with an open mind.

The Return of Consciousness (and Unconsciousness) to Science

A number of research programs returned to consciousness, often using euphemisms to avoid the professional taint of the term. Thus the topic of habit formation after repeated practice became very productive again, comparing 'strategic' versus 'automatic' mental processes. A small community of researchers refined the very tricky methodology of subliminal perception. Others focused on explicit problem solving, compared to implicit or 'intuitive' processes.

Even today many scientists still prefer to use a score of euphemisms for conscious experiences. This makes it difficult for students to understand the extraordinary range of conscious events we are studying, variously called 'perception,' 'episodic recall,' 'explicit problem solving,' 'awareness,' and even just 'cognition.' In practice all these terms are assessed in the same way – by way of voluntary report, often checked for accuracy. Thus empirically, all the euphemisms stand for the same construct. However, this Tower of Babel of euphemisms is ultimately a formula for confusion, not clarity of thought. In the long term it cannot be a good thing for healthy science.

Brain Science Often Confirms Subjective Reports

With the advent of new methods for observing the living brain with both spatial and temporal accuracy, scientists quickly found remarkable convergence between psychological functions and precisely localized brain events. The new field of

cognitive neuroscience provides scores of examples. Researchers also found that conscious processes (such as supraliminal percepts) had quite different brain effects compared to closely matched subliminal ones. Automatic and effortful tasks looked strikingly different in brain scans, even if the tasks are behaviorally identical. Even single neurons appeared to respond differently to conscious (compared to unconscious) stimulation. New insights into state differences (between unconscious sleep and waking, for example) opened up a host of novel questions and some possible answers. In short, the issue of conscious experience, rather than leading to controversy and confusion, appeared to lead to reliable findings and novel, productive questions. In that sense, the study of consciousness has simply turned into normal science.

Summary and Conclusions

While most psychologists and neuroscientists believe that the behavioristic rejection of consciousness is past, a former president of the American Psychological Association disagrees.

> Behaviorism is alive and well and nothing 'has happened' to it. . . . behaviorism is less discussed today because it actually won the intellectual battle. In a very real sense, all psychologists today (at least those doing empirical research) are behaviorists. . . . Behaviorism won. . . . Behaviorism is alive and most of us are behaviorists.
>
> (Roediger, 2004: 42)

However, that appears to be a fading belief among researchers. Francis Crick, the codiscoverer of DNA, devoted the last half of his career to reinvigorating the scientific study of consciousness. Just before his death he expressed satisfaction that

> A few years ago one could not use the word 'consciousness' in a paper, for, say, Nature or Science. But thankfully, times are changing, and the subject is now ripe for intensive exploration. . . . Consciousness is the major unsolved problem in biology. . . . Solving the problem of consciousness will need the labors of many scientists, of many kinds . . .
>
> (Francis Crick, 2004)

Many college students are still taught to be suspicious of conscious experiences, including their own. As wave upon wave of young apprentices adopt a misleading myth of origins, academic psychology continues to define itself by that imaginary historical moment, shortly after 1900, when superstition and wrong methods were cast into the outer darkness. Thus behaviorism continues to influence many scientists.

These continued misunderstandings are important. For example, some medical anesthesiologists claim that consciousness is not the crucial issue in surgical anesthesia, as long as patients cannot remember anything afterwards. That point raises ethical questions about patients who experience intense conscious pain, but who may not remember the fact afterwards due to memory-impairing drugs like benzodiazepines, which are routinely administered in surgery. Questions still rage whether newborn babies and late-term fetuses suffer pain. The developmental psychologist Jerome Kagan has claimed that newborns do not have sensory consciousness, and hence cannot feel pain (1998). Yet a study in the *Journal of the American Medical Association* shows that newborns utter distress cries during postnatal circumcision, unless they are given a local anesthetic. Consciousness thus gives rise to major ethical questions, which can be clarified using scientific methods.

The reality of private conscious experiences in human (and animal) minds still poses some of the most profound scientific and ethical questions in the world. On the scientific side we now see steady progress, which may help to address ethical questions as well. That is what keeps many scientists engaged in this fundamental enterprise.

See also: History of Philosophical Theories of Consciousness.

Suggested Readings

Baars BJ (1986) The Cognitive Revolution in Psychology. New York: Guilford Press.
Baars BJ (2002) The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Science 6(1): 47–52.
Edelman GM and Tononi G (2000) A Universe of Consciousness. New York: Basic Books.
Hilgard ER (1987) Psychology in America. New York: Harcourt Brace Jovanovich.
James W (1890/1983) The Principles of Psychology. Cambridge, MA: Harvard University Press.

Relevant Websites

http://plato.stanford.edu/entries/ancient-soul/#3.1 – Stanford Encyclopedia of Philosophy International Association.

http://plato.stanford.edu/entries/aristotle-psychology/ – Stanford Encyclopedia of Philosophy International Association.

Biographical Sketch

Bernard J. Baars is former senior research fellow in theoretical neurobiology at the Neurosciences Institute in San Diego (www.nsi.edu) His PhD is in cognitive psychology from UCLA. He is interested in human language, the brain basis of consciousness, volition, and a variety of related topics including the history of scientific studies of consciousness, and neuroethics. Baars pioneered a cognitive theory of consciousness called Global Workspace Theory, which is widely cited in scientific and philosophical sources. Together with William P. Banks, Baars has edited the journal Consciousness and Cognition for more than fifteen years (from Academic Press/Elsevier). With Nicole M. Gage, Baars has written an introductory text for cognitive neuroscience, called Cognition, Brain and Consciousness: An Introduction to Cognitive Neuroscience (Baars & Gage, Eds. San Diego, CA: Elsevier/Academic Press, 2007). Baars was founding president of the Association for the Scientific Study of Consciousness and has an ongoing research collaboration for large-scale cognitive modeling with professor Stan Franklin (University of Memphis, Institute for Intelligent Systems).

History of Philosophical Theories of Consciousness

W Seager, University of Toronto Scarborough, Toronto, ON, Canada

Introduction to the Problem

It is impossible to approach the problem of consciousness without having some basic answer to the question: what is consciousness? But this question is not altogether easy to answer. If someone really did not have any idea what consciousness is, it is hard to imagine how to convey anything informative to them (as Louis Armstrong once said when asked what jazz was, if you got to ask you ain't never gonna know). The term 'consciousness' covers a large range of mental phenomena. Sometimes it refers to the complex reflective appreciation of our own mental states as we experience them. This is a very sophisticated mental act. On the other end of the scale, consciousness is sometimes equated with mere wakefulness or sensory awareness (this is the definition employed by anesthetists). Presumably animals (at least some of them) enjoy this kind of consciousness while it is likely that only human beings can reflect on their own mental states. The most vivid examples of conscious states are indeed sensory states: the brilliant red of a ripe tomato, the piercing sound of the trumpet, the taste of a cold beer on a hot day, and so on through all the sense modalities. But the catalog of consciousness should also include conscious thought, and this involves both the contents of our thoughts and more nebulous aspects of cognition, such as the sense we have when we suddenly know that we understand something, a characteristic state of consciousness which transcends the content of what we have grasped. As to the content of our conscious thoughts, it remains a vexed philosophical question whether all thoughts come to consciousness clothed in some sensory garb (as it might be, mental images, either aural or visual) or whether there are some 'pure' intellectual apprehensions. The catalog further needs to include basic affective states of consciousness, the paradigms of which are pain and pleasure. The motivational power of such states makes them distinctive, and their fundamental importance to survival perhaps even suggests that they are in some way the most primitive of all the conscious states. Affect goes beyond the simple push–pull of pain and pleasure. It informs a vast array of emotional states of consciousness. Emotions such as anger, love, or fear make up some of our most intense conscious experiences and while these states involve, they go beyond the bare affective, sensory, or intellectual forms.

One might despair of finding any core or common feature in such a diverse list of mental states. The somewhat elusive essence of consciousness is subjectivity, the way that there is, to use the famous phrase of Thomas Nagel, something it is like to be a subject of a conscious state. The rock or the ashtray has no subjective aspect, no 'interior' life. Leaving aside the details of all the different kinds of mental states which populate consciousness, the core problem of consciousness lies in understanding the nature and origin of subjectivity itself.

It must also be appreciated that the problem of consciousness appears in the history of Western philosophy as part of a more general issue, which is usually labeled the mind–body problem. This is so despite the fact that consciousness is the distinctive feature of mentality that sets it apart from the rest of nature, so one would not expect there to be any problem about mind at all apart from the problem of consciousness. The larger problem arises because consciousness is an attribute – it is a characteristic of something and this 'thing' came to be called the mind. This issue is deeply tied up with the truly ancient question of whether there is anything which might survive the all too evident destruction of the body, which every human being eventually recognizes will bring an inevitable end to their earthly existence. From the earliest times there has been a strong tendency to believe (or hope) that we somehow survive death and, in order for this survivor to be us, it had to retain consciousness and of course it had to be a radically different kind of thing than the body.

Although the modern form of the problem of consciousness is a relatively recent development, which is, as we shall see, intimately connected to the rise of empirical science and the subsequent scientific view of the world, it has long been recognized that it is consciousness which makes us different from (the rest of) material nature. Thus to understand the problem of consciousness and the theories philosophers have proposed to solve it we have to look back at the long history of the mind–body problem.

Early History

The views of the pre-Socratic philosophers of the ancient Mediterranean basin are often mocked or derided as pathetically ludicrous: everything is made of water, nothing can move. But we owe them an immense debt for coming up with the idea that the world can and should be explained in terms of natural forces and investigated by reason. It is hardly surprising that the positive views of these knowledge pioneers were wildly wrong, more important was the outlook they embodied, which laid the early foundation of the scientific revolution.

The pre-Socratics's quasinaturalist outlook led them to ask a fundamental question about the structure and organization of the world, whose answer remains to this day in dispute. This is the opposition between emergence and what we might call (for lack of a better term) inherence. The world contains a vast number of different things exhibiting a staggering complexity of features and interactions. But it is evident that some of this complexity can be explained in terms of the interactions of the participating entities (this is most evident in the artifacts we construct and it is sometimes forgotten how highly skilled ancient people were at devising tools and simple machines). This observation suggests that perhaps everything in the world can be explicated in terms of a small number of basic features possessed of a few simple properties. Such a line of thought naturally invites the opposing challenge that there are some things which cannot be thus 'reductively' explained.

Someone who thinks that there are a few simple elementary features of the world whose basic properties and interactions give rise to everything else endorses a form of emergentism. Anytime an object possesses a property which its parts do not we have emergence. For example, the solar system exhibits a complex dynamical structure that none of its constituents do (or even could) possess. But the organization of the solar system is entirely the product of the law governed interactions of its constituents, following from their properties (e.g., mass, position, and velocity). Dynamical structure is thus an emergent feature. The most famous emergentist of the pre-Socratics was the famed atomist Democritus (c. 460–370 BCE) who held that everything could be 'reduced' to the simple properties and motions of the atoms.

But while it is easy to see that some features are plausibly regarded as emergent, it does not follow that everything is. Consider for example color. Supposing that the atoms are not colored, how does color emerge from the atomic interactions? Opposed to Democritean emergent atomism, Anaxagoras (c. 500–425 BCE) flatly denied that emergence was possible and opted instead for a view in which all qualities inhered in everything, the diversity of the world explained by the different mixing proportions peculiar to each particular thing. Modern atomists – that is, all of us – are not impressed by the example of color. Color is simply the effect on our visual systems of fundamentally noncolored constituents of matter. Perhaps Democritus was foreshadowing this response when he obscurely pronounced "by convention is color. . . in reality the void and atoms."

Within the mind–body problem this basic dichotomy generates two theories of the nature of consciousness: emergence and panpsychism. Emergentists believe that mind and consciousness arise from nonconscious parts derived from nonconscious precursors, whereas panpsychists hold that mind is a fundamental feature of the world and that everything possesses in some way and in some measure a form of consciousness. There are of course many ways to articulate each of these views. A panpsychist outlook might emphasize the idea that consciousness could not emerge from nonconscious components but hold back from the claim that absolutely everything has a mental aspect.

In one way or another, this dichotomy has formed the backdrop of the mind–body problem throughout its history. Let us consider the two

giants of ancient philosophy, Plato and Aristotle, in this light. Plato utterly lacked the reductionist spirit of the pre-Socratics and in fact explicitly attacked it in the Phaedo. Plato instead proposed that the mind was distinct from the body, which it somehow 'animated.' Naturally, Plato has some arguments for the view that the mind is separate from the body. One such argument is very abstract and stems from his famous doctrine of Forms. The Forms are perfect entities that material reality merely approximates. For example, in geometry we prove things about circles, but nowhere in the material world are there any such beings. All circular objects – and most certainly the sometimes very rough diagrams deployed by geometers – fail to be perfectly circular and hence are not, strictly speaking, circles at all. So how is it possible, Plato asks, for our minds to get into intellectual contact with these radically nonphysical entities? Our conceptual acquaintance with them suggests that the mind is similarly nonphysical. Plato also develops a less abstract version of this argument in his doctrine of recollection. Since we know the Forms but never encounter any instances of them in perception, our knowledge must be based on some other kind of awareness. Plato believed that this could only be explained in terms of a nonbodily phase of our existence, which we dimly recall in our bodily existence (Plato in fact went so far as to endorse a doctrine of reincarnation). While his arguments are not particularly convincing, they do point to a very curious feature of thought: its ability to be about objects independent of their existence (a property which later came to be called the intentionality of thought).

At first glance, Aristotle stands as the naturalistic antithesis of Plato. He denied the doctrine of Forms, substituting the idea that objects are to be understood in terms of his own distinction between form and matter, where by form Aristotle means something like organization or structure, while matter is whatever 'unstructured stuff' constitutes the object in question (thus the matter of water is hydrogen and oxygen, the matter of an ant colony is the ants). Aristotle defines the soul (or mind) as "the form of a natural body having life potentially within it." This suggests that the mind is a feature of us as material beings rather than some ethereal, other worldly stuff. However, Aristotle was not a thoroughgoing naturalist; like Plato he thought there was something very special about the mind, something so special that it could transcend physical matter. He sometimes likened the relation of mind to body to that of pilot to ship and explicitly held that at least one aspect of the mind – that part capable of rational thought, which he called the active intellect – could exist apart from the body. Aristotle deployed his conception of form in an argument for dualism. The mind can think about anything, and since it is the form of the object thought about that 'informs' the mind (i.e., when you think about a lion the form of lion informs the 'matter' of your mind) this means that the mind can take on any form. But no material organ can take on every form (Aristotle's examples are the eye which can take on only the forms of the visual, or the ear which can take on only the auditory.) and thus the mind cannot be material. It is interesting that Aristotle's argument, like Plato's recollection argument, focuses on the mysterious ability of the mind to get into contact with any potential being (or even nonbeings when we recall that it is no problem to think about the impossible).

Aristotle also introduced the still influential idea that all conscious mental states are in some measure about themselves or are self-representing. The argument for this was that if mental states are brought to awareness via a distinct mental state (which is about the first state) then there would have to be a third state to bring the second state to awareness, so generating a vicious regress to infinity. Notice that Aristotle is implicitly assuming that all mental states are such that we are consciously aware of them; otherwise the regress can stop with the first mental state, which is not a conscious state. This is not to fault Aristotle. It took a very long time indeed before the idea of unconscious mental states could even be regarded as coherent, let alone regarded as a serious hypothesis.

The Scientific Revolution

Aristotle's philosophy, with some considerable bending and squeezing, served the ecclesiastically constrained philosophy that dominated the middle ages. There was of course much debate about the nature and structure of the mind and its contents.

(It is to the medieval scholastics that we owe the crucial notion of intentionality or the directedness of mental states onto objects, which they are "about.") But the idea that there was some kind of problem about the mind in relation to the natural world was very far from center stage, and the notion that mind needed to be explained in naturalistic terms just did not arise. This complacent view of things changed radically when the scientific revolution burst upon Europe. Galileo (1564–1642) set the stage with his distinction between primary and secondary qualities. The former are the mathematically tractable properties of matter, such as shape, mass, and motion, while the latter are the manifestations in consciousness of arrangements of primary qualities. The secondary qualities are solely properties of the conscious mind. In *The Assayer* (published in 1623) Galileo wrote that

> I think that tastes, odors, colors, and so on are no more than mere names so far as the object in which we locate them are concerned, and that they reside in consciousness. Hence if the living creature were removed, all these qualities would be wiped away and annihilated.

This distinction made the world safe for science, insofar as it could then focus on the objective features of the material world as explicated in purely mathematical theories. But the question of the nature of the secondary qualities themselves was only temporarily avoided. The burgeoning scientific picture of the world, which then and always aspires to completeness, could not long ignore the elephant in the room.

The most famous response – that of René Descartes (1596–1650) – forthrightly imposed a total separation between the mind and the material universe in a theory that quickly came to be called Cartesian dualism. If one goes along with the assumption that mind is separate from body then Descartes's view is quite in line with commonsense. It allows that mind and body interact causally, so that a kick in the shin generates (after considerable physical machinations about which Descartes had very interesting, even prescient, things to say) a consciousness of pain and, conversely, a feeling of anger could cause one's arm to execute the retaliatory punch. Cartesian dualism also opens the door for standard Catholic doctrines of immortality of the soul and bodily resurrection – an important consideration in a culture that sometimes burned heretics and infamously forced Galileo to recant his defense of Copernicanism and consigned him to a life sentence of house arrest.

But Cartesian dualism faces a number of exquisitely difficult philosophical problems. The causal interaction between mind and body threatens the completeness of the scientific picture of the world. Every so often the mind intervenes and changes the physical world in a way that would be physically inexplicable and would violate rather basic laws, such as the conservation of energy. Second, and more metaphysically, how is it even possible for two realms that differ so fundamentally as the mind (nonspatial, without location) and the body to causally interact? This was noted as soon as Descartes put forth his theory. One of his royal correspondents, the princess Elizabeth in 1643, asked him outright: "how can the soul of man determine the spirits of the body, so as to produce voluntary action?" Descartes's official answer was that the mind–body union was a brute fact, instituted by god, which was humanly incomprehensible.

In response to such difficulties, a host of alternative dualist theories were devised. Spinoza (1632–1677) offered a dual-aspect account in which mind and matter were two of an infinity of attributes of a single unified underlying substance which he called god (a blasphemy to many as it implied that god possesses material properties). There is no interaction between diverse attributes but they stand as perfect mirror images of each other, thus presenting an appearance of causal commerce.

Leibniz's (1646–1716) theory avoided the objectionable pantheism of Spinoza by postulating an infinite host of individual minds (called monads) whose perceptions were set up by god (the supreme monad) to be in perfect preestablished harmony with the events of the physical world. Leibniz was also notable as the first to entertain the idea that some mental states, which he called "petite perceptions," were unconscious.

Another, rather curious, account of the mind–body relation is Malebranche's (1638–1715) occasionalism, which posits god as the intermediary between mind and matter so that every intentional action or sensation is literally a miracle.

Such a theory forcefully reveals the perceived intractability of the problem of mental causation.

Such dualist, antiemergentist accounts of the mind were the dominant approach of the time, but there were a few prominent thinkers who were at least branded as materialists (e.g., Hobbes (1588–1679), Gassendi (1592–1655)). However, it seems that the core of their materialist views was the denial of a separate mental substance which exists apart from the body. They did not go so far as to claim that mental attributes were identical to physical properties. For example, Hobbes says that human beings are material objects that can be fully explicated in terms of the motions of the particles which make up the brain and body, but he adds that ''the appearance or sense of that motion is that we either call delight or trouble of mind.'' Such appearances are the realm of consciousness and in these early materialists we have what is sometimes called property dualism (as opposed to the substance dualism of Descartes). They did not venture beyond the already radical idea that human beings and souls were entirely material objects.

The Rise of Scientific Philosophy

Nor, of course, did materialism form the core of any important philosophical developments at the time. Instead, philosophy pursued the antimaterialist and antiemergentist theory of idealism, which asserts that the totality of existence is fundamentally mental in nature. Various forms of idealism were promulgated, starting with Berkeley (1685–1753) and Kant (1724–1804), and idealism remained in a dominant position until the beginning of the twentieth century. Despite its immense significance for the history of philosophy, idealism is not of prime concern for this article since its acceptance of consciousness as the ontological ground for the physical world meant that idealism faced no problem of consciousness in the relevant sense. It is however important to note an offshoot of the idealist developments known as phenomenology (stemming from the work of Franz Brentano (1838–1927) and Edmund Husserl (1859–1938)), which sought to provide the exact structure and interior constitution of conscious experience as such via introspective examination.

Of more significance in the development of the problem of consciousness was the spectacular growth in the range and explanatory capability of the sciences from the time of Descartes to the end of the nineteenth century. In addition to the general extensiveness of science, two developments were of particular importance: the birth of evolutionary biology with the publication of Darwin's (1809–1882) *Origin of Species* in 1859 and the transformation of psychology from a branch of philosophy to an autonomous scientific discipline. The theory of evolution promised to explain how complexity could emerge from simpler forms, explicitly within the organic realm, but implicitly much further – up to and including the origin of life itself from nonorganic precursors. This forced into prominence the problem of emergence. There were those who denied the possibility that consciousness could be formed out of the nonconscious. Discussing the discontinuous leap from the nonconscious to the conscious, William Clifford (1845–1879) wrote ''we cannot suppose that so enormous a jump from one creature to another should have occurred at any point in the process of evolution.'' This 'genetic' argument against emergence retains advocates to this day. The basic problem stems from the observation that the normal mode of emergence, which evolutionary theory enshrines, is that of novel capabilities springing from greater organizational complexity (an example would be the difference between a hand calculator and a supercomputer). The argument proceeds to claim that the normal mode cannot explain the appearance of unique properties such as consciousness. While this argument has some intuitive force, the alternative is to accept that mentality is a fundamental feature of the universe, built in right from the beginning on the ground floor of creation. Such a panpsychist view is very far from intuitively acceptable – our experience strongly suggests instead that at some point consciousness emerges along with the growing complexity of nervous systems in animals.

Many thinkers in the later nineteenth century embraced the panpsychist option in one form or another (a very notable example is William James (1842–1910)). But the emergentists, including J. S. Mill (1806–1873) and C. Lloyd Morgan (1852–1936) of 'Morgan's Canon' fame, also developed a vigorous theory in which nature was seen as

hierarchically ordered in layers of emergent properties stemming from basic and irreducible laws of nature that governed the generation of the emergents as underlying complexity grew. This form of emergence breaks radically from the normal mode discussed above and is often called radical emergence to distinguish it from the more commonplace form. According to radical emergence, it would be impossible, even in principle, to deduce the emergent features from the submergent structure unless one also took into account the irreducible and unpredictable 'laws of emergence.' The emergentists regarded ordinary chemistry as a clear and uncontroversial example of radical emergence in the physical domain and extended the idea to include life and consciousness. Of course, this leaves the appearance of consciousness as a rock-bottom mystery even as it accepts the commonsense view that dogs but not amoebas are conscious. As Thomas Huxley (1825–1895) put the point "how it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp."

The doctrine of radical emergence was implicitly tied to classical physics. Thus the development of quantum mechanics appeared to undercut completely the linchpin example of chemistry when in 1925 the basis of chemical composition was in principle explained (and there has been no sign since then that the chemical properties of substances are not fully dependent on the physical properties and arrangement of their constituents without the magic of radical emergence). Developments in biology, culminating in the discovery of the chemical basis of heredity, then destroyed the plausibility of life being a radically emergent phenomenon, which left consciousness as the sole candidate for emergence with no other examples to lend it support.

The rise of psychology as a scientific discipline also worked against the idea of radical emergence. After initially embracing consciousness as a proper object of study, in the early twentieth century psychology turned away from the interior aspects of the mind and entered a bizarre and lengthy infatuation with behaviorism – the core (or sole) idea of which being that the only proper subject of study in psychology was the observable and objective bodily movements of subjects.

From a behaviorist perspective the problem of consciousness is real, but easily solved: either consciousness does not exist or it is definable in terms of the movements of certain physical bodies (or in the dispositions of physical bodies to move in certain ways under certain conditions). In philosophy, behaviorism took support from certain metaphysical and methodological doctrines of the Logical Positivists, in particular their verification theory of meaning. Verificationism entails that any statement which is not publicly verifiable by objective observation or measurement is in fact meaningless or nonsensical (one of Wittgenstein's (1889–1951); examples of such nonsense was the question of "when is it 5 o'clock on the Sun"). Verificationists were predictably annoyed when it was pointed out that there was no publicly observable method to verify their own theory of meaning.

Behaviorism had a huge impact on the scientific study of the mind, essentially blacklisting reference to internal mental states except insofar as they could be explicitly defined in terms of observable behavior. The study of consciousness as such was seen as unscientific at best, entirely pointless or akin to an attempt to study unicorns or interview Sherlock Holmes at worst. Behaviorism's impact on philosophy was also very great. The behaviorist 'zeitgeist,' if not its official doctrines, strongly influenced philosophers such as Wittgenstein and Gilbert Ryle (1900–1976), who both developed accounts of the mind more or less based on the idea that mental states words (like 'clever,' 'pain,' or 'imagine') referred to patterns of behavior rather than internal states of the subject. But perhaps a more important effect of behaviorism was the impetus it gave to the then nascent philosophical project of naturalization.

To naturalize something is to show how it fits smoothly into the scientific picture of the world without placing any strain on that world view. For example, in the nineteenth century, vitalism was a respectable scientific theory of life and vitalists like Hans Driesch (1867–1941) performed ground breaking work in fields such as embryology. Yet vitalism's assertion that life depended upon a mysterious and nonphysical 'elan vital' could not be reconciled with the materialist outlook favored by physically minded thinkers. One of the triumphs of twentieth century biology, that is, the discovery of

the chemical basis of the genetic code along with a host of discoveries about the chemical makeup of organic compounds and the physical nature of organic processes, served to naturalize life. Nowadays life is seen as a kind of chemical process (of great complexity to be sure) that does not violate natural law nor introduce any immaterial entities into nature.

Behaviorism offered a similar route toward the naturalization of the mind. If mental states really were just complex patterns of behavior of living creatures, and if the operations of living creatures was itself 'reducible' to the chemical processes going on within them along with the purely physical interactions between them and their environment, then the mind would pose no threat to the scientific picture of the world.

While behaviorism did not survive, the attendant urge to naturalize has been the core feature of philosophical theories of consciousness since the mid-twentieth century. One of the very powerful forces behind the naturalization program is the widespread acceptance of scientific realism – the doctrine which holds that hypothetical entities postulated by science, which are typically invisible to the human senses, are real. There was genuine resistance to this idea. Early work on chemical structure in the nineteenth century was regarded as simply a useful way of organizing information about measurable chemical proclivities. In general, the atomic theory was similarly regarded as a useful but unrealistic model, opposed by such eminent scientists as Wilhelm Ostwald (1853–1932) and Ernst Mach (1838–1916) into the beginning of the twentieth century. The work of Albert Einstein (1879–1955) and Jean Perrin (1870–1942) on Brownian motion is the traditionally cited coup de grace of antiatomism. In any event, thereafter the right and capacity of science to pronounce on the invisible, deep structure of the world became widely recognized. Although the existence of atoms was consistent with radical emergentism, the apparent in principle reduction of chemistry to quantum mechanics suggested instead that the world operated entirely on the basis of the interactions of a relatively few exotic quantum entities as governed solely by laws at the quantum level (along with, in a very uneasy partnership, the principles of general relativity for very massive objects).

The evident ability of science to pronounce on the features of the world that lay 'behind' the phenomena and were not directly observable made the strictures of behaviorism seem unnecessarily encumbering or even silly, and the doctrine withered to be replaced with a cognitive psychology happy to emulate the other sciences and postulate inner processes, states, and events which only indirectly and in concert generated behavior. In addition, the importance of internal states is revealed by the automaton argument. If behavior is all that matters for the ascription of mind, then a preprogrammed robotic creature which acted as if it possessed a mind would in fact have a mind, which seems to violate a strong intuition to the effect that the cause of the behavior matters. (There is a difference between a zombie and genuine human being.)

Recent Approaches to Consciousness

In response to the growing acceptability of scientific realism and the decline of behaviorism, philosophers (notably J. J. C. Smart and U. T. Place) responded in the 1950s with a radical form of materialism which asserted not only that human beings were entirely material objects but that mental properties were identical to physical properties, in particular, identical to properties (or states) of the brain. This central state identity theory asserted that states of consciousness, such as feelings of pain or experiences of color, were nothing more than states or processes going on in the brain. This was not the view that brain states correlate with or cause mental states, but the radical view that the property of pain just is a neurological property. Such a view obviously fits very comfortably with the scientific picture of the world (it was of course designed to fit), avoids the behaviorist's absurd denial that there are inner mental states, and clearly avoids all the problems of interactionist dualism. Note also that the theory is a form of emergentism, but not of the radical sort discussed above but rather a benign kind following the normal mode of emergence in which the emergents are explicable novelties. The scientific picture of the world is full of this kind of emergence (tornadoes,

thermodynamical properties, etc.) and the identity theorists could add this to the scientific virtues of their theory.

Nonetheless, serious problems with the identity theory were soon noted. The core notion of identifying mental and physical properties seemed metaphysically somewhat dubious, but proponents could point to a fair number of property identifications in science, the most frequently mentioned being the identity of a gas's temperature with the mean kinetic energy of its constituent molecules. Pondering the implications of such empirically discovered identities led to a number of important philosophical advances. But apart from metaphysical scruples, more straightforward objections were also advanced. One is the problem of 'neural chauvinism' (a term coined by Ned Block, one of the early proponents of this objection). Suppose that we met an alien race who by all appearances were intelligent and conscious beings. We then discover that their inner workings are radically different from our own – they do not have brains as we know them. Would this preclude them from possession of mentality? It seems intuitively reasonable that many different internal systems could generate consciousness, but that runs directly counter to the identity theory. Another weakness of the theory is that since it identifies mental and physical states it has profound difficulty explaining the nature of subjectivity. All physical states are entirely objective and in principle entirely knowable. Yet the subjective quality of conscious states does not seem to be revealed via knowledge of the physical structure and operation of the brain (this line of argument, which is powerfully general in scope, was initially raised by Thomas Nagel and extended by Frank Jackson). It thus seems quite mysterious why certain physical properties are subjective (we cannot simply say that they have subjective qualities or features since the identity theory's core claim is that all such features are strictly identical to physical properties).

A variant of the identity theory, labeled functionalism, diagnosed the core problem with the identity theory as a misidentification of the proper level of analysis. According to functionalism, mental states are not to be found in the neural hardware directly but are instead to be identified with the functional architecture of the brain. A frequent analogy is to the operation of the digital computer, leading to the functionalist slogan: as software is to hardware so mind is to brain. Just as the same program can be run on a variety of different kinds of computers (and, in principle, a computer can be made out of just about anything) so too mental states can be implemented or realized in any number of different physical substrates. All that is required is that the system of internal states interact in the appropriate organizational way, leading to both internal changes and external behavior consistent with the possession of a mind. Functionalism has an obvious appeal: it trades on a current technological analogy (always popular in explaining the mysteries of the mind), does not denigrate brain research while also encouraging an abstract field of mental or cognitive architecture, and it allows for a pretty clearly acceptable route toward the naturalization of the mind. It also avoids the neural chauvinism objection to the identity theory while avoiding, or at least, patching over the metaphysically disturbing property identifications characteristic of the identity theory. And it too, no less than the identity theory, fits in with current scientific theory, in particular with the young and burgeoning interdisciplinary field of cognitive science, which implicitly accepts the functionalist outlook as its core vision. It would be fair to say that in one form or another functionalism has become and remains the most widely accepted account of the nature of the mind and how it fits into the natural world.

Nonetheless, functionalism faces its own set of problems. The virtue of not being tied down to any particular physical substrate threatens to become the vice of a too liberal distribution of the mental. If any system which possesses the appropriate organization can realize conscious mental states then we can envisage some very bizarre realizations. We could, to adapt an example first developed by Ned Block, organize a (very large) group of people to pass text messages back and forth so as to mimic the functional architecture of a mind suffering intense pain. It is hard to believe that such a system would generate any suffering (apart from the probable excruciating boredom of the participants). On the other hand, if these bizarre realizations are ruled out then the question arises, as it does for the identity theory, of just

exactly what it is that accounts for the existence of subjective states of consciousness. Biting the bullet on bizarre realizations does not obviate the question however. As in the identity theory, there is a deep problem about why certain organizational structures of matter should possess (or realize or implement) subjective features of consciousness. In fact, the problem may be worse for functionalism than for the identity theory. How could mere organization generate Technicolor phenomenology?

Functionalism highlights another, rather subtle, worry that harks back to the time of Descartes. If mentality is an organizational feature then it is not identical to its material substrate but is, wherever and however it is realized, dependent on and determined by the substrate. In that case there is a problem about mental causation. Organization does not seem to cause anything by itself but only to 'borrow' causal efficacy from whatever stuff it may be that is organized. For a simple example, consider how hurricanes cause death and destruction. Hurricanes and such are normal mode emergents. It is plausible to regard them as organizational features of the atmosphere, totally dependent on and determined by the underlying atmospheric properties of temperature, pressure, etc. Note also that hurricanes are not to be identified with the underlying properties of the Earthly atmosphere – all sorts of gas configurations can produce hurricanes (they, or similar vortical disturbances, exist on Jupiter, for example, with a radically different atmosphere than Earth's). Yet if we ask where the causal efficacy of a hurricane resides, it is not in the hurricane as such but in the underlying features (e.g., wind velocity). Moving to the case of consciousness, the worry is that, for example, pains do not cause anything in virtue of being painful, but rather and only in virtue of the powers of the underlying substrate. This is deeply counterintuitive. It is hard to deny that painfulness in itself has causal efficacy. This problem seems even more acute if we ask how more abstract mental states, such as awareness of meaning, can have any genuine causal powers.

This 'causal drainage' from high-level to low-level properties is a general feature of naturalization efforts and it seems pretty harmless in domains distant from consciousness. We understand quite well how a hurricane wreaks havoc via the powers of its constituents. It is less easy to admit that our states of consciousness only do their work in virtue of the entirely nonconscious constituents of their realizing states. The case of conscious thought, as opposed to sensory consciousness, seems even worse. The currently most favored philosophical theories of how mental states acquire meaning essentially involves relations to entities external to the mind; hence, these views are labeled externalism. There is a famous philosophical thought experiment introduced by Hilary Putnam which postulates a distant planet superficially indistinguishable from Earth but differing in one particular way: our water (H_2O) is replaced on Twinearth with another substance (XYZ) which is similarly liquid, potable, etc. According to theories of meaning in which external relations are an essential part of the mechanism which assigns content to meaningful states, it is our relation to H_2O which determines the meaning of 'water' while on Twinearth it is the relation to XYZ that determines the meaning of 'water' as used on Twinearth. When my twin on Twinearth thinks "I would like a glass of water," his thought is different in content from mine, no matter how physically similar we might be (neglecting as philosophically irrelevant the inconvenient fact – a mere artifact of the chosen example – that we are mostly made of water while our twins our made of XYZ). There are additional seductive arguments for externalism, pioneered by Tyler Burge, which emphasize the social dimension of meaning which maintain that the contents of our thoughts are beholden to our linguistic milieu. For example, we can think about legal contracts even though we do not know everything that the concept entails. Somebody might mistakenly think, for example, that contracts have to be written documents. They are wrong but their thought is about 'our' legal contracts nonetheless. On Twinearth, as it might be, contracts do have to be written, so, our subject's twin is thinking a true thought about 'their' legal contracts.

In general, the externalist lesson about the content of thought is that two physically identical people could have thoughts with distinct contents if they are in distinct environments (and I emphasize this is supposed to be so even if the environmental differences have no effect on the physical state of the individuals involved). A crude analogy is the way that two identical pieces of paper could differ in

their status as 'genuine US dollars.' Arguably, causal powers reside in the physical structures which implement mental states, and these are not sensitive to such environmental differences. Thus, content-carrying mental states are not efficacious in virtue of their content (no more that the property of 'being a genuine US dollar' has intrinsic causal efficacy), which is, again, deeply counterintuitive when we think of the apparent causal power which even introspective access to the contents of our thoughts appears to require.

The problem of mental causation also appears in the influential theory of Donald Davidson (1917–2003) called anomalous monism. This theory – in some ways an updated form of Spinozism – asserts that the mental and physical realms are not and cannot be linked by any scientific psychophysical laws even though events can be given both mental and physical descriptions (reminiscent of Spinoza's dual aspect view). Davidson allows that mental and physical events causally interact however but not, it would seem, in virtue of the mental properties of events but rather because of their physical properties. While Davidson regards talk of 'causing in virtue of' with great suspicion, it is a perfectly commonplace feature of our understanding of causation. For example, suppose that a brick falls on your toe and you feel pain. It makes perfect sense to say that this was in virtue of the brick's mass but not its color. So the worry is that in anomalous monism the mental properties of events are causally inert (at least within the physical world).

The problems of consciousness, content, and naturalization have recently been linked in some exciting theories which take the representational power of the mind as constitutive of consciousness, thus reducing the problem of naturalizing consciousness to the supposedly easier task of naturalizing representation. Two of these approaches deserve a brief mention. One, in some ways harking back to Aristotle, defines a conscious mental state as one which is the target of another, higher-order mental state. That is, one's state, S, is conscious just in case one is having a thought to the effect that one is in S. (This latter thought need not be conscious, and generally will not be conscious, unless or until a yet higher-order thought (HOT) occurs about it.) Theories of this kind are, for obvious reasons, labeled HOT theories of consciousness. There are a number of distinct versions, but they all share the idea that a mental state becomes conscious by becoming the target of another mental state. It is a striking fact that our conscious states all seem to be instantly available to introspection. The HOT approach takes this as a defining feature of consciousness itself. The nature of a mental state's 'availability' to introspection remains under dispute with accounts ranging from equating availability with active introspection (as in David Rosenthal's original HOT theory), equating it with potential introspection (as in Peter Carruthers's updated HOT theory) or even with (in the account of Daniel Dennett) the state's gaining control of mental state describing speech production systems.

Opposed to the HOT approach is one that denies that any HOT is required for a state to become conscious but instead postulates that consciousness is to be explicated in terms of first-order representation (two notable proponents of which are Fred Dretske and Michael Tye). What we are conscious of just is what our mental states are representing. A famous argument in favor of this approach stems from the so-called transparency of mental states. Transparency is a notable feature of the phenomenology of consciousness – there seems to be nothing between the object of consciousness and our experience. Try this experiment. Look carefully at a nearby coffee mug. Now try to focus your introspective attention on the properties of the visual experience of seeing the mug. You will find that nothing appears to consciousness except the features of the mug itself. The experience is thus transparent. This encourages defenders of the first-order representational theory of consciousness to think that if we could find out how the brain represents things in perception or thought we would be well on the way to solving the problem of consciousness. And, of course, there is the hope that if consciousness could be reduced to representation then the prospects of naturalization look brighter insofar as 'representation' looks to be a more straightforward target than subjective consciousness (on this point, the first order and HOT approaches agree) and in fact there are a number of theories of naturalized representation currently on offer (though, typically for philosophical theories, they all suffer from severe flaws). One problem for this approach is that representation seems too easy

to find, too cheap to be the true gold of consciousness. If to be conscious is simply to represent, will not consciousness turn up just about everywhere? We can dismiss books, movies, and the like as derivative representation, dependent upon the conventions of conscious beings. But perhaps even original representation is too widespread and a kind of panpsychism threatens to invade our information-laden world. The standard reply is to mark out some special field of representation within cognitive systems as the home of consciousness. But then the recurring problem of explaining just why these representations enjoy the gift of subjectivity returns.

Champions of the first-order theories do not downplay the importance of introspective access to our own states of consciousness, but they do not think that such access is the hallmark of consciousness. Rather, introspection is an 'add on' optional accessory, available only to creatures with very sophisticated conceptual equipment. Consciousness itself is a simpler, more primitive feature of the mind. Thus an instructive contrast between the two approaches can be found in their attitudes toward animal consciousness. The first-order theories see nothing remarkable about animal consciousness. Animals possess cognitive systems which represent various features of their bodies and their environments (most especially the relative biological value of these features encoded at the most basic level in the information sources we call pain and pleasure). Thus it is to be expected that we and they share a fundamental sort of basic consciousness. The HOT approach finds animal consciousness not so easy to explain. Because of the requirement that a mental state is conscious only if there is a thought about it, animals will be conscious only if they possess the concepts needed to think that they are in mental states. That is, simply in order to feel pain animals will have to have mental state concepts, at least the concept of pain as a mental state. That animals possess such concepts seems – to put it mildly – intuitively unlikely. Various responses to this problem have been advanced, from the 'moderate' claim that the needed concepts are not, in the final analysis, really very sophisticated to the 'radical' claim that this just shows that animals are not conscious beings. (Here we find another curious echo of Cartesian doctrine, the infamous idea that animals are mere automata.)

An extremely interesting problem with these approaches stems from the recent appeal of externalist accounts of representation or meaning (see above) in which representational content accrues to a state only if that state is suitably connected to its referent. An immediate problem is that, just as a randomly created piece of paper that happened to be identical to a US dollar bill would not be genuine currency, so too a randomly created entity that happened to be identical to a human being would not possess any representational states. If consciousness is in some way identical to or even just dependent on representation then it follows that the random human would be entirely nonconscious. In the face of its behavior (indistinguishable from ours), neural processes (identical to ours), ability to (seemingly) converse intelligently, etc., this conclusion seems highly suspect.

The representational theories remain under vigorous development and, while they may not prove to be the answer to the problem of consciousness, they are shedding considerable light on it. But the difficulties facing all the attempts to naturalize consciousness have come to seem quite daunting and some philosophers have recently reventured into old metaphysical waters, defending a modern form of dualism (David Chalmers) or even a radical panpsychism (Galen Strawson).

Consciousness thus remains intractable, and the main problems retain historical familiarity. How can mental states cause things in the world? How do physical processes generate or underlie consciousness? Why does consciousness exist at all? And can (or should) consciousness be understood as an emergent feature of the world (and if so of what brand of emergence) or does it somehow stand as a fundamental and irreducible aspect of a world, which is not exhausted by the physical?

See also: Animal Consciousness; History of Consciousness Science; Mental Representation and Consciousness; The MindBody Problem; William James on the Mind and Its Fringes.

Suggested Readings

Burge T (1979) Individualism and the mental. Midwest Studies in Philosophy 4: 73–122. (Seminal article on the externalist theory of mental content).

Campbell N (ed.) (2003) Mental Causation and the Metaphysics of Mind: A Reader. Peterborough, ON: Broadview Press. (Collects together important articles on

the mental causation debate, especially with reference to Davidson's anomalous monism.)
Carruthers P (2000) Phenomenal Consciousness. Cambridge: Cambridge University Press. (Extensive development and defense of HOT theory of consciousness.)
Chalmers D (1996) The Conscious Mind. Oxford: Oxford University Press. (Important book by the philosopher who jumpstarted a resurgence of interest in the problem of subjective consciousness in both philosophy and the sciences; includes a defense of a form of dualism.)
Chalmers D (ed.) (2002) Philosophy of Mind: Classical and Contemporary Readings. Oxford: Oxford University Press. (Compendious collection of historical and current articles across the whole range of philosophy of mind.)
Dennett D (1991) Consciousness Explained. Boston: Little Brown & Co. (Intriguing if idiosyncratic theory of consciousness in the HOT theory mold.)
Descartes R (1641/1985) Meditations on first Philosophy. In: Cottingham J, Stoothoff R, and Murdoch D (eds.) The Philosophical Writings of Descartes. Cambridge: Cambridge University Press. (Essential for understanding the problem of consciousness and how it has been approached over the last 400 years.)
Dretske F (1995) Naturalizing the Mind. Cambridge, MA: MIT Press. (Important version of the first-order representational theory of consciousness.)
Kim J (1998) Mind in the Physical World: An Essay on the Mind-Body Problem and Mental Causation. Cambridge, MA: MIT Press. (Important effort to grapple with the problem of naturalizing the mind and mental causation.)
McLaughlin B (1992) The rise and fall of British emergentism. In: Beckermann A, Flohr H, and Kim J (eds.) Emergence or Reduction? Prospects for Nonreductive Physicalism. Berlin: De Gruyter. (Very interesting account of radical emergentism and its fate.)
Putnam H (1975) The meaning of 'meaning.' Minnesota Studies in the Philosophy of Science 7: 131–193. Reprinted in Putnam's Philosophical Papers, vol. 2. Cambridge, MA: Cambridge University Press. (Seminal article on the externalist theory of mental content.)
Rosenthal D (1986) Two concepts of consciousness. Philosophical Studies 49: 329–359. (Pioneering paper on the HOT theory of consciousness.)
Seager W (1999) Theories of Consciousness. London: Routledge. (Overview and criticism of main philosophical theories of consciousness.)
Strawson G (2007) Consciousness and Its Place in Nature. Exeter: Imprint Academic. (Defense of panpsychism with multiple replies by – mostly – nonpanpsychists.)
Tye M (1995) Ten Problems of Consciousness. Cambridge, MA: MIT Press. (Important version of the first-order representational theory of consciousness.)

Biographical Sketch

William Seager is a professor of philosophy at the University of Toronto. He was born and raised in Edmonton, Alberta. He received his BA and MA from the University of Alberta and then moved to Toronto for his PhD work where he has remained ever since. His main research interests are in the philosophy of mind, especially the problem of consciousness and the philosophy of science. His most recent book is *Theories of Consciousness* (Routledge, 1999) and his most recent article is 'The intrinsic nature argument for panpsychism' (in G. Strawson's *Consciousness and Its Place in Nature*, Imprint, 2007).

Hypnosis and Suggestion

A J Barnier, Macquarie University, Sydney, NSW, Australia
D A Oakley, University College London, London, UK; Cardiff University, Cardiff, UK

Glossary

Absorption – A personality characteristic involving an openness to experience emotional and cognitive alterations across a variety of situations.
Dissociation – The splitting off of mental processes or mental contents from the main body of consciousness, potentially allowing simultaneous streams of thoughts or behaviors.
Hypnotic analogue – In experimental psychopathology, the use of hypnosis to recreate in the laboratory clinical features of pathological conditions and model their underlying processes.
Hypnotic induction – The procedure used by a hypnotist to begin the experience of hypnosis, usually involving instructions to relax, concentrate on the hypnotist's words, and focus attention inward.
Hypnotizability – An individual's ability to experience the effects of hypnosis, often measured by standardized scales that sum the number of hypnotic suggestions that the person successfully responds to.
Instrumental hypnosis – Research that uses hypnosis as a tool to investigate phenomena outside its immediate domain.
Intrinsic hypnosis – Research that focuses on the phenomena and nature of hypnosis itself.
Posthypnotic suggestion – A suggestion given during hypnosis that asks the individual to show a particular behavior or have a particular experience after hypnosis in response to a specific cue.
Suggestion – A direct or indirect instruction given to an individual inside (or outside) hypnosis, that leads to responses experienced as involuntary and real.
Trance – The change in mental state said to be produced by an hypnotic induction, characterized by focused attention, absorption, relaxation, sense of automaticity and altered self-agency.

Introduction: the Phenomena and Phenomenology of Hypnosis

Hypnosis has been defined as:

> a process in which one person, designated the hypnotist, offers suggestions to another person, designated the subject, for imaginative experiences entailing alterations in perception, memory, and action. In the classic case, these experiences are associated with a degree of subjective conviction bordering on delusion and an experienced involuntariness bordering on compulsion. As such, the phenomena of hypnosis reflect alterations in consciousness that take place in the context of a social interaction (Kihlstrom, 2008: p. 21, chapter in Nash and Barnier, 2008).

This widely accepted definition captures two aspects of hypnosis: 'hypnosis-as-procedure,' what the hypnotist does, and 'hypnosis-as-product,' what the subject experiences.

The procedures of hypnosis are straightforward. To begin, the hypnotist tells the subject they will receive suggestions to experience themselves and the world in different ways and they should respond however they feel comfortable. This introduction aims to distinguish suggestions in the hypnotic setting from other, everyday forms of suggestion. The hypnotist then administers an 'induction.' This typically guides the subject to close their eyes, to relax and to narrow the focus of their attention and absorption. The particular details, length, and style of the induction are mostly irrelevant so long as it meets the

subject's expectations (i.e., that what they are involved in is hypnosis). Successful inductions encourage the subject to concentrate on the hypnotist's communications (reinforced, for instance, by a dimly lit room with few salient features or by closing their eyes) and motivate in the subject a 'cognitive preparedness to respond.' In such a highly motivating, yet impoverished context, the subject turns their attention inward, focuses exclusively on the hypnotist's voice, and keeps competing thoughts to a minimum. Following the induction (or sometimes embedded within it), the hypnotist gives a series of suggestions for alterations in perception, memory, action, thought, or emotion. He or she tests and cancels each suggestion in turn. Finally, the hypnotist administers the deinduction, a series of instructions to terminate the experience of hypnosis. Whereas most suggestions are administered, tested, and canceled in the time between the induction and deinduction, 'post-hypnotic suggestions' are administered during hypnosis, but tested and canceled after hypnosis. In this way, the effects of hypnosis can, for some people and in some situations, extend after hypnosis, over time and into different contexts.

Once the four elements of introduction, induction, suggestions, and deinduction have been administered, hypnosis-as-procedure has occurred. But it is the subject's reaction to them – hypnosis-as-product – that makes hypnosis exceptional. For some subjects with sufficient aptitude and positive attitudes, the induction acts as a 'switch' to enter or experience a state of hypnosis. Within this state – which combines cognitive, social, and interpersonal influences – the hypnotist's suggestions produce powerful changes in the contents of consciousness across the domain of hypnosis – what Ernest Hilgard described as the ''common topics that we study when we engage in hypnotic research'' (Hilgard, 1973, p. 972). These include: (1) ideomotor action and catalepsy; (2) hallucinations (both positive and negative, including analgesia and perceptual distortion); (3) age regression and dreams; (4) amnesia and hypermnesia; and (5) posthypnotic suggestion. To illustrate with some classic hypnotic phenomena, subjects given suggestions for 'analgesia,' 'hallucination,' 'age regression' and 'amnesia,' for instance, claim to feel no pain even when they are exposed to painful stimuli, report seeing objects in the room that aren't actually there, behave as if they were a small child again, and have trouble remembering things that just happened to them.

Across this domain, responding to suggestion during hypnosis feels unusual. In particular, hypnotized subjects experience hypnotic phenomena with 'involuntariness bordering on compulsion' and 'conviction bordering on delusion.' But it is not just that hypnotic responses happen easily or seem real. To the hypnotized person, they feel 'surprisingly easy' and 'surprisingly real.' Surprising ease is illustrated by a study of posthypnotic suggestion in which, during hypnosis, subjects received a suggestion that after hypnosis they would rub their right earlobe when they heard the hypnotist say ''Well, what did you think of that?'' As soon as one female subject heard the cue, her hand and arm began to move toward her ear, and as her hand and arm moved, she watched them with a look of surprise and puzzlement on her face. She experienced her (post)hypnotic ear rubbing response as surprisingly and overwhelmingly involuntary. Surprising reality is illustrated by a study of hypnotic delusions of sex change in which, during hypnosis, subjects received a suggestion to become the opposite sex (for male subjects to become female and for female subjects to become male). One male subject changed his name to a female name and described his appearance as female. When asked to open his eyes to look at an image of his actual self on a video monitor, he said ''That's not me, I don't look like that.'' He experienced his hypnotic sex change response as surprisingly, overwhelmingly and, in his words, 'disgustingly' real.

These sometimes flamboyant alterations in consciousness are the hallmarks of hypnosis. It is tempting to interpret them in extreme ways: as reflecting a fundamental shift in cognitive or neural processing that transforms or overcomes normal capacities in some way or, alternatively, as reflecting mere fakery and role-playing. The truth, distinct from the hype, of hypnosis is more complex and subtle. To provide background to this complexity, 'Fundamentals' covers fundamentals of hypnosis. This section: (1) sketches the history

of hypnosis, noting in particular its use as a window on consciousness; (2) discusses individual differences in susceptibility to hypnotic (and non-hypnotic) suggestions; and (3) describes current explanations of hypnosis in terms of the primary impact of hypnosis, the locus of its effects and the nature of the hypnotic illusion. 'New horizons' builds on this foundation to review recent 'intrinsic' and 'instrumental' collaborations between hypnosis researchers and neuroscientists that: (1) reveal the neural basis of hypnosis; and (2) use hypnosis to create temporary, functional disturbances of consciousness, belief, and action as controllable analogues of perplexing clinical conditions.

Before turning to these sections, it is useful to clarify one major issue. The definitions (above) of hypnosis-as-procedure and hypnosis-as-product take for granted (but do not address directly) the possibility that the hypnotic induction creates an altered state of consciousness – sometimes labeled 'trance' – that significantly influences the way suggestions are responded to during hypnosis. By this view, the induction element of the hypnotic procedure is a crucial precondition for the subsequent element of suggestion, because it produces a 'background state' that facilitates 'true' hypnotic responding. However, the interpretation of hypnosis as an altered state has been controversial in the field for a variety of reasons. Some theorists are uncomfortable appealing to, what they see as, a 'special process' to explain hypnosis. Other theorists argue that since hypnosis is not associated with a unique physiological signature (as dreaming during sleep is associated with rapid eye movements), it cannot truly be an altered state. In John Kihlstrom's view, "the conclusion that hypnosis reflects an altered state of consciousness seems unavoidable" (Kihlstrom, 2008, p. 35). Recent neuroimaging work (described in the 'New horizons' section) may help to resolve this issue, by starting to separate and reveal the neural consequences of an induction, of a suggestion and of an induction plus a suggestion. But as Kihlstrom notes, to fully understand the nature of hypnotic responding, we will need to look for convergence among four relevant variables: the administration of an induction, changes in subjective experience, changes in overt behavior, and physiological correlates.

Fundamentals

An Historical Overview

Hypnosis is usually dated to the work of Franz Anton Mesmer, a Viennese physician practicing in France in the 1770s and 1780s. He developed a therapeutic technique called 'animal magnetism' or 'mesmerism,' which he claimed could heal people. According to Mesmer, a magnetic force pervaded the universe, inside and outside people's bodies (which was consistent with physical science views of the time); disruption of the harmonious flow of this force through the body caused disease. Mesmer claimed that he could treat these disruptions with his 'magnetic techniques,' which included passing his hands close to the patient's body. This resulted in a 'crisis' (a kind of convulsive fit) and an apparent cure. During this crisis or trance, patients displayed many of the behaviors that we now identify with hypnosis.

The French medical establishment reacted with hostility to Mesmer's claims and techniques. In 1784, the King of France, Louis XVI, established a Commission of Inquiry into Animal Magnetism, known as the 'Benjamin Franklin Commission' because it was presided over by Benjamin Franklin, then American Commissioner to France. Using systematic and sophisticated methods of public observation, self study, case study and hypothesis testing – the first experimental studies of hypnosis – the Commissioners aimed to identify the true causes of the effects of animal magnetism. Although the Commissioners did not dispute that Mesmer's therapy might have therapeutic benefits, they disputed the cause. They found no evidence for a universal magnetic fluid, and instead argued that the effects of mesmerism were due to psychological factors: to touch, imitation, and imagination.

In 1843, a Scottish physician, James Braid, gave animal magnetism a new name and a new explanation. Braid was the first to coin the term 'hypnosis' and he explained it in terms of a physiological process: by his view, fixating attention on a bright moving object fatigued certain parts of the brain and caused a trance or, what he called, 'nervous sleep.' Initially he thought that sleep was involved, but later acknowledged (and we now conclusively know) that sleep is not involved in hypnosis.

During the mid-nineteenth century, evidence for the medical value of hypnosis accumulated. For instance, James Esdaile in India reported success with hypnosis as the sole anaesthetic in 345 major operations. This clinical role for hypnosis – managing pain, whether acute or chronic – is still relevant today. However, hypnosis became less popular after chemical anaesthetics were developed and widely available.

The next great surge of interest in hypnosis was in the 1880s and 1890s, particularly in France, where a debate raged between Jean-Martin Charcot, professor of neurology at La Salpetriere Hospital in Paris, and Hippolyte Bernheim, professor of medicine at the University of Nancy. Charcot believed that hypnosis was a form of latent psychopathology; he believed that hypnotizable people are vulnerable to mental illness, particularly to what was then known as hysteria. Charcot influenced some now famous names in the field of psychology, including Sigmund Freud and Pierre Janet. Janet developed his theory of dissociation after spending time with Charcot; he argued that in hypnosis and some pathological conditions, the mind can be split into different streams, with one stream or part hidden from consciousness. Bernheim, of the Nancy School, had a very different view. He argued that hypnosis relates to suggestibility – the degree to which a person will accept and act on suggestions; he believed that hypnosis was not pathological. This theme, that hypnotic behavior is simply a social reaction to the suggestions of the hypnotist, has continued to this day. In extreme versions, hypnosis is considered no more or less than faking.

Hypnosis as a window on consciousness

According to historian Henri Ellenberger, from the late eighteenth century to the early twentieth century (and particularly in the last decades of the nineteenth century) hypnosis was seen as the 'royal road to the unknown mind.' Hypnosis was an important source of data as researchers worked to explain the architecture and operation of the cognitive system, and how it related to psychological disorders. In a typical experiment of the 1880s, Edmund Gurney investigated posthypnotic phenomena using a planchette, which is an automatic writing device consisting of a small triangular or heart-shaped board supported by two castors and a vertical pencil. When lightly touched by the fingertips, it records muscle movements, including messages that were thought to be of unconscious, subconscious, or supernatural origin (as popularized by the 'Ouija' board). In these experiments, Gurney's assistant presented subjects with information during hypnosis, and suggested that they would be able to write the information with the planchette only after awakening. Following hypnosis, subjects were offered money (one guinea) if they could repeat the hypnotic information, but none ever recalled it. However, when their hand was placed on the planchette, which was hidden behind a screen, they were able to write the information.

Gurney interpreted these findings as demonstrating the existence of secondary streams of consciousness or subconscious 'personalities' in normal persons. In the 1890s Frederic Myers extended this interpretation into a model of cognitive processing and awareness in which he conceptualized consciousness as a spectrum, ranging from awareness of the automatic regulatory processes of the body, through subliminal consciousness (as demonstrated by hypnotic phenomena) and supraliminal (or everyday) consciousness, to awareness of parapsychological phenomenon. Most importantly, he believed that hidden streams of consciousness, as highlighted by hypnotic and posthypnotic responding, existed in normal individuals. This was different from Janet's view, based also in part on investigations of hypnosis with hysterical patients, that subconscious processes could only be found in psychologically disturbed individuals. Janet proposed that conditions of stress lead to 'désagrégation,' or the detaching of ideas from the mainstream of consciousness and the development of neurotic symptoms or secondary personalities. William James, a contemporary of both Gurney and Myers, was strongly influenced by their work when developing his views of consciousness in general and hypnosis in particular. To explain hypnosis experiments involving difficult cognitive tasks performed seemingly out of awareness, James proposed a cognitive mechanism that was stored in consciousness, but split off or dissociated from the rest of the individual's mind, only to reassert itself at the appropriate time. Although later challenged, these late nineteenth

century explanations of hypnosis – as creating a temporary disunity of consciousness or as reflecting the natural disunity of consciousness – strongly influenced a range of contemporary accounts (described below).

Individual Differences in Hypnotic (and Nonhypnotic) Suggestibility

From the earliest days of Mesmerism, workers in the field recognized that some people are more susceptible to hypnotic suggestions than others. In 1892, Albert von Schrenck-Notzing published the First International Statistics of Susceptibility to Hypnosis, reporting individual differences across 8705 people hypnotized by 15 clinicians from different countries. Just over 100 years later, Kevin McConkey and colleagues published individual differences across 4574 people hypnotized over a period of 10 years at Macquarie University in Australia. Although collected a century apart, the distribution of scores is remarkably similar with, for example, 15.1% in 1892 and 14.3% in 1996 identified as high hypnotizable, or 'somnambulistic' in the language of the 1890s.

Measuring individual differences

In modern research and practice, hypnotizability is measured by standardized scales, which were developed by Andre Weitzenhoffer and Ernest Hilgard at Stanford University in the 1950s. Their development was motivated by the view that reliable and valid measurement is essential to scientific advance in a field. Still in use today, these scales typically involve: an introduction, an induction, a set of standard suggestions (or items) that are scored according to predefined behavioral criteria, and a deinduction. Hypnotizability scores are calculated simply by summing the items that the individual passes according to the behavioral criteria, although subjective criteria sometimes are also used. For over 50 years these scales have provided a 'gold standard' metric for work on hypnosis and hypnotizability.

The concept of hypnotizability as an individual difference dimension, much like intelligence, is enshrined in these scales and supported by a substantial body of worldwide norming data (e.g., on the Harvard Group Scale of Hypnotic Susceptibility, Form A (HGSHS:A), a group adaptation of Weitzenhoffer and Hilgard's original Stanford Hypnotic Susceptibility Scale, Form A (SHSS:A)). These norms indicate that approximately 10%–15% of subjects are 'high hypnotizable' (passing all or most items on the scale), approximately 10%–15% are 'low hypnotizable' (passing only a few of the items), and the remaining 70%–80% are 'medium hypnotizable' (passing some, but not other items). In other words, hypnotizability appears normally distributed in the population. It is also remarkably stable over time (with 25 year test–retest reliability >.80) and remarkably consistent across scales, geography, and language.

Some people conceptualize hypnotizability, not as a dimension on which everyone falls somewhere, but as types of subjects who differ either in ability level or in underlying processes. For instance, some researchers have argued that very high hypnotizable people – known as 'hypnotic virtuosos' – are qualitatively different from all other subjects and are the only 'true' hypnotic responders. Other researchers have argued that the 10%–15% of high hypnotizable people in the population can be differentiated by their approach to and experiences of hypnosis: some highs are very active or 'constructive' in their response to hypnotic suggestions, and use their imaginative capacities to create compelling hypnotic responses; other highs are less active – 'concentrative' – waiting for the hypnotist's words to take effect, perhaps via dissociative processes; finally, some highs respond chiefly because they are highly motivated and hold a 'positive response set.'

Whether or not hypnotizability is best thought of in terms of a dimension or types, it interacts in important ways with the content of hypnotic items. Standardized scales of hypnotizability, as well as experimental and clinical uses of hypnosis, ask people to do essentially three different kinds of things during hypnosis: (1) to transform ideas into actions, such as in the hand lowering suggestion "your (outstretched) arm is feeling heavier and heavier and falling down"; these are known as 'ideomotor items'; (2) to challenge the reality of a suggested state of affairs, such as in the finger lock suggestion "your fingers are tightly interlocked; now try to take them apart"; these are known as 'challenge items'; and (3) to experience alterations

in perceptual and/or cognitive processing, such as in the fly hallucination suggestion "there is an (hallucinated) fly buzzing around your head"; these are known as 'perceptual-cognitive items.' Pass rates show that some types of suggestions are easier than others, whereas some types are the province of only the most talented hypnotic subjects. For instance, ideomotor items are experienced by a large proportion of hypnotic subjects, even some low hypnotizable people. Ideomotor items generally are easier than challenge items, even when both involve motor actions (e.g., a suggestion that your hands are moving apart versus a suggestion that your arm is immobilized and cannot move). Facilitating an action via suggestion appears easier than inhibiting an action. Perceptual-cognitive items are the most difficult items and generally limited to high hypnotizable people. This is especially true of perceptual-cognitive items such as negative visual hallucination (a suggestion that you won't see something that is actually present in front of you), where inhibiting aspects of genuine reality in favour of the suggested reality seems particularly difficult.

Drawing the link between individual differences and item content, a current view (based on sophisticated statistical analysis of large sets of data from standardized measures) conceptualizes hypnotizability as a combination of: (1) an underlying dimension of general hypnotizability along which everybody falls and differs, which predicts responding in general, plus (2) distinguishable component abilities or building blocks that some people may or may not have, which predict response to certain kinds of hypnotic alterations in consciousness.

The source of individual differences

To explain individual differences in hypnotizability, research has pointed tentatively both to genetic contributions and to developmental histories. In terms of genes, for instance, identical twins are more closely matched on hypnotizability scores than are fraternal twins and siblings. In terms of development, researchers have focused especially on imaginative play in childhood as one pathway to high hypnotizability, and perhaps particular types of high hypnotizability. However, more work is needed in this area.

Most work on individual differences has searched instead for personality or cognitive characteristics that predict hypnotic ability. The search for correlates has trawled through personality variables, such as those from the Minnesota Mutliphasic Personality Inventory, the Eysenck Personality Questionnaire, and the Big Five Inventory, as well as cognitive style variables, intelligence, and gender. But as hypnosis scholars have lamented, correlations with hypnotizability are typically weak, inconsistent, or nonexistent. Hypnotizability does correlate, however, with 'absorption,' a personality characteristic that involves an openness to experience emotional and cognitive alterations across a variety of situations. Although the relationship is modest ($r = .30$), it implies that hypnotic ability is related to a more general tendency to respond in characteristic ways to sensory, cognitive, and imaginative experiences.

Some researchers have raised doubts about the relevance of absorption because of 'context effects'; correlations between hypnotizability and absorption tend to be much lower and even nonsignificant when the two constructs are measured in separate contexts. But given that personality variables are inherently contextual, such fluctuations in the relationship do not rule out a role for absorption-like tendencies in hypnosis. Indeed, absorption correlates most strongly with perceptual-cognitive hypnotic items (and weakly with ideomotor and challenge items), which suggests that subjects may need absorption only for difficult hypnotic items. In other words, absorption may become apparent only in settings or in circumstances (whether hypnotic or hypnotic-like) that are consistent with, encourage, or require its expression.

Another line of research has focused on subjects' attitudes and expectations. People with negative attitudes may be unwilling to participate in hypnosis and, if they do, their negative attitudes may suppress performance, leading to low scores. People with positive attitudes, in contrast, may be low, medium, or high hypnotizable; positive attitudes are necessary but not sufficient. People's expectations about the nature and degree of their experience can also influence how they respond. But these expectancies are not always accurate. Sometimes low hypnotizable people are most accurate in predicting their responsiveness, whereas

mediums and highs underestimate their responsiveness. Other times mediums and highs are most accurate, whereas low overestimate. Despite such variations, some researchers have argued that expectancies mainly or even solely determine hypnotizability. Recent analyses of expectancy judgements collected throughout standardized scales confirm that expectancies play a significant role. However, underlying hypnotic aptitude still makes a central contribution to hypnotic performances.

A more recent line of research into individual differences has focused on cognitive processing of talented hypnotic subjects, especially in terms of attention and automaticity. Hypnotizable people have been tested, both inside and outside the hypnotic setting, on a range of cognitive tasks including variations of the color naming Stroop task (which measures interference in color naming from automatic word reading), negative priming tasks (which measure ability to suppress irrelevant information from working memory), and simple and go/no go reaction time tasks (which measure time taken to resolve conflicts in working memory). Initial results indicate that high hypnotizable people process information more automatically, learn and automatize tasks more quickly, can in some circumstances overcome seemingly automatic processes and resolve working memory conflicts differently to less hypnotizable people. These findings suggest that hypnosis and hypnotizability are closely connected to more general processes of attention.

Hypnotic versus nonhypnotic suggestibility

Hypnosis and suggestion have long been linked; we talk, for instance, of hypnotic 'suggestibility' as an individual difference ability and of hypnotic 'suggestions' to create particular responses. But is suggestibility in and out of hypnosis the same? And to what degree do different forms of suggestion belong within the domain of hypnosis? Some researchers have argued that hypnotic responding is quite different from responding due, for instance, to conformity, gullibility, and persuasibility. Other researchers have argued that hypnotic suggestibility is simply nonhypnotic suggestibility given a small boost from expectancy and motivation. By this second view, the hypnotic induction adds little or nothing, because (with or without an induction) subjects respond mainly on the basis of their nonhypnotic suggestibility. But the induction does add something. Hypnotic subjects certainly can experience hypnotic-like effects without one, but the onset of these effects is more rapid and their impact more compelling following an induction. Perhaps more tellingly, whereas absorption correlates most strongly with difficult hypnotic items (as noted above), nonhypnotic suggestibility (e.g., direct and indirect measures of placebo effects) correlates most strongly with easy hypnotic items. This implies that suggestibility inside and outside hypnosis is not the same, especially the hypnotic ability needed to respond to demanding perceptual-cognitive items.

Within the domain of suggestion more broadly, there is increasing evidence that the many different suggestibilities do not resolve into essentially one form of suggestibility. Researchers have distinguished between primary suggestibility (involving direct verbal suggestions for bodily movements), secondary suggestibility (involving indirect, nonverbal suggestions for sensory-perceptual experiences), and tertiary suggestibility (involving conformity, persuasion, and other forms of social influence), as well as between specific concepts such as placebo effects and interrogative suggestibility. Recent data collected across large sets of suggestibility tasks (including the HGSHS:A) have indicated that hypnotic suggestibility is an independent phenomenon – related to a set of unique abilities (perhaps cognitive, as described above), but unrelated to other measures of suggestibility.

Explanations of Hypnosis

Theories of hypnosis aim to explain the behaviors and experiences of the hypnotized subject. It is not enough to simply measure and explain what people do – lower their arms, fail to unlock their fingers, or swat at a nonexistent fly. There are many reasons why someone might do any of these things; not all of them hypnotic reasons. We need to measure and explain how this responding felt – whether involuntary or real. Theory building has been challenging because hypnosis essentially is a private experience of an altered sense of self and the world, and thus difficult to study directly. Given that many theorists have highlighted the

almost 'deluded' degree to which subjects come to believe in the events of hypnosis, how do we explain the fact that mere words from the hypnotist generate such compelling subjective changes in the contents of consciousness? Readers can find excellent discussions of the timeline of developing theoretical views in classic and modern texts, and readers can easily source full accounts of current theories of hypnosis (see 'Suggested Readings'). So it is useful instead to discuss current explanations in terms of three (related) conceptual issues: (1) the primary impact of hypnosis; (2) the locus of its effects; and (3) the nature of the hypnotic illusion.

The primary impact of hypnosis

There are two broad classes of possible effects of hypnosis. Hypnosis may influence a person's introspective awareness or consciousness of their 'system state' (i.e., his/her awareness of the state of their cognitive and physical systems) – labeled 'explicit' effects. An example is a person reporting that they feel no sensation or pain following a suggestion for hypnotic anaesthesia even when their hand is immersed in ice cold water. Alternatively, hypnosis may influence indicators that reflect the actual state of their cognitive or physical systems – labeled 'implicit' effects. An example is a person's physiological reactivity (e.g., heart rate) following a suggestion for hypnotic anaesthesia when their hand is immersed in ice cold water.

Irrespective of the details of their particular explanations, many theorists have argued that hypnosis influences introspective awareness (explicit effects), and not actual system state (implicit effects). Research findings overwhelmingly support this view. For instance, consider the impact of hypnosis on memory. The terms explicit and implicit were originally used in the memory literature to refer to two separate forms of memory: explicit memory refers to conscious recall of an event, whereas implicit memory refers to changes in performance on a task due to the original event, but without consciously recalling it. Research on posthypnotic amnesia shows differential effects on explicit and implicit memory. For example, in one study on posthypnotic amnesia of autobiographical episodes, high and low hypnotizable people were asked to recall their first day of high school and their first day of university. During hypnosis, they received a posthypnotic amnesia suggestion to forget the episodes either from high school or from university. After hypnosis, they completed category generation and social judgement tasks designed to measure implicit memory of the forgotten episodes, and then tried to explicitly recall them. Although high hypnotizable subjects had difficulty recalling the events targeted by amnesia, they completed the implicit tasks with information from these 'forgotten' episodes. This dissociation in explicit and implicit memory, reported consistently in posthypnotic amnesia studies, supports the view that hypnosis mainly affects introspective awareness of cognitive and physical systems.

The explicit/implicit distinction has been extended also to perception, where research on hypnotic blindness also shows that hypnosis impairs explicit, but not implicit, forms of processing. For example, in one study, during hypnosis high hypnotizable people received a blindness suggestion that they would not see anything in their visual field and then were shown a set of homophones with unusual spellings (e.g., 'stake' instead of 'steak'). After hypnosis, they completed a word spelling task designed to measure implicit perception, and then tried to explicitly recall the words. Although high hypnotizable subjects had difficulty recalling the 'unseen' words presented while they were hypnotically blind, they spelled these homophones consistent with having seen them (e.g., they spelled stake instead of steak). Again, this dissociation is consistent with hypnosis influencing introspective awareness. Finally, the explicit/implicit distinction can be extended to action, where motor suggestions impair consciously directed, explicit forms of action, but spare stimulus driven, automatic, implicit forms of action.

Because hypnosis appears to have mainly explicit, but not implicit, effects, it has been linked to a range of psychological and psychiatric disorders, which also feature explicit, rather than implicit, impairments of memory, perception, and action. For example, hypnosis is seen as a non-pathological analogue of dissociative disorders (such as dissociative identity disorder and dissociative amnesia), where patients are unable to consciously recall autobiographical memories and even their identity, but show implicit evidence of

their forgotten past. Hypnosis is seen as a nonpathological analogue of somatoform disorders (such as conversion disorder), where patients report distressing physical impairments that have no compelling medical basis, but show implicit evidence of intact functioning. And hypnosis is seen as a nonpathological analogue of abnormalities in the awareness of action (such as alien control delusion in schizophrenia), where patients report that their limbs move without their conscious control and awareness, but those actions appear normal and motivated.

Despite the majority view that hypnosis impacts only introspective awareness, some explanations and evidence raise the possibility that hypnosis impairs (also or instead) the actual operation of memory, perception, and action systems. For instance, one line of research on visual hallucinations and color processing suggests that hypnosis impairs primary visual perception. That is, hypnosis impairs early rather than late stages of processing. Still other explanations and evidence hint that hypnosis may confer unusual or extraordinary abilities. For instance, one line of research on hypnotic elimination of the Stroop effect suggests that hypnosis can de-automatize highly automatic cognitive processes such as word reading. However, evidence for implicit effects is outweighed by evidence for explicit effects, so more work is needed.

The locus of its effects

Within the domain of hypnosis, as well as outside, theorists have drawn a distinction between 'executive control,' which involves voluntary initiation and termination of thought and action, and 'executive monitoring,' which involves accurately representing objects and events in phenomenal awareness. This distinction is consistent with perspectives from outside hypnosis on dual-system models of action. Such models propose two complementary systems that manage the initiation and control of action. These can be seen in Figure 1, created by Erik Woody and Pamela Sadler to summarize dissociation theories of hypnosis. As Woody and Sadler describe, a higher, executive system (comprised of Executive Control and Executive Monitoring) modulates and monitors subsystems of control and is responsible for volitional acts. A lower system (Subsystems of Control) directly handles selection and tracking of behavior and is responsible for environmentally driven, routine acts. These models include feedback loops for communicating intentions and executive oversight of action (e.g., from Executive Control to Subsystems of Control (b), from Subsystems of Control to Executive Monitoring (e), from Executive Monitoring to Executive Control (d), and from Executive Control to Executive Monitoring (c)).

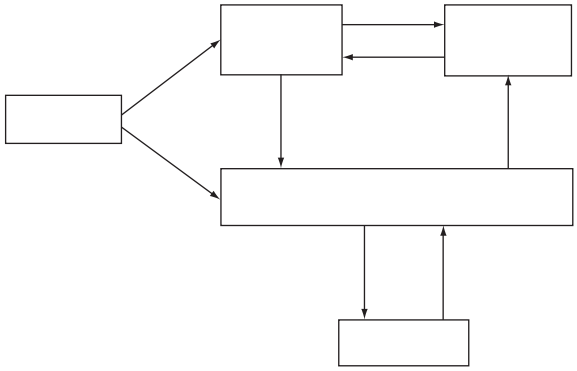


Figure 1 An integrative model of dissociation theories of hypnosis. Reproduced from Figure 4.2 from Woody EZ and Sadler P (2008) Dissociation theories of hypnosis. In: Nash MR and Barnier AJ (eds.), Oxford Handbook of Hypnosis: Theories, Research and Practice, pp. 82–110. Oxford: UK: Oxford University Press, with permission from Oxford University Press. Note. A theory of 'dissociated experience' involves the weakening of path c, and possibly of path e. A theory of 'dissociated control' involves the weakening of path b, and possibly of path a. A theory of 'second-order dissociated control' involves the weakening of path d.

A dominant view of hypnosis, both historically and currently, has proposed that the hypnotic induction initiates a temporary disruption or disunity of consciousness. Conceptualized within dual-system models of action, hypnotic responding is controlled in essentially normal, nonhypnotic ways, but executive monitoring is disrupted. This is represented in Woody and Sadler's Figure (Figure 1) and described by them as broken or weakened links from Executive Control to Executive Monitoring (c), and from Subsystems of Control to Executive Monitoring (e). By this view, Executive Monitoring is dissociated from important information about the causal role of self in hypnotic responding. However, different theorists propose different causes for this impairment in executive

monitoring. For instance, Ernest Hilgard argued that monitoring fails during hypnosis because it is isolated within a separate (dissociated) stream of consciousness. In contrast, Nicholas Spanos argued that social psychological factors cause the disruptions in monitoring. According to Spanos, subjects expect and receive strong social cues that 'good hypnotic responding' is involuntary, which leads them to misattribute the true cause of their hypnotic experiences.

A different set of theories propose instead that the hypnotic induction disrupts executive control and leaves monitoring essentially intact. This is represented in Woody and Sadler's Figure (Figure 1) and described by them as broken or weakened links from Executive Control to Subsystems of Control (b), and from Suggestion to Executive Control (a). By this view, suggestions from the hypnotist bypass executive control to directly activate subsystems of control such that Executive Control is dissociated from the effortful activation of behavior. Again, different theorists point to different factors as the cause of this shift in control. For instance, whereas Ken Bowers and Erik Woody argued that the hypnotic induction lessens executive control so that suggestions directly activate responding, Steven Jay Lynn and Irving Kirsch argued that expectancies directly activate response sets for hypnotic behavior. It is worth noting that theorists that posit changes in executive control generally do not view hypnosis as unleashing new abilities (e.g., overcoming automatic processing). A few theories argue for increased control or capacity, but more research is needed to test this possibility.

Although many theorists identify the locus of hypnotic effects in either executive control or executive monitoring, most recognize that both processes are implicated and related in hypnotic responding. Hilgard, for example, argued against artificially separating control from monitoring functions, since all initiated action is monitored and subject to ongoing correction (although Bowers drew a sharper distinction between the two). Some current theories unambiguously describe hypnosis in terms of feedback between the two processes. Hypnotic behavior may in fact involve impairments of control and/or monitoring depending on the particular circumstances, suggestions, and the individual's aptitudes and processing style. Hypnosis is probably best explained as the outcome of a range of interacting factors.

The nature of the hypnotic illusion

Hypnosis is often described as an illusion. But given different views of altered control versus monitoring, is there an illusion and, if so, what kind of illusion is it? The classic view is that the experience of involuntariness is an illusion; there is no real change in normal control following an hypnotic induction, but subjects mistakenly interpret their internally generated responses as externally caused ones. According to this view, cognitive effort in hypnosis is high but incorrectly experienced as low. Another view discussed above, is that the experience of involuntariness is not an illusion. There is a real change in normal control following an hypnotic induction; subjects genuinely are not in control during hypnosis and their experience accurately represents this. According to this view, cognitive effort in hypnosis is low and correctly experienced as low. A third view discussed above, is that hypnotized people's responses are automatically initiated via response sets. Again, the experience of involuntariness in hypnosis is not an illusion because subjects' responding genuinely is outside of their control. The twist here is that, by this view, people's everyday experience of voluntary control over their actions is considered an illusion; hypnosis temporarily reveals what our conscious life is actually like. According to this view, cognitive effort in hypnosis (and everyday life) is low, but only correctly experienced as such during hypnosis. (Daniel Wegner makes a related argument, that much of human behavior is experienced (perhaps incorrectly) as within our control. However, he makes less strong claims about the actual state of control (especially in the face of a complex and increasing body of data on human action), and focuses instead on the conditions that support the ubiquitous illusion of control.)

This discussion raises a final theoretical issue, which has generated significant controversy inside and outside the field (as referred to earlier). Is hypnosis a distinct or special state of consciousness? This is best addressed by turning to recent insights provided by neuroimaging techniques.

New Horizons

Research on hypnosis can be categorized as 'intrinsic' or 'instrumental.' Intrinsic research focuses on the phenomena and nature of hypnosis itself, whereas instrumental research uses hypnosis as a tool to investigate phenomena outside its immediate domain. Since at least the 1950s, when the introduction of hypnotizability measures transformed the scientific study of hypnosis, researchers have built a substantial database on the intrinsic nature and instrumental value of hypnosis. With the advent of powerful neuroimaging techniques (e.g., positron emission tomography (PET); and functional magnetic resonance imagining (fMRI)), hypnosis researchers and neuroscientists have collaborated to address questions of consciousness – attention, action, memory, belief, and self-awareness – both normal and abnormal. PET and fMRI detect and display images of changes in regional cerebral blood flow associated with brain activity. In general, they identify which areas of the living brain are active during particular tasks. In hypnosis, they offer an objective means to validate subjective reports and to examine the neural correlates of hypnosis and hypnotic phenomena. This new frontier of research allows us to address intrinsic questions about hypnosis in new ways, as well as to expand the reach of instrumental hypnosis.

Intrinsic Research: The Neural Basis of Hypnosis

Neuroimaging studies using PET and fMRI have provided very good evidence that suggestions given in hypnosis are associated with changes in corresponding areas of the brain. In other words, hypnotic alterations in behavior and experience parallel alterations at the neural level. This is illustrated by studies where subjects were given suggestions during hypnosis to experience painful heat in the absence of an actual heat stimulus. These suggestions not only produced subjective reports of heat pain, but also activity in widespread areas of the brain that normally respond to pain (thalamus, anterior cingulate, insula, prefrontal, and parietal cortices). The patterns of activation in these areas were very similar to those produced by an actual heat pain stimulus, but quite different from activation produced when the same hypnotized participants were asked to vividly imagine the painful experience. This outcome and others like it underline the 'as real' quality of the hypnotically suggested experience and underscore that hypnotic experiences are more than simply products of imagination (see next section on instrumental hypnosis for examples from studies of conversion disorder paralysis, auditory hallucinations and delusions of alien control).

Other neuroimaging studies show that specific hypnotic suggestions can have quite selective effects, not only on experience, but on brain activations. For instance, in one pair of neuroimaging studies, during hypnosis subjects were asked to place their hands in a hot water bath and then were given suggestions to modulate this painful experience; the suggestions told subjects that the painful heat stimulus would become either more or less unpleasant, while its perceived intensity remained the same. Notably, activation in one area of the brain associated with evaluating the emotional component of pain (anterior cingulate cortex) was modulated in direct relation to the direction of the suggested change in unpleasantness. In a related study, these researchers used the same procedures, but this time the hypnotic suggestions told subjects that the temperature of the water bath would either increase or decrease (in reality it did not change), while its unpleasantness remained the same. This time, activation in the primary somatosensory cortex, rather than anterior cingulate cortex, changed in direct relation to the direction of the suggested change in temperature. In a third experiment, during hypnosis, high hypnotizable subjects were given suggestions to either add color to a gray-scale stimulus display or to drain the color from a colored display. Subject's reported experiences of altered color processing were again accompanied by clear changes in brain activity; this time in occipital/parietal areas, which are normally associated with color processing. Taken together, these studies again underline the subjective 'reality' of hypnotic experiences and also show how the (experiential and neural) responses can be changed simply by changing the wording of the suggestion.

There is a lot of experimental evidence for brain changes following hypnotic suggestions. However, there is relatively little work on the

effects of the hypnotic induction procedure itself; that is, on changes in mental state independent of direct suggestions. Some researchers have proposed that the hypnotic induction leads to changes in attention, absorption, and critical thinking and that these changes are dependent on frontal cortical attentional systems. In principle, it should be possible to detect the underlying brain processes using neuroimaging techniques. In one study, where the primary aim was to investigate changes in pain experience, the researchers also asked participants to give subjective ratings of both absorption and relaxation during hypnosis. Increasing levels of reported absorption were associated with activity in widespread areas of the brain including anterior cingulate and frontal cortical regions, which the researchers described as an 'executive attentional network.' In contrast, increasing levels of reported relaxation were associated with reductions in activity in midbrain and thalamic areas known to be involved in the regulation of cortical arousal and with increases in activity in areas of anterior cingulate cortex, distinct from those associated with absorption.

There is emerging evidence also that there are differences in the 'resting' (or 'default mode') brain activity of nonhypnotized individuals engaged in a repetitive activity (e.g., watching a reversing checkerboard display) compared to when they are engaged in the same activity, but are hypnotized. In another study, as depth of hypnosis increased in the hypnotic resting state, there was a corresponding increase in activity in lateral prefrontal regions (involved in the maintenance of attention), but a reduction in activity in cortical midline areas of the brain (associated with the normal brain resting state). Ratings of mental state in the hypnotic and nonhypnotic resting conditions indicated that spontaneous conceptual thought, typically associated with the normal resting state, diminished along with increased attentional absorption in the hypnosis resting state. In a different study, which compared performance on word/color conflict procedure (Stroop task) in hypnotized and nonhypnotized conditions, activity increased in anterior cingulate cortex, typically associated with conflict-monitoring, without any change in activity in left frontal cortex, typically associated with cognitive control. The researchers concluded that the hypnotic induction procedure led to a spontaneous decoupling of the normal relationship between the processes of conflict monitoring and cognitive control – an outcome which explained earlier observations that performance on the Stroop task is frequently impaired following an hypnotic induction procedure.

Although the collection of evidence is at an early stage, these results support the view that hypnotic induction procedures recruit frontal attentional systems and alter executive control in individuals exposed to them. Given incidental observations from a number of studies that basic sensory and motor processes are not affected by hypnosis *per se*, these changes due to the induction procedure seem to occur at a relatively late stage of information processing. This suggests that hypnotic effects are the product, not so much of fundamental changes in the way the brain process information, but of influences on decision making within executive systems, which affect the mental contents that enter conscious experience. In other words, effects are on explicit rather than implicit mental processes. Since executive systems operate outside of awareness, resulting changes in perception and motor control will have subjective accompaniments of involuntariness, lack of initiative and passivity.

While it is helpful for clarity of exposition to consider the brain processes underlying hypnotic induction and suggestion separately, this leaves open the important question of what, if anything, emerges from the combination of the two processes. As noted above, some people at least can respond to suggestions outside hypnosis in apparently similar ways to inside hypnosis (i.e., after an hypnotic induction). But a case can be made nonetheless that induction adds something, especially in terms of how quickly the subject responds to the suggestions and/or how compelling their resulting experiences feel. There is very recent neuroimaging evidence that supports this notion of 'added value.' Studies have begun to compare the effect of exactly the same suggestions given to the same individuals inside and outside hypnosis. Results indicate (e.g., for both clinical pain and color processing) that although subjects' experiential reports of changes in response to the suggestions are quite similar across hypnotic and nonhypnotic

conditions, changes in related brain processes are much more marked when the suggestions are given following an hypnotic induction. This again indicates that at a brain level there is greater objective reality to the changes underpinning subjective experiences in hypnosis, which may explain why individuals frequently describe what happens to them during hypnosis as 'virtually real.'

Overall, results from neuroimaging studies are consistent with an identifiably different resting brain state following an hypnotic induction, but this is more a product of the redeployment of existing executive and attentional systems than the creation of fundamentally new forms of mental processing. The state of hypnosis is in this sense a 'normal' one, which depends on existing brain systems. Nevertheless, it enables the creation via suggestion of subjectively powerful and often 'abnormal' changes in conscious experience, which are underpinned by congruent changes in activity in relevant brain areas.

Instrumental Research: Hypnosis as a Controllable Analogue of Disordered Consciousness

Instrumental hypnosis research has provided new insights into processes underlying perplexing clinical conditions, such as hysterical blindness, dissociative amnesia, delusional beliefs, conversion disorder paralysis, anarchic limb and alien control syndromes, auditory hallucinations, and malingering. The combination of neuroimaging techniques with longstanding behavioral and phenomenological methods of experimental psychopathology has increased the power of hypnotic analogues of abnormal changes in conscious experience.

Hypnotic analogues of disrupted perception and belief

Experimental analogues aim to recreate in the laboratory the clinical features of pathological conditions and then model the processes that contribute to them. One example is research on functional (or hysterical) blindness. Functional blindness is a form of conversion disorder where a person experiences an unexpected loss or disruption in visual functioning in the absence of identifiable physical disorder or disease. People with functional blindness report they cannot see, yet their behavior suggests they are still processing visual information (e.g., they might navigate successfully in an unfamiliar room while claiming they cannot see). Hypnosis researchers have created similar patterns of visual experience by giving subjects a suggestion to not see a particular stimulus or to see nothing in their visual field. Just like patients with functional blindness, high hypnotizable people given an hypnotic blindness suggestion claim to see nothing, but respond on visual tasks as if they are still processing available visual information. And just like patients, they strongly defend their claims of blindness even when confronted with evidence of intact visual processing.

Another example of hypnosis recreating features and modeling processes is research on clinical delusions. Clinically deluded people report distorted beliefs that they hold with absolute conviction and maintain in the face of strong challenges. Hypnosis researchers have created similar patterns of false beliefs by giving subjects suggestions for fully formed delusional experiences (e.g., become the opposite sex, become a different person). For example, in one study developing a laboratory analogue of the neuropsychological condition of mirrored-self misidentification, researchers gave high hypnotizable subjects a suggestion during hypnosis to see a stranger, not themselves in a mirror. This created credible, compelling delusions strikingly like clinical cases of mirrored-self misidentification. When one hypnotized male subject opened his eyes to look in the mirror he said ''who's that, another person?'' as he looked around the room to find the person he believed was in the mirror. Current hypnotic work in this field is seeking ways to map the factors that contribute separately or in combination to the content of delusions and to the failure to reject them as implausible.

Hypnotic analogues with neuroimaging of disrupted action, perception, and belief

These laboratory models of functional blindness and clinical delusion focus on (normal and abnormal) patterns of behavior and experience. Other lines of research have added neuroimaging techniques to investigate the relationship between brain, behavior, and experience in analogues of

clinical conditions. One example is research on conversion disorder paralysis, where a person experiences an unexplained paralysis of a limb in the absence of identifiable physical disorder or disease. People with such paralyses report their symptoms as real and involuntary, yet their nerves and muscles are in perfect working order. In an initial neuroimaging (PET) investigation of this condition, researchers asked a woman with unexplained paralysis of her left leg to try to move it. Scans of activation in her brain (in the premotor cortex and cerebellum) suggested genuine attempts to prepare and try to move the limb. However, lack of activation in brain areas responsible for motor action (particularly primary sensorimotor) combined with increased activation in two others brain areas (right anterior cingulate cortex and right orbitofrontal cortex) suggested unconscious inhibition of her intended voluntary movements. In other words, at some unconscious level she was inhibiting the response that she prepared to make when asked to move her paralyzed limb.

Given the behavioral and experiential similarities between conversion disorder and hypnotic phenomena (as noted above), some researchers have wondered whether conversion disorder paralysis and hypnotically suggested paralysis are created in the same way in the brain. In a follow-up to their earlier PET study, these researchers repeated their procedures with a high hypnotizable man given hypnotic suggestions for paralysis of his left leg. Just like the woman with conversion disorder, they asked the hypnotized man to try to move his (hypnotically) paralyzed leg. His brain scans showed similar activation to the woman tested earlier. He too showed activation suggesting inhibition of voluntary attempts to move. Since the brain mechanisms were the same in the clinical and hypnotic conditions, a fair conclusion is that hypnosis is a powerful and controllable way to explore and potentially treat symptoms of conversion disorder.

A second example of hypnotic analogues with neuroimaging is research on hallucinations. In positive hallucinations, a person perceives something that is not present; in negative hallucinations, they fail to perceive something that is present. Hallucinations can be experienced in any sensory modality, but in clinical conditions such as schizophrenia, they are most often auditory or visual. People experiencing hallucinations (e.g., voices telling them to do things or commenting negatively on their behavior) describe them as 'as real as real,' and thus very disturbing or distressing. In a neuroimaging (PET) study, researchers used hypnosis to assess the reality quality of hallucinations. During hypnosis, they scanned brain activations of high hypnotizable people (who they identified in screening as either good hallucinators or poor hallucinators) and low hypnotizable people across three sets of trials; (1) when they listened to a taped auditory message; (2) when they were asked to imagine the tape playing; and (3) when they were told that the tape would play again but it did not. For this third set of trials, high hypnotizable good hallucinators hallucinated the message; they believed they heard the message again and rated it as clear as when they actually heard it. In contrast, high hypnotizable poor hallucinators and low hypnotizables heard nothing. Most importantly, when the high hypnotizable good hallucinators hallucinated the message, their brain activations were similar to when they heard the real message, and quite different from when they merely imagined it. Thus, hallucinated experiences (at least during hypnosis) not only feel real, but are real at a neural level. These findings reinforce the value of hypnosis for studying clinical conditions that otherwise are very difficult to bring into the laboratory.

A final example is research on delusions of alien control, where a person experiences one of their limbs as belonging to them, but believes that it is controlled by someone else. People with alien control (which is common in schizophrenia) report that their arm, hand, or leg, for instance, moves of its own accord, even though their nerves and muscles were intact and working as normal. Since hypnosis also involves disruptions in self-monitoring of actions, in a neuroimaging (PET) study, researchers used hypnosis to investigate brain processes underlying feelings of alien control. They scanned brain activations of high hypnotizable subjects under three conditions: (1) when subjects voluntarily moved their left arm up and down, (2) when their arm was passively moved up and down by a pulley system, and (3) when subjects were given an hypnotic suggestion that the pulley was again moving their arm up and down (even though it was not). This last condition was intended as an analogue

of alien control, where self-produced arm movements should be falsely attributed to the pulley.

As expected, subjects said that when their arm was moved by the pulley and when they thought it was moved by the pulley (during hypnosis), it felt more involuntary than when they moved it themselves. Most importantly, when subjects thought their arm was being moved by the pulley during hypnosis, their brain activations were similar in some ways to when they moved their arm themselves and similar in other ways to when the pulley moved their arm. Specifically, the hypnotic suggestion of the pulley moving their arm produced significant activation in brain areas associated with left-side movements (right sensorimotor cortex, premotor cortex, supplementary motor area and insula; bilateral basal ganglia and parietal operculum and left cerebellum), just like when they moved their arm themselves. However, the hypnotic suggestion also produced greater activation in brain areas that provide information about the outcome of movements (bilateral cerebellum and parietal cortex), just like when the pulley moved their arm. This suggests that feedback from these brain areas, which is normally quickly inhibited when the movement that generates it is predictable and voluntary, persisted in the case of the hypnotically generated movement giving rise to a subjective feeling of passivity and involuntariness.

Malingering in clinical conditions and hypnotic analogues

These findings of genuine changes in brain activity during hypnosis, which parallel clinical conditions, reinforce the view that hypnosis cannot be explained as mere faking or role playing. This issue has broader relevance, especially in clinical settings where malingering – intentionally feigning symptoms to avoid responsibilities or to gain compensation – is seen as a potential problem. For instance, conversion disorder patients, like hypnotized people, are often accused of deliberately faking by those who doubt the subjective reality of their experiences. Researchers addressed this issue directly in a neuroimaging (PET) study that extended work (described above) on conversion disorder paralysis. They scanned brain activations of high hypnotizable subjects who, half of the time, had hypnotically suggested paralysis of their left leg and, for the other half of the time, had both legs normal but were trying to convincingly fake left leg paralysis for a financial reward. Subjects did a good job. Even with repeated neurological examinations, an independent observer could not tell when subjects had hypnotic paralysis and when they were faking. But there were clear differences in brain activity. Hypnotic paralysis produced activation in brain areas not seen in faked paralysis (right orbitofrontal cortex, right cerebellum, left thalamus and left putamen), and faked paralysis produced activation in brain areas not seen in hypnotic paralysis (left ventrolateral prefrontal cortex and some right posterior cortical structures). These findings show that hypnotic paralysis is produced by different brain processes to faked paralysis and support subjects' claims that their hypnotic experiences are 'real.' Also, given the overlap in brain activations for hypnotic paralysis and conversion disorder paralysis, these findings suggest that clinical cases of paralysis are not simply the product of faking or malingering.

Conclusions: The Next Generations of Questions

For over 200 years interest in hypnosis has endured, perhaps because of its relatively unique status as a controllable, yet compelling, alteration in consciousness. In response to seemingly innocuous words from the hypnotist, hypnotized people experience major changes in themselves and the world, which they describe as both real and involuntary. Generations of hypnosis researchers have been incredibly productive in exploring and explaining these phenomena. In the twenty-first century hypnosis remains intrinsically interesting to psychologists, philosophers, cognitive scientists and (especially in recent years) neuroscientists. And hypnosis remains instrumentally useful in developing hypnotic analogues of psychological, neuropsychological, psychiatric, and neurological conditions.

There are clear and exciting directions for researchers and practitioners across the domain of hypnosis. These do not necessarily represent new questions about hypnosis, because many of our old questions are still catalysts for intrinsic and instrumental work. Rather, these directions represent

continuing questions best examined with fresh (often interdisciplinary) eyes and enabled by newly developed methods. In terms of individual differences, for instance, future work should profit from the 'building blocks' conceptualization of hypnotizability, which recognizes an underlying general capacity as well as identifies component abilities that predict certain kinds of hypnotic experiences; continue the more refined search for correlates of hypnotizability, including cognitive abilities such as attention (particularly given neuroscientific findings reported above); and map the developmental pathways of hypnotizable people, especially of different types of high hypnotizable subjects. In terms of explanations of hypnosis, future theorizing should continue connecting theories of hypnosis from inside the field with explanations and evidence of related phenomena and processes from outside the field (e.g., self-monitoring of action, and mental control and thought suppression, both of which have generated new, although still small, literatures within hypnosis); and consider the possibility that to understand hypnosis we might need different explanations for different subjects, at different times and/or for different hypnosis items. In terms of the neural basis of hypnosis, hypnosis researchers and neuroscientists should: collaborate to look for evidence of a trance state common to all hypnotic inductions, not just relaxation and in the absence of specific suggestions; consider how this hypnotic trance relates to other 'trance' states (e.g., meditation, mindfulness, daydreaming); determine what, if anything, the induction adds to hypnotic behavior and experience; and further study the effects of suggestions inside and outside hypnosis, especially the 'virtual reality' of suggested effects, which could then be harnessed in therapy and other applications. Finally, in terms of controllable hypnotic analogues, researchers and practitioners should continue working together to develop laboratory models of clinical conditions – there are many still to be explored – with a view to better understanding and improving treatments. Understanding can come simply from the act of trying to create an analogue. To recreate a particular clinical condition in the laboratory, the hypnotist needs to write effective hypnotic suggestions; getting the suggestions right demands a very clear understanding of what the patient is experiencing. When an hypnotic analogue fails, this highlights gaps in our understanding of the condition being modeled.

Hypnosis will endure as a topic of fascination at least for another 200 years because it is as involving and surprising to those observing or studying it as to those experiencing it. Hypnosis offers a great deal to the study of human behavior and experience: it provides a window on consciousness; it reveals important aspects of personality; it highlights the continuity and discontinuity of brain, behavior, and experience; and, perhaps most importantly, it is enormously useful.

See also: Implicit Social Cognition; Social Foundations of Consciousness.

Suggested Readings

Barnier AJ (2002) Post-hypnotic amnesia for autobiographical episodes: A laboratory model of functional amnesia? Psychological Science 13: 232–237.

Barnier AJ and McConkey KM (1996) Action and desire in posthypnotic responding. International Journal of Experimental and Clinical Hypnosis 44: 120–139.

Bowers KS (1976) Hypnosis for the Seriously Curious. New York: Norton.

Ellenberger HF (1970) The Discovery of the Unconscious: The History of Evolution of Dynamic Psychiatry. New York: Basic Books.

Hilgard JR (1965) Hypnotic Susceptibility. New York: Harcourt, Brace and World.

Hilgard JR (1973) The domain of hypnosis: With some comments on alternative paradigms. American Psychologist 23: 972–982.

International Journal of Clinical and Experimental Hypnosis (2003) Special issue: Hypnosis and the brain, vol. 51, Issues 2 and 3.

Jamieson GA (ed.) (2007) Hypnosis and Conscious States: The Cognitive Neuroscience Perspective. Oxford: Oxford University Press.

Kihlstrom JF (1980) Posthypnotic amnesia for recently learned material: Interactions with ''episodic'' and ''semantic'' memory. Cognitive Psychology 12: 227–251.

Nash MR and Barnier AJ (eds.) (2008) Oxford Handbook of Hypnosis: Theory, Research and Practice. Oxford: Oxford University Press.

Noble J and McConkey KM (1995) Hypnotic sex change: Creating and challenging a delusion in the laboratory. Journal of Abnormal Psychology 104: 69–74.

Oakley DA (2006) Hypnosis as a tool in research: Experimental pychopathology. Contemporary Hypnosis 23: 3–14.

Oakley DA and Halligan PW (in press) Psychophysiological foundations of hypnosis and suggestion. In: Rhue JW, Lynn SJ, and Kirsch I (eds.) Handbook of Clinical Hypnosis, 2nd edn. American Psychological Association.
Rainville P, Duncan GH, Price DD, Carrier B, and Bushnell MC (1997) Pain affect encoded in human anterior cingulated but not somatosensory cortex. Science 277: 968–971.
Woody EZ, Barnier AJ, and McConkey KM (2005) Multiple hypnotizabilities: Differentiating the building blocks of hypnotic response. Psychological Assessment 17: 200–211.

Biographical Sketch

Amanda Barnier is an associate professor and Australian Research Council (ARC) Australian research fellow in the Macquarie Centre for Cognitive Science, Macquarie University, Sydney, Australia. She completed her PhD in psychology (1996) at the University of New South Wales (UNSW), and her postdoctoral work at the University of California, Berkeley. Amanda then returned to Australia as an ARC postdoctoral fellow and later as an ARC Queen Elizabeth II Fellow. Amanda has earned an international reputation for her work on hypnosis and on memory, with research grants of more than $3 million. She has published 50+ articles/book chapters on hypnosis and memory and presented 50+ symposia, papers, posters, colloquia, or invited presentations at national and international conferences. Her current research focuses on hypnotic analogues of clinical delusions. Amanda's work has been recognized by Early Career Awards from the Australian Psychological Society (2001) and the American Psychological Association (Division 30, 2003) as well as the Young Tall Poppy Award from the Australian Institute of Political Science (2001) and the Australian Skeptics Eureka Prize for Critical Thinking (1997). Amanda has recently coedited (with Michael R. Nash, University of Tennessee, Knoxville) the Oxford Handbook of Hypnosis published by Oxford University Press in March 2008.

David Oakley is a chartered clinical psychologist, a professor emeritus in the Psychology Department at University College London and an honorary professor in the School of Psychology at Cardiff University. He is one of the directors of Hypnosis Unit UK, an independent training organization, and a former editor of the journal Contemporary Hypnosis. David is closely involved in research and teaching on the nature of hypnosis and its uses when applied to psychology, medicine, and dentistry. A current

interest is in the use of suggestion in hypnosis as a research tool, in particular to create experimental analogues for clinically relevant problems in pain, volition, motor control, and self-awareness and to explore their brain mechanisms with functional neuroimaging. David is a fellow of the Royal Society of Medicine and the British Psychological Society and has recently coedited two books: *Malingering and Illness Deception* with Peter Halligan and Chris Bass (published in 2003 by Oxford University Press) and *The Highly Hypnotizable Person* with Mike Heap and Richard Brown (published in 2004 by Brunner-Routledge).

Implicit Learning and Implicit Memory

A Cleeremans, Université Libre de Bruxelles, Bruxelles, Belgium

Glossary

Finite-state grammar – A finite-state grammar is a simple directed graph consisting of nodes connected by labeled arcs. Sequences of symbols can be generated by entering the grammar through a 'begin' node, and by moving from node to node until an 'End' node is reached. Each transition between a node and the next produces the label associated with the arc linking the two nodes. Finite-state grammars have been used both in the context of sequence-learning studies and in the context of artificial grammar learning studies.

Introduction

Implicit learning and implicit memory both refer to the nonconscious effects that prior information processing may exert on subsequent behavior. In general, memory for previous events can be expressed explicitly, as a conscious recollection, or implicitly, as automatic, unconscious influences on behavior. Thus for instance, when one recalls a particular event such as a cocktail last week, one typically consciously reexperiences many details of the event: The moment when Charles dropped a glass of wine upon unexpectedly meeting his ex-wife; the vibrant crimson color of the dress that the dean was wearing that evening; the political overtones of the president's speech; the awful taste of a spoiled shrimp. These many details are encoded in such a way that recalling them elicits not only a reexperiencing of the original context in which they were first experienced but also the distinct conscious experience that one is now recalling a particular episode of our own past.

This form of explicit, episodic memory contrasts strongly with implicit memory. For instance, you may find yourself avoiding seafood for a month or so; or you may find yourself drawn to crimson-colored items when choosing new furniture. Crucially, such changes in behavior may occur without conscious recollection of the original episode that installed them and are thus properly described as a form of nonconscious, implicit memory. Likewise, all literate adults are able to decide whether a string of letters constitutes a word without having any conscious recollection of the circumstances in which this word has first been learned. The same goes for much of what we know, suggesting that implicit knowledge is pervasive and forms the foundation of semantic and procedural memory.

The distinction between implicit and explicit learning taps onto the same difference as the difference between implicit and explicit memory, but additionally refers to situations where one becomes sensitive not only to particular instances but rather to whole structured ensembles of relationships between stimuli. Thus for instance, one can consciously learn about the features that define particular species of birds, – say, gulls. Gulls are medium-to-large birds, typically gray or white in color, and found at sea or in coastal areas. Gulls wail or squawk, have long, sometimes brightly colored bills, webbed feet, and so on. However, this conscious, explicit knowledge about the shared features of gulls as opposed to, say, terns, does not seem to be at play when an expert birdwatcher quickly identifies a Herring Gull amongst a group of seabirds. When prompted, an amateur ornithologist may come up with a list of defining features, but most likely he will experience difficulty verbalizing his decision criteria, claiming instead that it is just obvious that this bird is a gull. The knowledge involved in reaching a classification decision in this instance depends on a vast network of memorized features and relationships between features that may or may not be readily available for conscious report: Over months and years of experience, the expert has accumulated

intuitive knowledge that appears to be dissociated both from the original episodes in which it was first learned and also from his ability to muster explicit reports about it.

The same observation applies to a wide range of cognitive skills. Indeed, everyday experience suggests that we often seem to know more than we can tell. Riding a bicycle, playing tennis or driving a car, for instance, all involve mastering complex sets of motor skills, yet we are at a loss when it comes to explaining exactly how we perform such physical feats. These dissociations between our ability to report on cognitive processes and the behaviors that involve these processes are not limited to the motor skills of athletes or to the honed perceptual skills of birdwatchers, but extend to higher-level cognition as well. Most native speakers of a language are unable to articulate the grammatical rules they nevertheless follow when uttering expressions of the language. Likewise, expertise in domains such as medical diagnosis or chess, as well as social or aesthetic judgments, involves intuitive knowledge that one seems to have little introspective access to. Thus, while it is commonly accepted and hence unsurprising that we have little access to the cognitive processes involved in mental operations, it also appears that knowledge itself can remain inaccessible to report yet be causally efficacious.

We also often seem to tell more than we can know. In a classic article that appeared in 1977, social psychologists Nisbett and Wilson reported on many experimental demonstrations of the fact that accounts of our own behavior frequently reflect reconstructive and interpretative processes rather than genuine introspection. In one such study, Nisbett and Wilson asked patrons of a department store to choose the best quality item among four identical pairs of nylon stockings in the context of what subjects believed was a consumer survey. There was a strong position effect: Items located on the right, which were inspected last, were chosen much more often than items in other positions. Yet, when asked to motivate their choice, not a single subject mentioned position as a relevant factor, and many denied that it might have had an effect when asked directly about it. More generally, while it is generally agreed that cognitive processes are not in and of themselves open to any sort of introspection, Nisbett and Wilson further claimed that we can sometimes be ''(a) unaware of the existence of a stimulus that importantly influenced a response, (b) unaware of the existence of the response, and (c) unaware that the stimulus has affected the response'' (p. 231). This hints at important and subtle distinctions between awareness during perception, encoding, and retrieval of information during information processing episodes.

Paradigms and Theories of Implicit Memory and Implicit Learning

Learning and memory, in addition to perception (e.g., subliminal perception), are the two domains in which dissociations between ability to report and ability to use knowledge have been most extensively explored in the laboratory. Each is examined in turn below. Because of the considerable challenge that assessing awareness represents, both domains also pose substantial methodological, conceptual, and theoretical puzzles. These will be approached in the section titled 'Definitional, methodological, and conceptual challenges.'

Implicit Memory

The study of memory in general has a long history that extends all the way back to the ancients, with for instance Plato's metaphor (and rejection) of memory as a wax tablet, blank at birth and upon which experiences leave marks that affect its shape. Specific discussion of nonconscious forms of memory can be first traced to Descartes. Modern investigations of implicit memory begin with observations by Korsakoff and Claparède, around the early twentieth century. Both worked with amnesic patients, who, despite being severely impaired in their ability to consciously recollect previous experiences, nevertheless showed continued sensitivity to such prior experiences. Perhaps the most famous example of such early clinical work on implicit memory is Claparède's observation that a patient suffering from Korsakoff's syndrome whom he had covertly pricked the hand of using a pin reflexively withdrew her hand when Claparède extended his towards her on the next day. When asked to motivate her

refusal, the patient indicated that perhaps there was a pin hidden in Claparède's hand, though she failed to consciously recall her recent painful experience. The patient thus exhibited a simple form of operant conditioning without conscious awareness. Claparède subsequently tested this patient using Ebbinghaus's 'savings' method, by which one measures, on different occasions, the number of words that can be recalled after different number of repeated readings of these words. Despite failing to remember the repeated readings, the patient exhibited substantial savings over the course of several testing sessions, requiring fewer and fewer repetitions to memorize the list. This suggests a clear dissociation between the ability to consciously recollect a prior episode and the potency of such prior episodes in influencing subsequent behavior.

Recent research with amnesic participants has tended to be more focused on skill acquisition. The famous patient HM, for instance, was shown by Brenda Milner to exhibit preserved ability to learn a novel skill – mirror tracing, in which one has to draw a figure looking only at the reflection of one's hand in a mirror. Other studies with other patients have yielded similar results, indicating that amnesic patients exhibit preserved ability to learn about novel skills involving procedural memory in the absence of a corresponding ability to consciously recall the circumstances in which the novel skills were learned.

With normal participants, some studies have also explored the extent to which patients undergoing general anesthesia are able to learn novel associations, and one study has even demonstrated learning in fetuses, which are presumably unaware and definitely unable to report on their experiences.

The bulk of recent research on implicit memory has, however, been focused on priming. Most nominal implicit memory paradigms share the following elements:

1. A study phase in which participants are exposed to and asked to process an ensemble of (typically unstructured) stimuli, such as a list of unrelated words or pictures. Instructions may or may not require participants to intentionally encode the material, that is, participants may be asked to directly memorize the material, but it may also be the case that participants are merely required to process the material in some way – deciding whether each picture depicts an animal or not, deciding whether each word is a noun or not, and so on.
2. A test phase during which participants perform the same or a different task, thereby providing a measure P of the extent to which they exhibit sensitivity to the stimuli they have been exposed to in the study phase. Importantly, the test does not make direct reference to the study phase. Relevant tests include, for instance, stem or fragment completion, whereby participants are asked to complete fragments of words with the first word that comes to mind (e.g., CONSCIO————————?). Assessing the probability that experimental rather than control participants will complete a stem with a previously seen word rather than with an unseen word thus provides a measure of priming.
3. An awareness test C that directly assesses the extent to which participants are able to consciously remember the items memorized in the study phase. This may consist, for instance, of a recall or recognition task in which participants are simply asked to report as many items from the study list as possible (recall) or to decide whether an item is familiar or not (recognition).

P may thus be construed as an indirect measure of participants' knowledge of the study material, for it fails to make direct reference to the study episode itself. This measure, crucially, neither refers to nor requires knowledge of the material processed in the study phase. C, on the other hand, constitutes a direct measure of participants' conscious knowledge of the study material, for it specifically asks participants to intentionally recall the studied material. Implicit memory, according to a simple quantitative dissociation logic, may thus be inferred whenever P reveals more knowledge than C. Despite the fact that this reasoning is not without problems, as will be examined below, the vast majority of implicit memory studies have relied on this design.

Numerous studies have thus demonstrated dissociations between implicit (P) and explicit (C) memory in normal participants, essentially by documenting the selective influence that certain

variables exert on the expression of previously learned novel information. For instance, depth of processing (i.e., in the case of word stimuli, the extent to which semantic analysis is required) during encoding leaves implicit memory (e.g., priming) unaffected but strongly influences explicit memory, with deeper levels of processing promoting better conscious recall. Other important factors that have been extensively explored include attention and intentionality during study, transfer between different modalities (e.g., the study phase includes auditorily presented words, whereas the test concerns written words), or manipulations of the context. Other studies, mostly in social psychology, have also explored different dependent variables. For instance, the well-known 'mere exposure' effect refers to changes in the preference that participants express towards unfamiliar and familiar stimuli. Typically, participants prefer familiar stimuli even when unable to consciously recall prior exposure.

This vast array of experimental findings has received different theoretical interpretations. Among these, one may distinguish between activation, processing, and multiple-memory systems accounts. Activation views simply characterize priming as resulting from increased activation of the relevant items through prior exposure. Processing views ascribe differences between implicit and explicit memory to differences in the cognitive processes engaged during encoding and retrieval, thus focusing on the influence and differential effects of incidental versus intentional encoding; on the differences between voluntary and involuntary recall; and so on. Multiple memory-systems views are essentially motivated by incontrovertible findings that different forms of memory are subtended by separate neural pathways and structures in the brain. Whether such systems usefully differentiate along an implicit versus explicit continuum remains very much open for debate, however.

Implicit Learning

Arthur Reber, in a classic series of studies conducted in 1965, first suggested that learning might be 'implicit,' to the extent that people appear to be able to learn new information without intending to do so and in such a way that the resulting knowledge is difficult to express. Implicit learning is a complex, multifaceted phenomenon that has received as many as 15 different definitions over the years. Three experimental paradigms have been explored most extensively: Artificial grammar learning, dynamic system control, and sequence learning.

In Reber's seminal study of artificial grammar learning, subjects were asked to memorize a set of meaningless letter strings generated by a simple set of rules embodied in a finite-state grammar (Figure 1).

After this memorization phase, they were told that the strings followed the rules of a grammar, and were asked to classify novel strings as grammatical or not. In this experiment and in many subsequent replications, subjects were able to perform this classification task better than chance would predict, despite remaining unable to describe the rules of the grammar in verbal reports. This dissociation between classification performance and verbal report is the finding that prompted Reber to describe learning as implicit, for subjects appeared sensitive to and could apply knowledge (the rules of the grammar) that they remained unable to describe and had had not intention to learn.

In a series of experiments that attracted renewed interest in implicit learning, Donald Broadbent

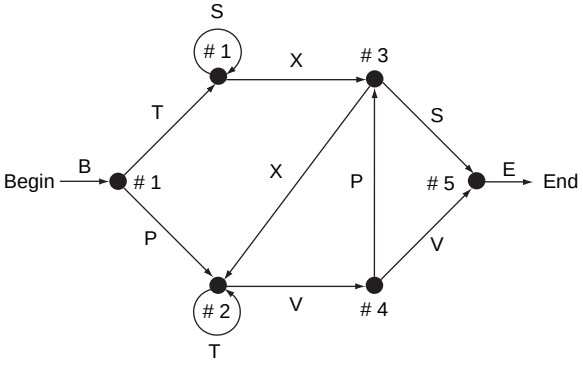


Figure 1 A finite-state grammar is a simple directed graph consisting of nodes connected by labeled arcs. Sequences of symbols can be generated by entering the grammar through a 'begin' node, and by moving from node to node until an 'End' node is reached. Each transition between a node and the next produces the label associated with the arc linking the two nodes. Concatenating the symbols together produces strings of symbols, in this case, letters of the alphabet. Finite-state grammars have been used both in the context of sequence-learning studies and in the context of Artificial Grammar Learning studies.

showed that success in learning how to control a simulated system (e.g., in one set of experiments, a system described to subjects as a 'sugar factory') so as to make it reach certain goal states was independent from ability to answer questions about the principles governing the subject's inputs and the system's output: Practice selectively influenced the ability to control the system, whereas verbal explanations about how the system works selectively influenced the ability to answer questions.

Today, a third paradigm – sequence learning – has become dominant in the investigation of implicit learning. Nissen and Bullemer first reported on sequence learning in a landmark 1987 article. The paradigm can broadly be defined as consisting of several types of situations that share the common features of presenting participants with a speeded, typically visual, task during which (1) they have to respond to the location of a target stimulus that may appear at one of several possible locations on a computer screen on each trial, and (2) in which the series of locations follows a regularity that is not revealed to participants. Although in most studies the series of locations is structured to follow a fixed and repeating sequence, some authors have used sequential material generated on the basis of a complex set of rules from which one can produce several different alternative deterministic sequences, or based on the output of probabilistic and noisy finite-state grammars similar to that illustrated in Figure 1.

In versions of this task where a simple sequence repeats, sequence-specific facilitation can be assessed by substituting the repeating sequence by a different one at some point during training. Any observed reaction time cost observed on this transfer block can be interpreted as reflecting the influence of sequence-specific knowledge acquired over the blocks of training preceding transfer. When a probabilistic sequence is used, a different method may be used to assess performance: One simply needs to compare reaction times on predictable and unpredictable stimuli of the sequence. Sequence learning is demonstrated when predictable stimuli, that is, stimuli that can be anticipated based on the temporal context set by previous stimuli, elicit faster reaction times than unpredictable stimuli.

In general, the results obtained in each of these paradigms have uniformly shown that the participants' performance with the serial reaction time (SRT) task expresses sensitivity to the sequential constraints regardless of the nature of the generation rules, and that this sensitivity is not necessarily accompanied by conscious awareness of the relevant sequential constraints when assessed by a comparable direct measure. Numerous subsequent studies of this effect have indicated that subjects can learn about complex sequential relationships despite remaining unable to fully elicit this knowledge in corresponding direct, explicit tasks.

Additional implicit-learning paradigms include probability learning, hidden covariation detection, acquisition of invariant characteristics, or contextual cueing. In the latter, participants are presented with a large visual array of randomly arranged distractors (e.g., multiple instances of the letter 'L' in various orientations), and are instructed to indicate the orientation of a single different stimulus (e.g., the letter 'T' oriented to the left or to the right). Unknown to participants, some of the displays repeat, such that if participants learn something about the relationship between particular (random) arrangements of the distractors and the location of the target, one may expect faster reaction times reporting on the orientation of the target. This is indeed what is typically found, despite the fact that participants claim not to have perceived any relationship between distractors and target location, and despite the fact that they perform at chance when asked to decide if a particular display is familiar or not.

The central difference between implicit memory and implicit-learning paradigms is thus that the latter involve sensitivity not to specific memory traces but rather to what could be dubbed 'family resemblances' – the ensemble of rule-governed regularities shared by many different instances. Implicit learning thus concerns a different level of abstraction than implicit memory.

Just as for implicit memory, these different paradigms all share a basic design, which typically consists of the following elements:

1. a learning phase during which participants are exposed to some complex rule-governed stimulus environment under incidental learning conditions (i.e., the true purpose of the experiment is not revealed to participants);

2. a testing phase during which participants perform the same or a different task, thereby providing a measure P of the extent to which they express sensitivity to the regularities they have been exposed to in the learning phase; and
3. an awareness test C that assesses the extent to which participants are aware of the learned regularities.

In implicit-learning paradigms, P and C can take many different forms, partly because the paradigms themselves are so different from each other. Typical measures of performance include the ability to decide whether a string is grammatical or not in artificial grammar learning experiments, ability to control the system in system control tasks, and sequence-specific reaction time facilitation in sequence-learning tasks.

Measures of awareness in implicit-learning paradigms have elicited considerable controversy, as will be described in the section titled 'Definitional, Methodological and Conceptual Challenges.' In contrast to the relatively restricted range of such measures in implicit memory research, there is a large array of possible measures of awareness in implicit-learning research. The simplest of these methods consists, rather naturally, of asking participants to verbally report on whether they are aware of the relevant regularities. For instance, one can ask participants in an artificial grammar learning experiment whether they were aware of the fact that all the training strings had been generated based on a grammar. In a sequence-learning experiment, one can likewise ask participants to indicate whether they have noticed that stimuli did not follow each other randomly, but rather in a specific sequence. The limitations of such reports, however, are obvious: Not only are verbal reports rather insensitive measures (e.g., participants may fail to report on knowledge held with low confidence) but they may also completely miss the mark in that the knowledge revealed by verbal reports may simply fail to be necessary for successful performance, so rendering dissociation findings moot. For instance, participants can perform above chance in an artificial grammar learning situation without any knowledge of the grammar itself, just as adult speakers of a natural language may be perfectly capable of deciding whether a proposition is grammatically correct without any formal knowledge of grammar. Hence, finding that a participant is unable to verbalize knowledge of the rules of an artificial grammar does not tell us much about the extent to which the knowledge actually used during the task was unconscious or not.

Experimental findings in these different paradigms continue to elicit controversy insofar as their theoretical interpretation is concerned. To wit, there are many different things that a participant engaged in an artificial grammar learning experiment may learn about the stimulus material. Some possibilities are illustrated in Figure 2. The finding that participants perform above chance in classifying new letter strings as grammatical or not may thus, a priori, stem from any of the following

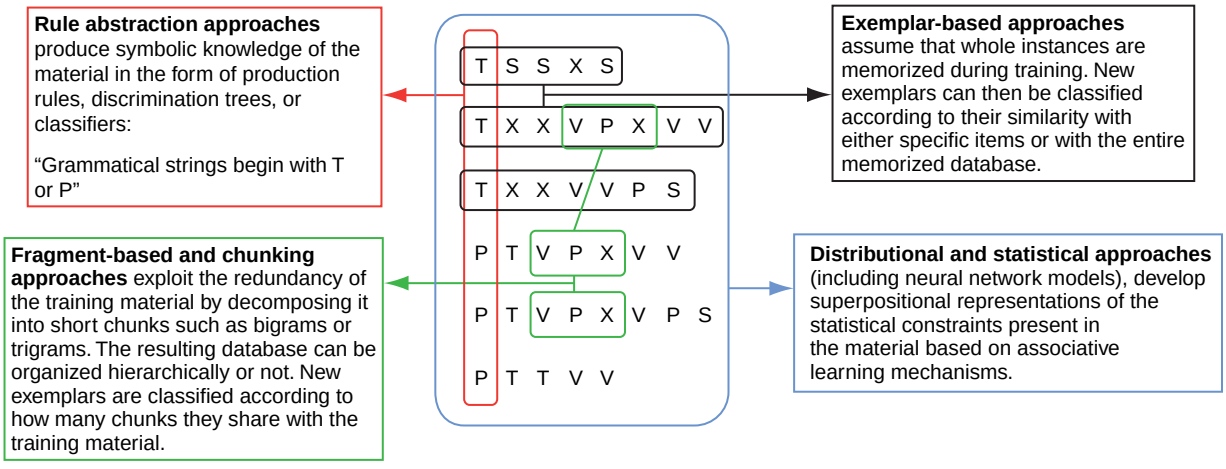


Figure 2 A representation of different theories of artificial grammar learning (see text for details).

possibilities: (1) they may have learned something about the rules of the grammar from which the strings have been generated, (2) they may have memorized frequently occurring fragments or chunks of the strings, (3) they may have memorized entire strings from the learning phase, or (4) they may have become sensitive to the statistical structure of the entire set of training exemplars. Similar discussions have taken place concerning the other paradigms of implicit learning, and the issue is largely unsolved today.

Thus, while there is general agreement that people learn more than mere statistics in implicit-learning experiments, there is continuing debate about just how abstract the acquired knowledge is. Even though early accounts of implicit learning as involving unconscious, incidental acquisition of formal rule systems have now been toned down considerably, there is no definitive demonstration to the contrary either. The major challenge is that it is very difficult to define exactly what a 'rule' is, so that debates about the extent to which one may learn about rules unconsciously overlaps with debates about how abstract, rule-based knowledge should be defined and detected. Partly as a result of such debates, new theoretical perspectives about implicit learning have emerged over the past decade.

These new perspectives have been largely motivated by methodological concerns about both the purported unconscious and abstract character of knowledge acquired in typical implicit-learning situations. Thus, many recent studies have in fact reported associations between performance and awareness. Likewise, it now appears that simple associative learning or chunking mechanisms, rather than rule abstraction processes, are largely sufficient to account for performance in all three main paradigms.

Such findings have prompted many authors to question the very existence of implicit learning. For instance, Shanks and St. John concluded their 1994 critical review article with the statement that 'Human learning is systematically accompanied by awareness,' and suggested that implicit and explicit learning should instead be distinguished based on their information-processing characteristics. Other critics have likewise suggested that implicit learning is just ordinary learning without awareness of the effects of this learning, and that research should therefore be focused not on awareness but on other features such as the role of intention during learning or the congruence between task demands during learning and the subsequent use of knowledge.

Thus, in the space of a few years, our general perspective on implicit learning has changed from one that assumes the existence of some potentially mysterious processes of passive, automatic, and unconscious acquisition of abstract and tacit knowledge to one that aims to highlight the fact that implicit learning is merely a side effect of ongoing processing, and that awareness systematically accompanies learning. The continuing debate illustrates the challenges involved in assessing awareness. These are reviewed in the section titled 'Definitional, methodological and conceptual challenges.'

Definitional, Methodological, and Conceptual Challenges

Because there is no accepted operational definition of what it means for an agent to be conscious of something, complex measurement challenges arise in the study of the relationships between conscious and unconscious cognition. Three challenges may be distinguished: A definitional challenge (how do we define awareness?), a methodological challenge (how do we devise an appropriate measure of awareness?), and a conceptual challenge (how do we interpret dissociation findings?). Each is outlined in turn below.

First, consciousness is not a single process or phenomenon, but rather encompasses many dimensions of experience. A first important challenge thus arises in delineating which aspects of consciousness count when assessing whether a subject is aware or not of a particular piece of information: awareness of the presence or absence of a stimulus, conscious memory for a specific previous processing episode, awareness of one's intention to use some information, or awareness that one's behavior is influenced by some previous processing episode. Different aspects of conscious processing are engaged by different paradigms. In subliminal perception studies, for instance, one is concerned with determining whether stimuli

that have not been consciously encoded can influence subsequent responses. In contrast, implicit memory research has been more focused on retrieval processes, that is, on the unintentional, automatic effects that previously consciously perceived stimuli can exert on subsequent decisions. In studies of implicit learning, it is the relationships between ensembles of consciously processed stimuli that remain purportedly unconscious. These subtle differences in which specific aspects of the situation are available to awareness illustrate the need to carefully distinguish between awareness during encoding and awareness during retrieval of information. Further, both encoding and retrieval can concern either individual stimuli or relationships between sets of stimuli, and both can either be intentional or not.

A second important challenge is to devise an appropriate measure of awareness. As described above, most experimental paradigms dedicated to exploring the relationships between conscious and unconscious processing have relied on a simple dissociation logic aimed at comparing the sensitivity of two different measures to some relevant information: A measure C of subjects' awareness of the information and a measure P of behavioral sensitivity to the same information in the context of some task. As discussed above, unconscious processing, according to the simple dissociation logic, is then demonstrated whenever P exhibits sensitivity to some information in the absence of correlated sensitivity in C. There are several potential pitfalls with the simple dissociation logic, however. First, the measures C and P cannot typically be obtained concurrently. This 'retrospective assessment' problem entails that finding that C fails to be sensitive to the relevant information need not necessarily imply that information was processed unconsciously during encoding, but that, for instance, it might have been forgotten before retrieval. A second issue is to ensure that the information revealed through C is indeed relevant to perform the task. As Shanks and St. John, in a severe critique of the field published in 1984, have suggested, many studies of implicit learning have failed to respect this information criterion. For instance, successful classification in an artificial grammar learning task need not necessarily be based on knowledge of the rules of the grammar, but can instead involve knowledge of the similarity relationships between training and test items. Subjects asked about the rules of the grammar would then understandably fail to offer relevant explicit knowledge. A third issue is to ensure that C and P respect the sensitivity criterion, that is, that both be equally sensitive to the relevant information.

At first sight, and despite their known limitations, verbal reports and other subjective measures such as confidence ratings, or as proposed recently, wagering, would appear to offer the most direct way through which to assess the contents of subjective experience, particularly when they can be obtained concurrently with decisions. According to this framework, learning is implicit when subjects who perform above chance in a direct test lack metaknowledge, either because they believe they are guessing (the guessing criterion) or because their accuracy is unrelated to their confidence judgments (the zero-correlation criterion). Thus for instance, if a participant in an artificial grammar learning experiment performs better than chance would predict but claims to be guessing on each correct decision, this is prima facie evidence for the fact that unconscious knowledge is at play. Such subjective measures are nevertheless often difficult to operationalize in a sufficiently controlled manner. For instance, people might simply refrain from reporting on knowledge held with low confidence or might offer reports that are essentially reconstructive in nature, as Nisbett and Wilson's experiments indicate.

For this reason, many authors have advocated using the so-called objective measures of awareness. Objective measures of awareness include forced-choice tests such as recognition, presence–absence decisions, or identification. Objective measures, however, may be criticized on other grounds, namely, the fact that they may themselves be influenced by unconscious knowledge.

In other words, it might prove elusive to hope to be able to obtain measures of awareness that are simultaneously exclusive and exhaustive with respect to knowledge held consciously. Thus, finding null sensitivity in C, as required by the dissociation paradigms for unconscious processing to be demonstrated, might simply be impossible because no such absolute measure exists. Even a test such as recognition, for instance, is likely to reflect a mixture of conscious and unconscious influences,

as it is perfectly possible to decide that an item is 'old' based not just on conscious, episodic memory but also on familiarity. A significant implication of this conclusion is that, at least with normal participants, it makes little sense to assume that conditions exist where awareness can simply be 'turned off.' Much of the ongoing debate about the existence of subliminal perception can be attributed to a failure to recognize the limitations of the dissociation logic.

It might therefore instead be more plausible to assume that any task is always sensitive to both conscious and unconscious influences. In other words, no task is process-pure. Two methodological approaches that specifically attempt to overcome the conceptual limitations of the dissociation logic have been developed. The first was introduced by Reingold and Merikle, who suggested that the search for absolute measures of awareness should simply be abandoned in favor of approaches that seek to compare the sensitivity of direct measures and indirect measures of some discrimination. Direct measures involve tasks in which the instructions make explicit reference to the relevant discrimination, and include objective measures such a recognition or recall. In contrast, indirect measures, such as stem completion in implicit memory tasks, make no reference to the relevant discrimination. By assumption, direct measures should exhibit greater or equal sensitivity than indirect measures to consciously held task-relevant information, for subjects should be expected to be more successful in using conscious information when instructed to do so than when not. Hence, demonstrating that an indirect task is more sensitive to some information than a comparable direct task can only be interpreted as indicating unconscious influences on performance. Successful examples of application of this reasoning can be found in both the implicit memory and implicit-learning literatures.

The second approach – Larry Jacoby's 'Process Dissociation Procedure' — constitutes one of the most significant advances in the study of differences between implicit and explicit memory. It is based on the argument that, just as direct measures can be contaminated by unconscious influences, indirect measures can likewise be contaminated by conscious influences: Particular tasks can simply not be identified with particular underlying processes. The process dissociation procedure thus aims to tease apart the relative contributions of conscious and unconscious influences on performance. To do so, two conditions are compared in which conscious and unconscious influences either both contribute to performance improvement or act against each other. For instance, subjects might be asked to memorize a list of words and then, after some delay, to perform a stem completion task in which word stems are to be completed either so as to form one of the words memorized earlier (the inclusion condition) or so as to form a different word (the exclusion condition). Inclusion and exclusion thus only differ with respect to instructions. In inclusion, participants can either respond based on conscious, episodic memory of the corresponding word, or failing conscious recollection, they can respond based on familiarity or intuition. The point is that in both cases, inclusion performance will tend to improve. This stands in contrast with the exclusion task, in which conscious and unconscious influences work in opposition to each other. Indeed, the only way for participants to produce the requested different completion is to consciously recall the memorized word and to exclude it. If the stems nevertheless tend to be completed by memorized words under exclusion instructions ('exclusion failure'), then one can only conclude that memory for these words was implicit, since if subjects had been able to consciously recollect them, they would have avoided using them to complete the stems. Numerous experiments have now been designed using the process dissociation procedure. A recent study, for instance, has applied the method to sequence learning, demonstrating exclusion failure in generation after training with short response-to-stimulus intervals. While the 'Process Dissociation Procedure' has elicited controversy because of the competing theoretical models that may be used to compute quantitative estimates of the relative influence of conscious and unconscious processes on performance in a given task, the method remains useful even when limited to comparisons between inclusion and exclusion performance.

The third challenge is a conceptual one. Even when dissociations are found that fulfill the relevant methodological criteria, when is it the case that we

can conclude that separable, distinct, independent systems are involved in subtending performance? In an often-ignored landmark article published in 1988, Dunn and Kirsner pointed out that even crossed double dissociations between two tasks do not necessarily indicate the involvement of separable, independent processes, for it is logically compatible with the operation of a single system. The only dissociation pattern that is actually incompatible with single-system accounts is called reverse association, whereby one observes a dissociation between the two variables of interest under certain experimental conditions, and an association between the same two variables under a different set of experimental conditions. Very few studies currently fulfill this criterion.

Further, many authors have described nonmodular architectures that can nevertheless produce double dissociations. Simulation studies using connectionist networks, for instance, have clearly indicated not only that a single system is capable of producing dissociations between different variables but also that single parameter changes (when applied to models that incorporate inherent variability, which is undoubtedly the case for the brain) can sometimes result in one pattern of dissociation and at other times in the opposite pattern of dissociation. Further, the dissociations often exhibit a gradual character. The moral of both Dunn and Kirsner's theoretical analysis and of the simulation studies is that interpreting a double dissociation as reflecting architectural specialization requires extreme caution and is often unwarranted, even with neuropsychological patients. Rather, it appears that many such dissociations can instead merely reflect functional specialization (functional modularity). The importance of this point cannot be understated in the context of implicit memory and implicit-learning research.

Theories of the Psychological Unconscious

Despite the methodological challenges noted above, the numerous findings in the domains of implicit learning and implicit memory all suggest that unconscious influences on behavior are pervasive. This raises the question of how to best characterize the relationships between conscious and unconscious processes, and in particular whether one should consider that mental representations can be unconscious. The idea that mental life includes both conscious and unconscious events has been expressed most clearly by Freud, whose psychoanalytical theory has profoundly influenced both scientific thought and public conceptions of the mind. Even though one can find earlier characterizations of cognition in which computations that are not accompanied by awareness contribute to conscious decisions in Leibniz and other authors such as de Biran, most thinkers up until Freud, including René Descartes and, arguably, William James, considered that mental life consists exclusively of conscious events. This position is still endorsed by some authors today, even though there is wide consensus about the idea that cognitive processing is largely unconscious.

Freud's characterization of the 'dynamic unconscious,' however, makes very specific assumptions – specifically that there exist unconscious mental representations and that these representations can reflect semantic and affective dimensions of processing. Further, the unconscious, as Freud depicted it, is as dynamic and causally efficacious as the conscious – it is a mental bubbling cauldron of sorts, replete with repressed instincts, thoughts and feelings vying to get access to awareness. Few thinkers would endorse Freud's characterization of the unconscious in full today – often choosing instead to deny its existence altogether or considering that only shallow aspects of processing can take place unconsciously. In particular, while popular belief has often tended to ascribe powerful abilities to the unconscious, empirical exploration of the level of analysis at which processing occurs in the absence of consciousness has failed to offer convincing demonstrations that unconscious processing can be as flexible or as deep as conscious processing. Rather, the evidence suggests that unconscious processing can bias conscious processing in a way that reflects strong or habitual responses.

An important issue in this context is to determine the extent to which conscious and unconscious processing depend on separable neural systems, and what their relationship should be. Contemporary theories of consciousness make widely different assumptions about its underlying

mechanisms. Such theories may be organized along two dimensions defined by (1) whether the theory assumes that consciousness depends on the involvement of certain processes or rather on the properties of certain mental representations, and by (2) whether the theory assumes that consciousness depends on the involvement of specific modules or whether it can occur anywhere in the brain. Computational models of information processing play a significant role in fostering the development of novel theories of consciousness. In neural network – or connectionist – models, for instance, task-relevant knowledge is embedded in the same structures that support processing itself. In contrast to traditional information processing frameworks, knowledge is thus implicit in such models to the extent that it cannot be separated from the mechanisms that subserve processing.

If consciousness does not depend on neural systems specifically dedicated to subserve subjective experience but should instead be viewed as an emerging property of the processing conducted by many different regions of the brain, one might expect to find indications that consciousness fails to be a unified, all-or-none phenomenon. Congruently, the British psychologist Anthony Marcel described a relevant series of experiments probing the unity of consciousness in normal subjects. In these experiments, people were asked to detect changes in the luminance of very faint stimuli in three different ways: by blinking, by pressing on a button, or by a verbal response ('yes'). When these responses had to be produced simultaneously, they often tended to be dissociated, with subjects answering positively, for instance, through a blinking response, but negatively through the other modalities. Another experiment indicated that the probability of correct guessing varied significantly across modalities, thus suggesting that they each afforded varying degrees of availability to conscious awareness.

Treating Consciousness as a Variable: Qualitative Differences

Rather than pursuing the elusive goal of demonstrating unconscious cognition, a more fruitful approach to exploring the relationships between conscious and unconscious cognition thus consists of treating consciousness as a variable, that is, of exploring the functional, neural, and computational differences that exist between tasks performed with awareness and without. This leads to the design of experiments in which qualitative differences between cognition with and without awareness can be established in carefully controlled conditions. In this respect, the 'contrastive approach' advocated by Bernard Baars offers great promise in helping solve the methodological challenges described above. Numerous qualitative differences between conscious and unconscious processes have been reported. For instance, on the basis of behavioral methods, Arthur Reber and others have suggested that implicit cognition involves processes that are more resistant to insult or injury, that tend to be less sensitive to individual differences or affective state, and less sensitive to attentional demands. Recent advances in brain imaging methods make it possible to go beyond purely behavioral methods and to actually integrate different online measures of awareness and of brain activity when attempting to contrast conscious and unconscious cognition.

Numerous studies combining various online measures of brain activity have likewise attempted to identify which regions of the brain are most involved in processes such as conscious recollection, intentional retrieval, or nonconscious influences in different memory tasks. While it is too early to draw definite conclusions from such studies, it appears, for instance, that anterior prefrontal regions of the brain appear to be most clearly associated with processes of effortful, conscious retrieval. This and similar findings should help us further develop theories of consciousness that have a firm rooting in our knowledge of the brain.

Conclusions

The study of implicit learning and memory constitute a privileged gateway to our understanding of the differences between conscious and unconscious processing. Conscious and unconscious processing differ on several dimensions, including

depth and specificity of processing. Unconscious processing tends to reflect habitual or strong responses, but, as implicit-learning research clearly demonstrates, is not limited to familiar stimuli. Research in this domain faces many challenges, ranging from definitional issues to often intricate measurement and interpretational issues. Contemporary views of implicit learning and memory recognize that both involve continua rather than dichotomies; as well as processing-oriented rather than architectural distinctions. Functional brain imaging methods, when used together with sufficiently sensitive behavioral methods that combine first-person and third-person data, offer the promise to elucidate the relationships between conscious and unconscious processes in cognition.

See also: Inner Speech and Consciousness; Meta-Awareness; Perception: Subliminal and Implicit.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. Cambridge: Cambridge University Press.
Berry DC and Dienes Z (1993) Implicit Learning: Theoretical and Empirical Issues. Hove, UK: Erlbaum.
Cleeremans A (1993) Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing. Cambridge, MA: MIT Press.
Cleeremans A, Destrebecqz A, and Boyer M (1998) Implicit learning: News from the front. Trends in Cognitive Sciences 2: 406–416.
French RM and Cleeremans A (2002) Implicit Learning and Consciousness: An Empirical, Computational, and Philosophical Consensus in the Making. Hove: Psychology Press.
Jacoby LL (1991) A process dissociation framework: Separating automatic from intentional uses of memory. Journal of Memory and Language 30: 513–541.
Marcel AJ (1993) Slippage in the unity of consciousness. In: Bock GR and Marsh J (eds.) Experimental and Theoretical Studies in Consciousness (Ciba Foundation Symposium 174), pp. 168–186. Chichester: John Wiley & Sons.
Nisbett RE and Wilson TD (1977) Telling more than we can know: Verbal reports on mental processes. Psychological Review 84: 231–259.
Perruchet P and Vinter A (2002) The self-organizing consciousness. Behavioural and Brain Sciences 25: 297–330.
Reber AS (1993) Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious. Oxford, UK: Oxford University Press.
Reingold EM and Merikle PM (1988) Using direct and indirect measures to study perception without awareness. Perception and Psychophysics 44: 563–575.
Schacter DL (1987) Implicit memory: History and current status. Journal of Experimental Psychology: Learning, Memory, and Cognition 13: 501–518.
Shanks DR and St. John MF (1994) Characteristics of dissociable human learning systems. Behavioral and Brain Sciences 17: 367–447.
Stadler MA and French PA (1998) Handbook of Implicit Learning. Thousand Oaks, CA: Sage Publications.
Weiskrantz L (1997) Consciousness Lost and Found: A Neuropsychological Exploration. Oxford: Oxford University Press.

Biographical Sketch

Axel Cleeremans is a research director with the Fonds de la Recherche Scientifique (F.R.S.-FNRS) and a professor of cognitive psychology at the Université Libre de Bruxelles, where he heads the Consciousness, Cognition and Computation (CO3) Group. Trained under the supervision of Jay McClelland at Carnegie Mellon University where he obtained his PhD in 1991, his research is essentially dedicated to the differences between information processing with and without consciousness, particularly in the domain of learning and memory. He is currently the president of the European Society for Cognitive Psychology and a member of the executive committee of the Association for the Scientific Study of Consciousness. He has authored and edited several books as well as numerous articles dedicated to consciousness. He is the editor, together with Tim Bayne and Patrick Wilken, of the *Oxford Companion to Consciousness*.

Implicit Social Cognition

Y Bar-Anan and B A Nosek, University of Virginia, Charlottesville, VA, USA

Glossary

Behavior priming – The automatic activation of a behavior as a result of perceiving a stimulus related to that behavior.
Goal priming – Automatically activated pursuit of a goal as a result of perceiving a stimulus related to that goal.
Implicit attitudes – Automatic associations between concepts and valence (e.g., good, pleasant).
Implicit self-concept – Automatic associations between the self and attributes (e.g., athletic, intelligent).
Implicit social cognition – Mental operations, related to the self or other people, that happen automatically and can occur without awareness.
Implicit stereotypes – Automatic associations between social groups and attributes.

Introduction

During a periodical health checkup, medical doctors ask patients questions about their health and conduct physical examinations such as listening to lungs and examining the ear canal. After hearing the patients' report of their health, why does the doctor bother with time-consuming physical examinations? Is it reasonable to think that patients can report about their own bodies well enough to make the doctor's examination irrelevant? Of course not. People do not have complete self-knowledge about the health of their own bodies. A doctor's examination provides a way to assess health without relying on patients' self-report because patients do not know everything about their health, and because their reports are not always reliable.

While it is easy to understand that people do not have perfect self-knowledge about their physical state, it is less obvious that people do not have perfect self-knowledge about their thoughts and feelings toward themselves and others. People often believe that they know their own feelings and thoughts quite well. However, some thoughts and feelings are explicit in that they are reportable by the person, but others are implicit in that they are not reported. They are not reported either because people are unaware of having them, or because people are aware of them but do not consider them to be their 'true' thoughts or feelings.

Implicit cognition is closely related to the concept of automaticity – thoughts or actions that occur spontaneously or uncontrollably. Thoughts and actions that occur automatically may occur without attention or awareness. For example, after traveling the route between home and office many times, a driver might be able to travel the route automatically. The driver's attention can drift away to think about other things. On occasion, attention may suddenly leap back into driving awareness with recognition that the mind was 'somewhere else.' Did I stop at the stop sign back there? Why am I heading home when I wanted to stop by the store first? In this way, the same behavior – driving – can occur with or without awareness because the relevant thoughts and skills have been automatized.

This example emphasizes how automatic thoughts and actions can occur without awareness, and are efficient in that they can happen while thinking about something else. Automatic thoughts and actions can also be difficult to control. For example, some automatic thoughts and actions can be difficult to prevent from starting. A person with spider phobia may not be able to prevent the thought that there might be spiders in his shoes every time he goes to put them on. Automatic thoughts and actions can also be difficult to stop once they have started. Now that the thought of a spider being in the shoe has

come to mind, it may be difficult for the person to avoid checking the shoe even if he knows that the thought is irrational. These features are important for implicit cognition because thoughts that happen efficiently, without control or without awareness, can be quite different than what the person explicitly wants to think or do.

Implicit Social Constructs

When people are asked "Do you like Canadians?," "Do you believe that Muslims are dangerous?," "Are you a good person?," or "Do you want to help me solve this problem?," they can thoughtfully consider the question and generate a response. This is explicit social cognition – the self-assessment of one's own feelings or thoughts about the self or other people.

Implicit social cognition focuses on automatic processes that relate to these same concepts such as attitudes (feelings about whether something is good or bad), stereotypes (beliefs about what people are like), self-concept (feelings or beliefs about oneself), and goals (the ends toward which one directs effort).

Implicit Attitudes

Implicit attitudes are associations between concepts such as 'Asians,' 'seafood,' or 'freedom' and valence such as good or bad, like or dislike, pleasant or unpleasant. When a concept comes to mind, an associated valence can be activated automatically, and may influence how people think about and act on the concept. Valence can even be activated and influence behavior without the person being aware that it happened.

Implicit and explicit attitudes toward the same concept can be quite different. For example, more people show implicit negativity toward Blacks compared to Whites than are willing to report or endorse such feelings explicitly. Despite their distinctiveness, both the implicit and explicit reactions might predict behavior. One study, for example, found that implicit reactions were predictive of people's nonverbal behaviors in an interaction with a Black experimenter – such as their apparent comfort with the interaction – but not what the person said. The researchers surmised that nonverbal behaviors are harder to control, and so implicit attitudes may be more likely to influence them, whereas explicit attitudes may be more influential when the behaviors are easier to control.

In other cases, implicit and explicit attitudes can be similar. For example, self-reported and implicit liking of political parties and candidates tend to be highly related. People who say that they like liberals also tend to show implicit favorability for liberals, and likewise for conservatives. Even so, the strong correspondence between implicit and explicit attitudes does not mean that they are totally redundant for predicting behavior, even in political domains. Studies have shown that the implicit attitudes of explicitly undecided voters predict who they will eventually vote for.

Implicit attitudes are not more correct than explicit attitudes, and the explicit attitudes are not just lies that people use to cover their implicit attitudes. Both implicit and explicit responses can be 'true,' even when in conflict. Behavior is a product of both explicit intentions and automatic reactions.

Implicit Stereotypes

Implicit stereotypes are associations between a social group and an attribute – such as an association between women and nurturing, or between men and aggressiveness. Like attitudes, such associations can exist in memory even if they are not believed. Deciding whether an association is true or false is a deliberate decision. As a consequence, if stereotype exists in memory, and is automatically activated, it could influence perception, judgment, or action related to the stereotype even against one's intentions. For example, studies find that, on average, both men and women associate math with male and liberal arts with female implicitly – an implicit academic stereotype. Women who have a strong math = male association report participating less in math activities, like math less, and perform more poorly than women who have a weak math = male association. This illustrates that implicit stereotypes about groups (math is for men) might even affect thoughts and feelings about oneself (e.g., I am not cut out for a math career).

Implicit Self-Concept

An association between an attribute and the self is an implicit self-concept. Perhaps the most important implicit self-concept is implicit self-esteem – associations between the self and valence. People usually associate themselves with positivity more strongly than negativity – a positive implicit self-esteem. Because most people have positive associations with the self, a reliable way to increase a person's liking for something is to get them to associate it with themselves. For example, people tend to like fictional characters in books and films that share their name, people are slightly more likely to marry others with the same first initial as theirs, and people are even slightly more likely to select an occupation that is similar to their own name (e.g., Dennis is slightly more likely to become a dentist than Larry; Larry is slightly more likely to become a lawyer than is Dennis). Amazing, but true!

Automatic Social Behavior and Goal Pursuit

Associations may also link perception to actions. For instance, when we see a red traffic light (perception), we tend to stop our car (action). According to the psychologist John Bargh, the perception of stimuli may automatically elicit behavior related to the stimuli, even without awareness of the stimulus or of its relation to the behavior. For instance, in one study, participants did a word task that included some words related to old-age such as 'gray,' 'wise,' and 'Florida.' Afterward, the researchers discreetly timed how quickly the participants walked down a hallway compared to another group that did not see the old-age related words beforehand. The 'old-age' primed participants tended to walk more slowly down the hall, suggesting that the 'old' words made them behave in line with a behavior related to aging. The age words are called 'primes' because they make the observed behavior accessible. When asked afterward if their walking speed was influenced by reading the age-related words, the participants thought the researchers were crazy and insisted that such priming surely did not affect them. The participants were unaware of the influence of the primes on their behavior.

Similar studies show that priming methods can activate goals implicitly. For example, researchers primed the goal of helping by putting words such as 'nurse' and 'firefighter' in a word search puzzle. This made participants more likely to pursue helping-related goals, compared to participants who were not primed, such as helping the experimenter pick up a dropped tissue. Priming can make people behave as if they had a goal, such as be helpful, even if they do not report it or know why they had it.

According to one perspective, a social function of automatic behavior and goal pursuit is to improve social interaction. For instance, people tend to unconsciously imitate the body gestures and the language use of their conversation partners. When one person puts her hands on her hips, the other is more likely to adopt a similar posture. This imitation promotes feelings of shared understanding and companionship. However, if people become aware that they are being imitated, then it can feel creepy and disingenuous, the complete opposite result. Lack of awareness of mimicking and being mimicked appears to be necessary for it to improve social relations. This illustrates that there can be advantages for having parts of social thinking and behaving occurring implicitly.

Formation and Change of Implicit Constructs

Implicit constructs are thought to reflect the summary of a person's social experience. Associations form from exposure to pairings of concepts and attributes. The mind is prepared to form associations because learning "what goes with what" can be very helpful for planning new behaviors. Forming associations between rotten milk and illness or between a hot stove and danger will help to avoid those things in the future. If those associations did not form, then people would not learn from past mistakes, and would find themselves having sore tummies and hands on a regular basis.

Every moment of the day offers opportunities to create associations that might assist with planning and understanding the future. Even so, we might not agree with every association that we observe. However, because forming associations is

so important, the mind does not ask permission to create them. As a consequence, people can possess associations that they would rather not have. For instance, if a certain social group is persistently portrayed as negative in news and entertainment media, then that association may form, even in the minds of people who belong to that social group.

Because implicit constructs are based on experience, new information can change them. And, even though implicit constructs occur automatically, they are surprisingly sensitive to the immediate social environment. For example, thinking about a strong woman for a few minutes can temporarily reduce automatic gender stereotypes of women as weaker than men. Also, trying to get along with a Black person can reduce automatic negativity toward Black compared to White people.

Repeated exposures to associations that differ from what already exists in mind can have a longer lasting effect on implicit constructs. In one study, students in women's college showed reduced automatic stereotyping of men as leaders and women as supporters because they had numerous women as their instructors. Also, treating people with spider phobia with a technique that has lots of safe exposures to spiders is effective at reducing automatic negative responses toward spiders.

These examples suggest that while implicit constructs can have an unintended influence on behavior, the undesirable ones can be altered by making deliberate changes to the associations that are accumulated in one's mind – perhaps by changing what is experienced in everyday life.

Measurement of Implicit Constructs

Measurement of explicit attitudes, stereotypes, self-concepts, and goals has a long history and is easy to comprehend. If you want to know how a person feels about bananas, ask them something like ''How do you feel about bananas?'' and provide response options such as ''I love them'' and ''I hate them.''

A key feature of implicit constructs is that it may not be sufficient to ask for direct responses because people may be unaware of or unwilling to report all of their relevant thoughts and feelings. How then can implicit constructs be measured? Early attempts at implicit measurement date back to Sigmund Freud. Freud developed methods such as free association and dream analysis to circumvent the limitations of self-report for revealing implicit thoughts and feelings. While tremendously innovative, Freud's approaches are not popular in modern experimental psychology because of their unreliability and questionable validity.

Modern methods assess implicit thoughts and feelings indirectly by having participants do a task and inferring their implicit reactions from their performance on the task. For example, to measure implicit attitudes, a method called evaluative priming presents words or pictures on a computer screen in rapid succession. Two items are presented in a row and respondents judge whether the second item (e.g., the words 'horrible' and 'wonderful') is good or bad. The speed with which a person can decide whether the second item is good or bad is influenced by the item that was presented right before it. Someone with negative implicit feelings about Black people, for example, might be faster to rate a bad word and slower to rate a good word if a black face is presented right beforehand. Seeing the black face activates negative feelings automatically, and those feelings influence the judgment of the items that are presented next. The difference in time required to judge good words and bad words following the presentation of black faces (or other social groups and concepts) is the indirect assessment of the participants' implicit feelings about Black people.

Another popular measure called the Implicit Association Test (IAT) is based on similar principles. Using a computer, items representing four different categories (e.g., black faces, white faces, good words, and bad words) are presented on the screen one at a time and the items are categorized into groups with two keys on the keyboard. There are two important conditions in the IAT. In one condition, every time a black face or a good word appears, the respondent hits one key, and every time a white face or a bad word appears, the respondent hits the other key. In the other condition, the rules change. White faces and good words are categorized with one key, and black faces and bad words are categorized with the other. The main idea is that it should be easier to categorize

the items when the concepts on the same response key are associated in memory. So, if the respondent is faster categorizing white faces with good words compared to black faces with good words, it indicates an implicit preference for Whites compared to Blacks. But, if the respondent is faster when categorizing black faces with good words compared to white faces with good words, then it indicates an implicit preference for Blacks compared to Whites.

The IAT can be adapted to measure a variety of associations. This example is for measuring an implicit attitude – associations between racial groups and valence. If the 'good' and 'bad' concepts were changed to 'academic' and 'athletic,' then it would be a measure of implicit stereotyping – associations between racial groups and attributes. If the 'good' and 'bad' concepts were changed to 'self ' and 'other,' it would be a measure of implicit self-concept – associations between the self and attributes; in this case how much the respondent automatically associates the self with Blacks versus Whites.

Conclusions

Behaviors and thoughts often do not involve conscious awareness or conscious control. They can affect thought, action, and feeling without people's knowledge, and are often detected only by using subtle experimental manipulations such as priming and implicit measures, which do not rely on self-report. Implicit processes affect the way we perceive, evaluate, and behave toward others and ourselves. Because these effects often happen without awareness or control, the social psychologist Timothy Wilson suggests that we can be ''strangers to ourselves.'' Conscious experience and deliberate actions are an important part of one's identity, but there is a sea of activity happening behind the scenes that also shapes 'who we are' and 'how we behave.'

See also: Perception: Subliminal and Implicit; Perception: Unconscious Influences on Perceptual Interpretation; Social Foundations of Consciousness; A Social Neuroscience Approach to Intergroup Perception and Evaluation.

Suggested Readings

Bargh JA (1994) The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In: Wyer RS and Srull TK (eds.) Handbook of Social Cognition, vol. 1, pp. 1–40. Hillsdale, NJ: Erlbaum.

Bargh JA, Chen M, and Burrows L (1996) Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. Journal of Personality and Social Psychology 71: 230–244.

Bargh JA, Gollwitzer PM, Lee-Chai A, Barndollar K, and Trötschel R (2001) The automated will: Non-conscious activation and pursuit of behavioral goals. Journal of Personality and Social Psychology 68: 768–781.

Cesario J, Plaks JE, and Higgins ET (2006) Automatic social behavior as motivated preparation to interact. Journal of Personality and Social Psychology 90: 893–910.

Dasgupta N and Asgari S (2004) Seeing is believing: Exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. Journal of Experimental Social Psychology 40: 642–658.

Dovidio JF, Kawakami K, and Gaertner SL (2002) Implicit and explicit prejudice and interracial interaction. Journal of Personality and Social Psychology 82: 62–68.

Fazio RH, Jackson JR, Dunton BC, and Williams CJ (1995) Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? Journal of Personality and Social Psychology 69: 1013–1027.

Greenwald AG and Banaji MR (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review 102: 4–27.

Greenwald AG, McGhee DE, and Schwartz JLK (1998) Measuring individual differences in implicit cognition: The Implicit Association Test. Journal of Personality and Social Psychology 74: 1464–1480.

Lakin JL and Chartrand TL (2003) Using nonconscious behavioral mimicry to create affiliation and rapport. Psychological Science 14: 334–339.

Nisbett RE and Wilson TD (1977) Telling more than we can know: Verbal reports on mental processes. Psychological Review 84: 231–259.

Nosek BA, Banaji MR, and Greenwald AG (2002) Math = Male, Me = Female, therefore Math ^ = Me. Journal of Personality and Social Psychology 83: 44–59.

Pelham BW and Swann WB, Jr. (1989) From self-conceptions to self-worth: The sources and structure of self-esteem. Journal of Personality and Social Psychology 57: 672–680.

Wilson TD (2002) Strangers to Ourselves: Discovering the Adaptive Unconscious. Cambridge, MA: Harvard University Press.

Biographical Sketch

Yoav is a student at the University of Virginia, expected to receive his PhD by May 2009. He holds an MA in social psychology from Tel-Aviv University. Yoav studies self-knowledge, implicit cognition, automaticity, the origin of thought, different effects on abstract versus concrete thinking, and methods that tap the associative map in people's mind. He is the Israeli collaborator in Project Implicit's multicultural collaboration (https://implicit.harvard.edu/implicit/israel/).

Brian Nosek received his PhD from Yale University in 2002 and is an associate professor in the Department of Psychology at the University of Virginia. In 2007, he received early career awards from the International Social Cognition Network (ISCON) and the Society for the Psychological Study of Social Issues (SPSSI). He directs Project Implicit (http://projectimplicit.net/), an Internet-based multiuniversity collaboration of research and education about implicit cognition – thoughts and feelings that exist outside of awareness or control. Nosek's research interests include implicit cognition, automaticity, social judgment, attitudes and beliefs, ideology, morality, identity, memory, and the interface between theory, methods, and innovation.

Inner Speech and Consciousness

A Morin, Mount Royal College, Calgary, AB, Canada

Glossary

Inner speech – Subvocal speech – talking to oneself in silence.
Left inferior frontal gyrus (LIFG) – LIFG (Broca's area) – brain area that has been shown to be active during inner speech production.
Private speech – Speech-for-self emitted out loud by children for self-regulatory purposes.
Self-awareness – Awareness of external and internal stimuli, which includes a sense of self.
Self-regulation – Self-guidance – setting immediate and distant goals, problem-solving, planning, and decision making.
Self-talk – Refers to both inner and outer self-directed speech.
Working memory – System that keeps a limited amount of information in an active state for a short period of time.

Introduction

Consciousness partially consists of a silent running verbal commentary describing one's current perceptual, sensory, motor, cognitive, mnemonic, and emotional experiences. This activity of talking to oneself in silence is called inner speech and is part of the wider process of intrapersonal communication, which also includes mental imagery. Inner speech occupies a significant portion of consciousness, as people report that approximately one-fourth of their conscious waking life involves silent verbal thinking. A host of equivalent terms are used to refer to the phenomenon of inner speech: self-talk (which includes both inner and outer self-directed speech), propositional thought, subvocal speech, covert speech, self-referent speech, internal dialogue, internal monologue, auditory imagery, subvocalizations, utterances, self-verbalizations, and self-statements. The expressions working memory, verbal rehearsal, and phonological loop specifically apply to inner speech used for mnemonic purposes. A communicative speech has been suggested as an umbrella term covering all forms of speech-for-self; this proposal is somewhat problematic however, since the self actually does communicate with itself when engaging in inner speech.

Jean Piaget utilized the term egocentric speech to refer to self-talk emitted aloud by children in social situations without any preoccupation of being understood by others; his view was that egocentric speech served no function whatsoever and simply represented a manifestation of children's cognitive immaturity. In sharp contrast, Lev Vygotsky used the expression private speech to designate speech-for-self emitted out loud by children for self-regulatory and problem-solving purposes. He thus postulated that self-talk played an important cognitive function and reflected intellectual development – not egocentrism. Echolalia constitutes a primitive form of private speech where young children repeat others' words in an automatic fashion for the mere pleasure of using words. The nightly soliloquies that some children produce between 1 and 3 years of age before they fall asleep are called crib speech. The term 'embedded private speech' specifically denotes adults' use of private speech during public lectures for strategic restructuring and self-regulation goals (e.g., ''Let's see, where was I?''; ''Do I have all the transparencies?''). Inner speech writings are rapidly recorded notes for self in notebooks, personal journals, shopping lists, etc. These memos usually consist of single words or phrases, or full paragraphs of highly condensed and often cryptic writing.

Theoretical Perspectives

What is the exact nature of inner speech? Why do we talk to ourselves? How does inner speech

develop? What is the relation between inner speech and thought? These are some of the key questions philosophers and psychologists have been raising for centuries. Most of these are empirical in nature and will be addressed in forthcoming sections. A brief summary of the inner speech and thought debate is presented below, followed by an equally concise introduction to Vygotsky's influential sociocultural theory of inner speech.

Two opposing and rather extreme positions have been proposed concerning the relation between thought and language: (1) language (including inner speech) literally is thought and (2) pure thought can exist without language. The first view was held by Plato, who wrote that ''When the mind is thinking, it is simply talking to itself, asking questions and answering them.'' Behaviorists such as John Watson also believed that thought should be equated with inner speech. Charles Darwin obviously embraced that position when he wrote that ''A long and complex train of thought cannot be carried on without the aid of words, whether spoken or silent, than a long calculation without the use of figures or algebra.'' Similarly, Vygotsky proposed that thought is not expressed in words – it comes into existence through them. A somewhat weaker version of this first view is provided by the Sapir–Whorf hypothesis, according to which any particular language influences the habitual thought of its speakers. Different language patterns lead to different patterns of thought, so that the use of vocabulary that is specific to one's native language, for instance, will color one's perception of the world. The second position was supported by the Wurzburg's school of thought founded by Oswald Kulpe in the late 1800s: pure thought can exist without language, and thus inner speech; thought can be imageless. Karl Buhler's work at the beginning of the twentieth century aimed at defending this view. A proverb was read to participants and the experimenter would ask them to press on a button once they understood its meaning. Participants reported that no verbal thoughts or images were present when they pressed the button. Hence, understanding and thought seemed to precede language. Contemporary cognitive scientists and linguists hold neither of these extreme positions and agree that the question should not be 'Does language cause thought?' or 'Does thought cause language?' Rather, it should be 'How does language affect thought processes?'

A somewhat related controversy opposes Ludwig Wittgenstein and Jerry Alan Fodor. The former proposed that we think in words using natural language (e.g., real symbols written on paper; inner speech), whereas the latter suggests that thought requires Mentalese – innate cognitive, and more complex, abstract mental representations that differ from natural language.

Quite a few theories of inner speech have been put forward, and Russian scholars must be credited for having formulated the most comprehensive, innovative, and coherent proposals. Vygotsky's work in particular, written in the mid-1900s, represents a landmark in that respect, with its emphasis on culture, language, and internalization. In Vygotsky's view, culture contributes to children's intellectual development in two ways. First, through culture children acquire much of the content of their thinking, that is, their knowledge. Second, culture shapes children's higher mental functions by not only teaching children what to think, but also showing them how to think. Cognitive development grows out of a dialectical process whereby children learn through problem-solving experiences shared with social agents such as parents, teachers, siblings, peers, etc. There is a difference between what children can do on their own and what they can do with help. Vygotsky called this difference the zone of proximal development. At first people interacting with children assume most of the responsibility for guiding the problem-solving process, and then gradually this responsibility transfers to the child. Language constitutes the main form of interaction through which social agents transmit information to children. As learning evolves, children's own language comes to serve as their primary tool of intellectual development. Eventually, they come to use private speech (and later, inner speech) to guide and control their own behavior. This is internalization, the process of using tools of thought that first exist outside children. Again, according to Vygotsky, this happens by and large through inner speech.

In essence, Vygotsky suggested that inner speech has its origins in social speech and that it serves an important self-regulatory function – a

notion that has received much empirical support. For instance, the internalization process entails that children will first talk to themselves aloud (private speech) and that this self-guiding talk will gradually go underground as inner speech. This is indeed the case.

Measurement Techniques

Like most psychological inquiries, initial attempts to study inner speech relied on introspection. For example, Alfred Binet asked his two daughters to work on various problems and then asked them how they were able to solve them. He noted that the daughters would often report things like "Well, I told myself this. . ." or "I said to myself that. . .," and concluded on that basis that most thinking was mediated by internal speech.

Since then, measurements have been refined, leading to advancements in our understanding of inner speech. Because spontaneous emission of private speech by children can accurately be recorded and quantified, it has been extensively studied in natural settings (e.g., in the classroom) and in the laboratory in various situations (e.g., with others vs. alone, working on goal-directed vs. unfocused activities). Private speech may be quantified as follows: raw utterance counts, utterances per minute, proportion of total speech or total private speech, or ratio of social to private speech. Most studies code and classify verbalizations into different categories, which are then correlated with behavior or performance. Examples of such categories are (1) task-irrelevant private speech that includes word play, affect expressions, and comments to imaginary others; (2) task-relevant private speech that contains statements about the task or the child's ongoing or future task-related activity; and (3) partially internalized private speech made up of inaudible muttering, whispers, and silent, verbal lip movements.

The think out loud method consists of recording the verbalizations of adult participants who are explicitly instructed to vocalize their thoughts while engaging in a given task. The assumption is that these verbalizations will reflect actual inner speech activity, or at least will provide a representative sample of it. For this sample to be as unbiased and natural as possible, directives clearly specify not to censor thoughts or to worry about making sense. The videotape reconstruction procedure involves showing volunteers video recordings of their behavior in specific situations (e.g., during social interaction) and asking them to report (i.e., to reconstruct) inner speech activity. This technique is less intrusive than the think out loud method but presents a problem of its own: video cameras are notorious for inducing public self-focus when directly facing the participant, which actually inhibits reports of personal thoughts; this can be easily avoided by positioning the video device sideways. With the thought listing method participants are invited to catalog their verbal mental activity after completion of a task. The thought-sampling technique aims to obtain a typical sample of people's inner speech in natural settings. Volunteers wear a paging device that delivers auditory signals at random intervals throughout the day; they are instructed to stop upon hearing the signal and to note the content of their consciousness, including inner speech use. In all the assessment methods mentioned above, inner speech is coded and classified into various groups that are then correlated with behavior or task performance.

The most popular tool for measuring inner speech is questionnaires consisting of self-statements along various possible dimensions, for example, anxious versus nonanxious ("This is too much"; "I can cope"), positive versus negative ("I feel good"; "I wish I could die"), social phobia ("I have nothing intelligent to say"); participants indicate their frequency of self-talk use on a Likert-scale. Unlike the time-consuming and relatively complicated think out loud method and related variations, questionnaires can be easily and rapidly administered to large groups of individuals. However, because such scales contain a predetermined set of self-statements, they seriously limit the range of spontaneous inner speech that participants can report. In technical terms, questionnaires lack ecological validity.

Electromyographic recordings of movements of the lips and tongue have also been used to assess inner speech frequency during problem-solving tasks. Alexander Sokolov devotes an entire book to this method. Electromyography is a technique for evaluating and recording physiological

properties of muscles. This is performed with an electromyograph that detects the electrical potential generated by muscle cells when these cells contract, and also when the cells are at rest. The premise here is that movements of the lips and tongue produced during overt speech are also observed (albeit with much less amplitude) during covert speech, so that these articulations can be taken as objective outer manifestations of inner speech activity. Recordings are typically made with suction electrodes placed on the tongue, sublingual horseshoe electrodes positioned under the tongue, or surface electrodes affixed to the lower lip. Electrodes translate articulatory movements into electrical signals of various amplitudes that convey information about intensity of inner speech activity as a function of time during completion of a multitude of mental tasks, for example, mental arithmetic, silent reading, listening to speech, recollection of verbal material, and manipulation of graphic-visual material. If articulatory movements are observed during subvocal speech, then substantial interference of these movements should lead to inner speech disruptions. This last method is called articulation suppression: participants are asked to perform some task (e.g., understanding or memorizing speech) while simultaneously reciting verses or mentally counting backward from 100. Articulatory suppression obviously does not represent a measure of inner speech *per se*; it is nonetheless very instructive to learn what one cannot do without inner speech.

Each assessment technique has its advantages and disadvantages; the nature of the problem being investigated should ultimately dictate what method to use. To illustrate, in a preliminary phase of a study, an open-format procedure (e.g., thought listing, think out loud method) would be adequate for a researcher interested in gathering freely generated verbalizations from participants experiencing social anxiety. These verbalizations could then be used to build a validated questionnaire that could be administered to large groups of individuals in a subsequent stage of the study. It is also common (and highly recommended) to employ multiple method assessment, for example, to measure inner speech with the videotape reconstruction procedure and the think out loud method in a single study.

Development and Characteristics

Vygotsky's hypothesis regarding the social origin of inner speech finds support in studies that report strong positive correlations between rates of social interaction and private speech in children. Children raised in environments that are low in verbal and social exchanges show a delayed development of private speech. Conversely, children exposed to rich language environments and cognitive stimulation at home – situations more typical of families of higher socioeconomic status – appear to use and internalize private speech earlier than children from families of lower socioeconomic status. Also, mostly in agreement with Vygotsky's original views, both cross-sectional and longitudinal investigations confirm that the frequency of children's private speech follows an inverted-U relation with age, peaking at 3–4 years of age, decreasing at 6–7 years of age, and virtually disappearing at age 10. The reduction in private speech is accompanied by corresponding increases in the frequency of partially internalized manifestations of inner speech, such as whispers and inaudible muttering. Private speech of bright children gets internalized into inner speech earlier, with girls usually showing a faster private speech development than boys. Children become aware of engaging in private speech at around age 4. Concerning the aforementioned ontogenetic pattern in frequency of private speech, it should be noted that (1) it is often observed only among certain subtypes of private speech rather than in all forms of self-talk, (2) age-related changes in children's private speech use in naturalistic classroom settings seem to be more extended and gradual than those recorded in laboratory studies, and (3) similar curvilinear trends in private speech usage repeat themselves microgenetically as children of different ages master new challenging tasks.

Very little is known regarding potential cultural differences in private speech development and frequency. Only one study examined private speech in British and Saudi Arabian children and found no differences in frequency of private speech.

Vygotsky originally postulated that once self-talk has been fully internalized as inner speech, it does not resurface as external speech for self. However, recent work demonstrates that healthy adults do use private speech when alone for self-regulatory

purposes, as well as for spatial navigation and search, concentration, and affective discharge and control. In one study, 96% of all adult participants reported sometimes talking to themselves aloud.

In a classic series of experiments on verbal mediation, Alexander Luria studied the extent to which both external and self-generated verbal commands effectively regulate children's behavior. Youngsters were instructed to press a rubber bulb as they were told by the experimenter to start, stop, or coordinate presses with a flashing light and with his own words. Luria observed the following developmental sequence. At 1½ to 2½ years, the initiating function of speech by the experimenter (start) was effective but not the inhibiting function (stop). Self-initiating and inhibiting functions of speech at that age were absent. At 3 to 4 years, both the initiating and inhibiting functions of the experimenter's speech and the initiating function of the child's own speech were observed, but not the self-inhibiting function. The full regulating function of the child's own speech was present at 4½ years.

Studies of spontaneous self-regulatory private speech in children indicate that at first it follows action, then it occurs simultaneously with behavior, and finally it precedes it. Children's self-talk becomes gradually more self-regulatory in nature between 3 and 4 years of age. Caregivers who initiate dialogues which actively engage children as collaborative partners in problem-solving activities promote the development of self-regulatory private speech. Such conversations allow children to incorporate mental strategies and build individualized verbal statements adapted to solving problems encountered when alone.

An important transition in the way in which children spontaneously use private speech seems to take place between the ages of 3 and 4. Three-year-old children tend to talk to themselves across a very wide range of situations (e.g., goal-directed and unfocused activities, sustained or rapidly changing activities), whereas 4-year-old children's private speech tends to occur specifically during self-selected but focused, sustained, goal-directed activity. This suggests that 4-year-old children may be using private speech in situations in which it fruitfully serves self-regulatory purposes.

Once private speech has been fully internalized as inner speech, it develops a life of its own with unique qualities different from those of social speech. Overall, the semantic dimension of speech becomes most salient while its syntactic and phonological aspects fade away. The most important characteristic of inner speech is that it is predicative – syntactically crushed, condensed, and abbreviated. Since the context of speech is always implicit to the talking agent, the subject of a thought does not need to be explicitly stated. This predicative quality of inner speech is responsible for individuals experiencing it not as a sequence of fully formed utterances, but instead as a fragmentary series of verbal images. This explains why the rate of internal speech is much more rapid than that of overt speech. There is also a prevalence of sense over meaning in inner speech, which refers to the way that the personal, private significance of words takes precedence over their conventional meanings. Agglutination involves the development in inner speech of hybrid words signifying complex, subject-specific concepts.

Inner speech also contains remnants of the dialogic quality of social speech. Since social speech essentially constitutes a dialogue between two people, then some speech-for-self should possess a dialogic structure, with a speaking self (the generative, producing, inner voice) and a self talked to (the perceptual, auditory, inner ear). Some segments of inner speech consist of a series of alternating lines – questions and answers or directives and answers – a format that closely resembles social verbal interactions. Many aspects of overt speech perception (e.g., sex, loudness, accent, and dialect) are absent during self-induced internal speech. Note that the dialogical or social view of inner speech has been contested by the foundationalist or reflection view, which states that the conversational duality of inner speech is more apparent than real. It is only the reflecting self that does the talking; inner speech is monologue, not dialogue.

Neuroanatomy

Brain-imaging techniques such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) technique have made it possible to identify one main brain area that is more active during inner speech production: the left

inferior frontal gyrus (LIFG). The LIFG reliably gets activated when participants are asked to silently articulate sentences or single words; similarly, the LIFG is recruited when volunteers engage in working memory tasks that require covert rehearsal of verbal material. Neuropsychological evidence further confirms that accidental destruction of the LIFG disrupts inner speech. Repetitive transcranial magnetic stimulation (rTMS) interferes with normal brain activity; when applied to the LIFG it blocks internal speech. rTMS applied to the motor cortex of the left hemisphere, which controls mouth and tongue muscles, also inhibits inner speech. This observation is consistent with the view that like overt speech, but to a much lesser extent, covert speech requires articulation. Other brain structures get activated during inner speech tasks, especially temporal regions bilaterally, as well as the cerebellum.

The LIFG is also known as Broca's area, left ventrolateral prefrontal cortex, and left frontal operculum; it corresponds to Broadmann's areas 44, 45, and 47. Broca's area represents the neurological basis of both outer and inner speech production. Interestingly, inner signing in deaf individuals also activates the LIFG. The most anterior part of Broca's area (BA 45) is involved in retrieval of words for their meaning while its posterior part (BA 46/47) is specialized in getting access to words through an articulatory code. The LIFG has been shown to serve various additional functions such as cognitive control (the ability to orchestrate thoughts and actions in accordance with internal goals), working memory (temporarily storing and manipulating information), selection among competing alternatives (choosing among rival sources of information to guide response, e.g., classifying pictures according to one of many different attributes), and interpreting actions of others by looking at hand and mouth movements.

Functions

Self-Regulation

Self-regulation (e.g., verbal self-guidance), which includes setting immediate and distant goals, problem-solving, planning, and decision-making, has been the most extensively studied function of private and inner speech. Tasks that require the elaboration of complex behavioral sequences and the simultaneous appreciation of multiple behavioral options are usually better performed with the aid of self-talk. Four effective categories of problem-solving self-verbalizations have been identified: (1) a precise definition of the problem ("Ok. What's the problem? What am I supposed to do?"); (2) an effective approach to the problem ("I must think of ways to solve this problem"); (3) a sustained focus on the problem ("No. That's not important, I must not focus on this. I must work on that"); and (4) a progress evaluation that includes praise or strategy readjustment ("Good! I did it!"/"No. That's not it. That's OK. I must try again and take my time").

As already mentioned, children first learn to respond to adult verbal commands to orient and control their own behavior; this regulatory function of language gradually gets internalized and becomes increasingly self-generated. Private speech use increases linearly with task difficulty. Examples of typical tasks that are employed in research are puzzles, memory tasks, picture classification and discrimination, and sequencing tasks. Children's private speech is maximized under circumstances when there is a need for executive control and there is a relative absence of regulation provided by others. More specifically, children are more likely to self-regulate with private speech in the following situations: (1) when they are engaged in goal-directed, academic, or problem-solving activities compared to free play or other activities; (2) when the problem-solving task is challenging but achievable as opposed to easy; (3) when they are either alone or with peers as opposed to in the presence of an adult who is regulating their behavior; (4) when they are working with an adult who is appropriately scaffolding their problem-solving activity compared to an adult who is highly directive. Scaffolding describes a type of assistance that involves high assisting with only those skills that are beyond children's capability.

So private (and presumably inner) speech use increases with task complexity. Does self-talk actually enhance performance as well? Evidence is contradictory – two key factors are task difficulty and concurrent versus future performance. When a task is too difficult, private speech is likely to

interfere and be associated with task failure; when the task is within the child's zone of proximal development (i.e., within the child's ability range), it will enhance performance. In addition, private speech is frequently correlated with future rather than with concurrent success. The relationship between private speech and task performance is mostly delayed, or diachronic, as opposed to synchronic. As a result, the influence of private speech on task performance is often observed during task sessions subsequent to the time at which private speech was emitted.

Self-regulatory use of self-talk in athletes during training and competition has been widely studied in a broad range of sports, for example, water-polo, golf, skating, gymnastics, basketball, and wrestling. Self-talk use has been compared to other cognitive strategies, especially mental imagery. Results largely establish that self-verbalizations effectively enhance sport performance. The following dimensions of self-talk are typically assessed or manipulated in studies. (1) Valence, that is, positive self-statements that keep the focus of attention in the present, not on past errors or in a distant future, and negative self-statements that interfere with performance because they are inappropriate, irrational, counterproductive, or anxiety-producing. Research indicates that negative self-talk tends to lead to poorer performance, rather than positive self-talk increasing performance. (2) Overtness, that is, how athletes' self-statements are verbalized, up to what point they are overt as opposed to covert. Although direct evidence is lacking, overt self-talk has been postulated to be more effective at improving performance, partially because it helps the athlete to impersonate his or her coach. (3) Self-determination, that is, self-selected, freely chosen self-talk as opposed to assigned by a trainer. Self-determined talk is assumed to have greater motivational power. (4) Self-instruction versus self-motivation. Self-instructional talk represents guidance, how-to-perform talk that is best used in practice settings, whereas self-motivational talk constitutes I-can-do-it talk particularly effective in competitive settings. (5) Frequency of self-talk. Frequency tends to increase across phases of a sporting season and is associated with improved performance, but excessive frequency (called paralysis by analysis) is detrimental.

Language

All aspects of normal language functions (e.g., reading, writing, speaking, and calculating) require intact inner speech, and indeed, loss of inner speech following brain damage invariably leads to aphasia, agraphia, alexia, acalculia, and impaired verbal short-term memory. Recent experiments show that speakers monitor their own inner speech in order to detect and repair phonological, lexical, or grammatical errors before they are spoken. Inner speech use, and even crib speech, have been linked to spontaneous pronunciation practices and grammatical drills. Several lines of evidence corroborate the association between inner speech and silent reading. To illustrate, electromyographic recordings of lip movements show significant increase during silent-language recitation tasks; in patients with frontotemporal dementia and Gilles de la Tourette's syndrome, loss of control over inner speech as evidenced by coprolalia (emitting undesirable vocalizations in social settings) is associated with an inability to read in silence.

The scientific literature reports one case of an individual with no inner speech who nonetheless retained most language functions. Although mute, the patient exhibited normal speech perception, reading, writing, memory, and calculation. How could this patient perform such language tasks without inner speech? Extensive examination revealed that he used mental imagery as a compensatory strategy. Spoken or written words triggered vivid visual images of corresponding pictures. That is, the patient could actually picture words in his mind (as opposed to verbalizing them) and talk to himself that way. A somewhat related question is language and inner speech use in deaf people. Since congenitally deaf individuals never get to hear sounds, including spoken words, they cannot acquire normal language abilities and eventually internalize these as inner speech. So deaf people cannot speak with or think in words, but those who learn sign language not only develop highly adequate interpersonal communicative skills but they also effectively talk to themselves using sign language. In orally competent individuals, articulatory suppression impedes performance on tasks requiring inner speech use. In

deaf individuals proficient in sign language, such a decline in task performance is achieved by having signers grip building blocks tightly in their hands, thus inhibiting self-sign language use. Deaf individuals have been observed to spontaneously think aloud with fluttering hands when working on difficult test questions.

Memory

Inner speech is intimately associated with memory functions, especially working memory. Working memory is a system that allows us to maintain a limited amount of information (1–10 items, e.g., a phone number) in an active state for a short period of time (up to 60 s) and to manipulate that information. It is considered to be necessary for higher cognitive processes such as reasoning, decision making, problem solving, and language understanding. Working memory manipulates verbal and spatial information differently in independent neuroanatomical systems. The basic architecture of each system consists of three different functional components with distinct neural substrates: (1) a pure storage component whose contents decay rapidly; (2) a rehearsal component that can reactivate, or refresh, the rapidly decaying contents of the previous component; and (3) an executive component that regulates the processing of the contents of working memory. Inner speech specifically refers to the rehearsal component of working memory. A simple example of this component could be repeating a phone number to facilitate later recall.

Another form of memory that involves inner speech is autobiography. Although there is no doubt that we store and recall events with images, there is increasing evidence that personal episodes (autobiographical information) are also encoded and retrieved in words. This means that we often recall personal events as self-narratives. A tangential question is inner speech use in bilinguals: In what language does a bilingual individual talk to himself or herself? This issue has precisely been addressed in the context of autobiographical memory. In one cross-cultural study the internal language of autobiographical memory was assessed in Polish people who emigrated to Denmark 30 years ago. Overall, participants reported retrieving personal memories in Polish for the decades prior to immigration and in Danish after immigration. Not surprisingly, these results suggest that personal events that are stored in one language are best retrieved in the same language. Although all immigrants had spent 30 years in Denmark, early immigrators (averaging 24 years old at the time of immigration) reported more current inner speech behaviors in Danish, whereas late immigrators (averaging 34 years old at the time of immigration) indicated more use of Polish.

Other Functions

Inner speech clearly serves other purposes besides self-regulation, language, and memory. For instance, inner speech has been shown to play a role in task-switching performance (i.e., in one's ability to switch back and forth between two mental operations such as when adding and subtracting numbers). Private speech helps children to distinguish their own voice from those of others in social contexts. Self-talk is also used as a tool to rehearse person-to-person communicative encounters in preparation for social performance. It has been linked to selective attention, concept formation, and remembering the goals of actions, and it often represents a vehicle for emotional expression and release. Although speculative, it can be postulated that praying is mediated by inner speech. Last but not least, quite a few theorists have ascribed a role play, or fantasy function, to inner speech. Children often engage in fictive conversations with imaginary friends and describe their own actions and feelings. This would be part of a larger process of differentiating the self from others and enhancing awareness of one's own existence.

Dysfunctional Self-Talk

Overview

Inner speech can be compared to a double-edged sword: on one hand it is associated with very constructive consequences such as self-regulation, and on the other hand distorted self-talk may lead to – or at least maintain – psychological disorders. Conditions such as test anxiety, bulimia, anorexia,

lack of assertiveness, insomnia, social anxiety, agoraphobia, compulsive gambling, male sexual dysfunctions, low self-esteem, and depression have been shown to involve frequent repetitive negative and interfering cognitions. More benign transitory negative states such as worry, guilt, and shame are most likely mediated by inner speech. Defective use of inner speech can be linked to self-deception; some theorists have proposed that dysfunctional self-talk plays a causal role in suicide, criminal activity, and child abuse, but scientific evidence is lacking.

According to the content-specificity hypothesis, all maladaptive behaviors listed above involve negative inner speech related to relevant dysfunctional themes. For instance, anxiety activates ideas of physical or mental harm and doubts about the future (e.g., ''Will I make it?''), whereas depression is accompanied by thoughts of loss, failure, rejection, incompetence, and hopelessness (e.g., ''My future's bleak''). Studies of compulsive gamblers have focused on their irrational thoughts about control of the game. These erroneous beliefs have been captured with the think aloud technique while participants were playing slot machines or roulette. Compared with noncompulsive gamblers, problem-gamblers tend to emit significantly more inadequate verbalizations indicating expectations of success surpassing the laws of probability (e.g., ''I am going to bet on those rows again, this is a good game.'')

A central finding in cognitive clinical psychology is the existence of a basic asymmetry between positive and negative self-statements. Negative self-verbalizations have a more significant dysfunctional impact than positive ones on coping. That is, mentally apprehending events (e.g., an important surgical procedure or exam) is more detrimental than mentally imagining positive outcomes. Healthy individuals are characterized by a 1.7 to 1 ratio of positive to negative self-statements, whereas dysfunctional individuals' ratio is around 1 to 1. This asymmetry is further illustrated as follows: psychotherapy outcome studies assessing successful cognitive change show a decrease in negative thoughts without a corresponding increase in positive thoughts; negative thoughts such as 'mutilation' increase heart rate, whereas positive ones such as 'peace' do not. This observation obviously violates the popular belief of positive thinking, as it may be more important to eliminate negative thoughts than to establish positive ones.

Another disconcerting fact is that attempts at suppressing unwanted thoughts in inner speech (called negative self-referent thoughts, e.g., ''I must stop thinking that I dislike this about myself'') not only cause the thoughts to become hyperaccessible (the negative thoughts are experienced more frequently), but generate more anxiety, depress mood, and lower self-esteem. This ironic phenomenon has been labeled the rebound effect and has been largely documented in both laboratory and naturalistic studies.

Schizophrenia

The intriguing phenomenon of auditory verbal hallucinations in schizophrenic patients is now being increasingly explained in terms of deficient monitoring of their own self-generated subvocal activity. It is undeniable that the voices some of these patients hear in their head are the product of their own inner speech: the brain area that has been shown to be active when schizophrenic patients are experiencing verbal auditory hallucinations is identical to the one responsible for the production of inner speech, namely, the LIFG. The exact nature of the deficit involved is still unclear. The most accepted view suggests that in healthy individuals speech production initiated in the LIFG creates a corollary discharge that sends a message to the left temporal lobe where speech and verbal thoughts are perceived. It is this communication between the frontal and temporal lobes that presumably accounts for the intact speech self-monitoring in normal individuals. Schizophrenic patients seem to experience corollary discharge dysfunctions during speech; auditory hallucinations can be linked to this dysfunction, which prevents patients from recognizing their own inner speech as self-generated. In support of this hypothesis, schizophrenic patients with auditory hallucinations exhibit activity in the LIFG when engaging in inner speech, but unlike healthy participants, fail to show activity in the left temporal cortex. Note that it is unlikely that verbal hallucinations result from an inner speech deficit *per se*:

Schizophrenic patients with severe auditory hallucinations nonetheless show normal performance at short-term memory tasks that require inner speech use. Therefore, the deficit is likely due to monitoring.

A purely psychological explanation of auditory hallucinations has also been proposed. Because inner speech becomes considerably abbreviated as it develops, it tends to lose some of the full-blown dialogic structure of social speech. In healthy individuals, the subjective experience of inner speech thus substantially differs from the experience of conversation. In schizophrenic patients with hallucinations, it is postulated that inner speech has been incompletely abbreviated or is normally abbreviated but temporarily reexpands into a full inner dialogic speech. The phenomenological result of this abnormal developmental process would be perceiving the voices in the dialogue as having an external origin.

Hyperactivity

Hyperactivity in children was originally thought to be partially caused by a lack of self-regulatory private (and inner) speech leading to inadequate self-control. On the basis of cognitive-behavioral approach, a host of procedures were developed to teach agitated youngsters to talk to themselves in order to effectively engage in verbal self-guidance. Therapies typically consisted of gradual steps leading to the internalization of self-regulatory speech: modeling, overt external guidance, overt self-guidance, faded overt self-guidance, and covert self-guidance. Recent reassessments of this method indicate that it is mostly unsuccessful. For one thing, the cognitive-behavioral approach shows short-term gains limited to specific tasks that fail to generalize to broader academic and interpersonal behaviors. In other words, what is being taught is self-control (i.e., copying adults' commands) as opposed to genuine self-regulation (i.e., self-generating flexible plans for action). More importantly, hyperactive children are actually not deficient in spontaneous production of private speech. On the contrary, research now shows that there is an increased private speech use among children with poor self-control compared to healthy kids.

Autism

Autism represents a neurodevelopmental disorder characterized by social, communicative, and imaginative abnormalities in the absence of severe cognitive deficits. Its main feature is lack of social insight and self-awareness. One study suggests that children with autism make limited use of their inner speech. For instance, they do not construct internal verbal codes for pictorial information during memory tasks. It is still unclear if inner speech deficits in autistic patients result from a lack of inner speech, a delay in its development, or poor awareness of how to use inner speech. These deficits may well account for autistic individuals' overall lack of mentalizing abilities, which presumably require verbal labeling of internal mental states. Note that a recent experiment failed to replicate the results reported in the aforementioned study and rather concludes that autistic children's use of inner speech is normal.

Inner Speech, Consciousness, and Self-Awareness

How do language and consciousness relate to each other? More specifically: Do language and, by extension, inner speech play a causal role in consciousness? It all depends on how one defines consciousness. If consciousness is described as being awake and aware of external stimuli, then the answer is clearly no. Prelinguistic infants and nonverbal animals can effectively interact with their environment without having to talk to themselves or to others about it; aphasic patients exhibit a large variety of deficits but always remain fully alert and conscious. In human animals, overt or covert verbal activity may accompany, follow, or even precede conscious experience; consciousness will nonetheless occur in the absence of such verbalizations. However, being conscious is usually defined as including an awareness of internal stimuli and a sense of self, in which case it is more appropriate to employ the term 'self-awareness.' The general consensus is that language is likely required for the emergence of self-awareness, although some have questioned this assertion on neuroanatomical grounds. To illustrate, brain-imaging studies show significant right hemisphere superiority for self-face recognition. Since the right mute hemisphere seems

specialized for a few self-tasks, it is proposed that self-awareness is produced by the right hemisphere and does not necessitate language. Of course such an argument is misleading, as one should not reduce self-awareness to self-recognition. In fact, based on his famous split-brain studies, Michael Gazzaniga rather concludes that conscious awareness arises in the left verbal hemisphere – the Interpreter. In typical experiments testing the cognitive functions of the disconnected hemispheres of commissurotomized patients, visual information presented in the left visual field is exclusively perceived by the right hemisphere and visual information projected in the right visual field is solely seen by the right hemisphere. It is thus possible to flash visual instructions to the right hemisphere and have it generate a given behavior (e.g., 'clap hands'). In such situations the left hemisphere consistently tries to make sense of these behaviors elicited by the right hemisphere – it naturally wants to explain what is happening and always comes up with a plausible (but understandably incomplete) justification. The right nonverbal hemisphere never engages in such interpretational work. Gazzaniga suggests that the left speaking hemisphere in the split-brain patient is trying to preserve an overall feeling of integration and unification that is central to conscious awareness; he further proposes that the left hemisphere in healthy individuals serves the same main purpose: to generate a conscious experience.

Several lines of evidence suggest the existence of a link between language and self-awareness. For example, archeologists have identified a period called the Middle-Upper Paleolithic transition (around 40 000 years ago) during which a cultural Big Bang occurred, characterized by the emergence of the first burials and body adornments, boat-making, more sophisticated tools, and more refined cultural practices. They associate all these changes with the development of self-focused thoughts; interestingly, experts also date the appearance of human language at about this same period. Also relevant is the observations related by Helen Keller, who was blind and deaf but nonetheless managed to learn to use language. Keller states, of the time before she was taught a language, that

> Before my teacher came to me, I did not know that I am. I lived in a world that was a no world. . . When I learned the meaning of 'I' and 'me' and found that I was something, I began to think. Then consciousness first existed for me.

Julian Jaynes put forward a highly controversial theory of self-awareness that is worth mentioning here. Jaynes asserted that until as recently as 3000 years ago, humans were not self-aware. Instead, individuals were guided by mental commands believed to be issued by external gods; however, these instructions were emanating from individuals' own minds. In other words, ancient people experienced verbal hallucinations like modern-day schizophrenic patients. Rather than making conscious evaluations in new or unexpected situations, individuals would hear a voice or god giving admonitory advice and obey these voices without question. Jaynes called this process the bicameral mind. He inferred that these voices came from the right brain counterparts of the left brain language centers – specifically, the counterparts to Wernicke's and Broca's areas. These regions are somewhat dormant in the right brains of most modern humans, but Jaynes noted (incorrectly) that some studies show that auditory hallucinations correspond to increased activity in these areas of the brain. He theorized that a shift from bicameralism marked the beginning of self-awareness as we know it today. The bicameral mind began malfunctioning during the second millennium BC. Jaynes speculated that primitive ancient societies tended to collapse periodically due to increased societal complexity that could not be sustained by this bicameral mindset. The mass migrations of the second millennium BC created a rash of unexpected situations and stresses that required ancient minds to become more flexible and creative. Self-awareness was the culturally evolved solution to this problem. Thus, cultural necessity forced humanity to become self-aware or perish and self-awareness emerged as a neurological adaptation to social complexity.

Support for the notion of an association between inner speech and self-awareness, although far from definitive, is substantial. A strong positive correlation has been repeatedly reported between various validated scales assessing frequency of self-focus and use of inner speech. Studies measuring brain activity during processing of self-information consistently show activation of the medial prefrontal cortex and portions of the left prefrontal lobe that include the LIFG. This implies inner speech

activity during self-awareness tasks. A recent review of the literature found that the LIFG was more frequently recruited during conceptual tasks (e.g., identifying one's emotions or personality traits) than during perceptual tasks (e.g., self-recognition), which further signifies that more abstract self-aspects need to be verbalized in order to be fully brought to consciousness. [Note that this last observation does not imply that inner speech (or language in general) constitutes a perfect cognitive vehicle. Some nonverbal stimuli and experiences (e.g., faces, the taste of a wine) are especially difficult to translate into words (to oneself in inner speech or to others in social speech), and doing so actually interferes with performance (e.g., on a face recognition task). This phenomenon is called verbal overshadowing.] Loss of inner speech caused by brain damage seems to negatively affect self-awareness, as the following quotation by a former aphasic patient suggests:

> I had lost the ability to converse with others, I had also lost the ability to engage in self-talk. In other words, I did not have the ability to think about the future – to worry, to anticipate or perceive it – at least not with words. Thus for the first four or five weeks after hospitalization I simply existed.

Various theories have given a central role to inner speech in self-awareness. One particularly dominant view has been George Herbert Mead's sociological theory. In many respects, Mead's proposal is highly consistent with Vygotsky's theory and complements it very well. Rooted in social interactionism, the theory states that the mind and self emerge from the social process of communication. People act toward things based on the meaning those things have for them; meaning is derived from social interaction and is modified through interpretation. Particularly important in the process of self-awareness development is perspective taking: human beings can imagine how others perceive them and thus can gain an objective point of view on themselves. For Mead, existence in community comes before individual self-consciousness. First, one must participate in the different social positions within society and only subsequently can one use that experience to take the perspective of others and thus become self-conscious. An important distinction is drawn between the 'I' and the 'me': the 'I' represents the subjective self and the 'me' constitutes the objective self that emerges from the perspective taking process. Inner speech is postulated to mediate this process to a great extent because it often initiates a fictional dialogue where verbalization of an objective, and thus different point of view about ourselves is possible. People sometimes engage in self-talk in which they state to real or imaginary persons (Mead's generalized other) their motives for having behaved in a given fashion or for possessing some personal characteristics. When, in response to the expected reactions of others, people explain their actions or describe themselves in self-talk, they take others' perspectives into consideration and thus gain a relatively objective view of themselves.

More contemporary approaches echo Mead's original proposal. To illustrate, Narrative Theory speculates that the self is composed of many I-positions, each of which interacts with the others and each of which has a unique perspective on the person's experience. The I-positions occupy an embodied real or imaginary time and space. Each I-position has a unique psychological quality in addition to a specific spatial perspective, originating from previous experiences and the voices of significant others. Thus, the self is inherently social because the real or imagined I-positions can discourse with each other. The self emerges out of this dialogue as the 'speaker' attempts to clarify his or her perspective to the 'listener.' Daniel Dennett's view of the self as a center of narrative gravity – a verbal autobiography – is highly consistent with Narrative Theory. Another account is that one becomes aware of a mental state (e.g., boredom) when one verbally generates a higher-order thought about that state ("I'm bored"). Others suggest that the process of labeling and categorizing depends on inner speech; this process in turn not only makes it possible for a person to represent internal states and experiences – it brings about the capacity to reflect on them. Reflections can be communicated and discussed with the self in inner dialogues as well as with others. Without inner speech, self-awareness remains relatively primitive, vague, and unelaborated.

Conclusion

Inner speech represents a phenomenon not only central to consciousness but to psychology in

general. The multifunctional dimension of self-directed speech suggests that it plays a fundamental role in initiating, shaping, guiding, and controlling human thought and behavior. The development of reliable assessment techniques have made it possible to empirically investigate many relevant aspects of inner speech activity: neurological bases, characteristics, functions, distorted use and dysfunctional impacts, participation in higher forms of consciousness, and ontogenetic pattern in frequency of private speech, to name a few. Yet compared to other central psychological concepts, inner speech remains neglected and, in fact, is not even mentioned in handbooks of neurolinguistics or introductory psychology textbooks – a remarkable state of affairs indeed.

See also: Intuition, Creativity, and Unconscious Aspects of Problem Solving; Language and Consciousness; Philosophical Accounts of Self-Awareness and Introspection.

Suggested Readings

Allen P, Aleman A, and McGuire PK (2007) Inner speech models of auditory verbal hallucinations: Evidence from behavioural and neuroimaging studies. International Review of Psychiatry 19: 409–417.

Berk LA (1992) Children's private speech: An overview of theory and the status of research. In: Diaz RM and Berk LE (eds.) Private Speech: From Social Interaction to Self-Regulation, pp. 17–53. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carruthers P (2002) The cognitive functions of language. Behavioral and Brain Sciences 25: 657–674.

Fuson KC (1979) The development of self-regulating aspects of speech: A review. In: Zivin G (ed.) The Development of Self-Regulation through Private Speech, pp. 135–218. New York: Wiley.

Hardy J (2005) Speaking clearly: A critical review of the self-talk literature. Psychology of Sport and Exercise 7: 81–97.

Kendall PC and Hollon SD (1981) Assessing self-referent speech: Methods in measurement of self-statements. In: Kendall PC and Hollon SD (eds.) Assessment Strategies for Cognitive-Behavioral Interventions, pp. 85–118. New York: Academic Press.

Larsen SF, Robert W, Schrauf RW, Fromholt P, and Rubin DC (2002) Inner speech and bilingual autobiographical memory: A Polish–Danish cross-cultural study. Memory 10: 45–54.

Levine DN, Calvanio R, and Popovics A (1982) Language in the absence of inner speech. Neuropsychologia 20: 391–409.

Mead GH (1934) Mind, Self, and Society. Chicago: University of Chicago Press.

Meichenbaum D (1977) Cognitive-Behavior Modification: An Integrative Approach. New York: Plenum Press.

Morin A (2005) Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. Journal of Consciousness Studies 12: 115–134.

Schwartz RM (1986) The internal dialogue: On the asymmetry between positive and negative coping thoughts. Cognitive Therapy and Research 10: 591–605.

Sokolov AN (1972) Inner Speech and Thought. New York: Plenum Press.

Vygotsky LS (1943/1962) Thought and Language. Cambridge: MIT Press.

Zivin G (1979) Removing common confusions about egocentric speech, private speech, and self-regulation. In: Zivin G (ed.) The Development of Self-Regulation Through Private Speech, pp. 13–50. New York: Wiley.

Biographical Sketch

Alain Morin received his PhD from Laval University, Quebec, in 1992. Between 1991 and 2001 Dr. Morin taught various courses and conducted research in a host of universities and colleges in the Maritimes and Quebec. At present, he teaches principles of psychology, theories of personality, and social cognition at Mount Royal College, Alberta. Dr. Morin's field of expertise is self-awareness. It includes the cognitive bases of self-reflection with an emphasis on inner speech, levels of consciousness and self-awareness, the neuroanatomy of self-processes, and self-recognition. He is also interested in self-awareness, fame, and self-destruction; the antecedents of self-consciousness; the split-brain phenomenon; and neurophilosophy. Dr. Morin publishes his work in Brain Research Bulletin, Cortex, Consciousness & Cognition, Brain and Behavioral Sciences, Journal of Consciousness Studies, and Journal of Mind and Behavior; he also often contributes to Science & Consciousness Review.

An Integrated Information Theory of Consciousness

G Tononi, University of Wisconsin, Madison, WI, USA

Introduction

Everybody knows what consciousness is: it is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream. It is also all we are and all we have: lose consciousness and, as far as you are concerned, your own self and the entire world dissolve into nothingness.

Neuropsychological and neurophysiological observations indicate that both the quantity and quality of our consciousness depend on specific parts of the brain working in specific ways. Widespread destruction of the cerebral cortex leaves people permanently unconscious (vegetative), whereas the complete removal of the cerebellum, even richer in neurons, hardly affects consciousness. Neurons in the cerebral cortex remain active throughout sleep, yet during certain periods, consciousness fades while at other times we dream. Different parts of the cortex influence different qualitative aspects of consciousness: damage to certain parts of the cortex can forever eliminate the experience of color, whereas other lesions may prevent you from experiencing visual shapes. These observations raise the question of what are the necessary and sufficient conditions that determine the quantity and the quality of consciousness.

Phenomenology: Consciousness as Integrated Information

The integrated information theory (IIT) of consciousness claims that, at the fundamental level, consciousness is integrated information, and that its quality is given by the informational relationships generated by a complex of elements. These claims stem from realizing that information and integration are the essential properties of our own experience. This may not be immediately evident, perhaps because, being endowed with consciousness most of the time, we tend to take its gifts for granted. To regain some perspective, it is useful to resort to two thought experiments, one involving a photodiode and the other a digital camera.

Information and the Photodiode

Consider the following: You are facing a blank screen that is alternately on and off, and you have been instructed to say 'light' when the screen turns on and 'dark' when it turns off. A photodiode – a simple light-sensitive device – has also been placed in front of the screen. It contains a sensor that responds to light with an increase in current and a detector connected to the sensor that says 'light' if the current is above a certain threshold and 'dark' otherwise. The first problem of consciousness reduces to this: when you distinguish between the screen being on or off, you have the subjective experience of seeing light or dark. The photodiode can also distinguish between the screen being on or off, but presumably it does not have a subjective experience of light and dark. What is the key difference between you and the photodiode?

According to the IIT, the difference has to do with how much information is generated when that distinction is made. Information is classically defined as reduction of uncertainty: the more numerous the alternatives that are ruled out, the greater the reduction of uncertainty, and thus the information. It is usually measured using the entropy function, which is the logarithm of the number of alternatives (assuming they are equally likely). For example, tossing a fair coin and obtaining heads corresponds to $\log_2(2) = 1$ bit of information, because there are just two alternatives; throwing a fair die yields $\log_2(6) = 2.59$ bits of information, because there are six.

Let us now compare the photodiode with you. When the blank screen turns on, the mechanism in

the photodiode tells the detector that the current from the sensor is above rather than below the threshold, so it beeps 'light.' In performing this discrimination between two alternatives, the detector in the photodiode generates $\log_2(2) ¼ 1$ bit of information. When you see the blank screen turn on, on the other hand, the situation is quite different. Though you may think you are performing the same discrimination between light and dark as the photodiode, you are in fact discriminating among a much larger number of alternatives, thereby generating many more bits of information.

This is easy to see. Just imagine that, instead of turning light and dark, the screen were to turn red, then green, then blue, and then display, one after the other, every frame from every movie that was ever produced. The photodiode, inevitably, would go on signaling whether the amount of light for each frame is above or below its threshold: to a photodiode, things can only be one of two ways, so when it beeps 'light,' it really means just 'this way' versus 'that way.' For you, however, a light screen is different not only from a dark screen, but from a multitude of other images, so when you say 'light,' it really means this specific way versus countless other ways, such as a red screen, a green screen, a blue screen, this movie frame, that movie frame, and so on for every movie frame (not to mention any sound, smell, thought, or any combination of the above). Clearly, each frame looks different to you, implying that some mechanism in your brain must be able to tell it apart from all the others. So when you say 'light,' whether you think about it or not (and you typically will not), you have just made a discrimination among a very large number of alternatives, and thereby generated many bits of information.

This point is so deceivingly simple that it is useful to elaborate a bit on why, although a photodiode may be as good as we are in detecting light, it cannot possibly see light the way we do – in fact, it cannot possibly 'see' anything at all. Hopefully, by realizing what the photodiode lacks, we may appreciate what allows us to consciously 'see' the light.

The key is to realize how the many discriminations we can do, and the photodiode cannot, affect the meaning of the discrimination at hand, the one between light and dark. For example, the photodiode has no mechanism to discriminate colored from achromatic light, even less to tell which particular color the light might be. As a consequence, all light is the same to it, as long as it exceeds a certain threshold. So for the photodiode 'light' cannot possibly mean achromatic as opposed to colored, not to mention of which particular color. Also, the photodiode has no mechanism to distinguish between a homogeneous light and a bright shape – any bright shape – on a darker background. So for the photodiode, light cannot possibly mean full field as opposed to a shape – any of countless particular shapes. Worse, the photodiode does not even know that it is detecting a visual attribute – the 'visualness' of light – as it has no mechanism to tell visual attributes from nonvisual ones, such as sounds or smells, not to mention which particular sounds or smells, and so on. As far as it knows, the photodiode might just as well be a thermistor – it has no way of knowing whether it is sensing light versus dark or hot versus cold.

In short, the only specification a photodiode can make is whether things are this or that way: any further specification is impossible because it does not have mechanisms for it. Therefore, when the photodiode detects 'light,' such 'light' cannot possibly mean what it means for us – it does not even mean that it is a visual attribute. By contrast, when we see 'light' in full consciousness, we are implicitly being much more specific: we simultaneously specify that things are this way rather than that way (light as opposed to dark), that whatever we are discriminating is not colored (in any particular color), does not have a shape (any particular one), is visual as opposed to auditory or olfactory, sensory as opposed to thought-like, and so on. To us, then, light is much more meaningful precisely because we have mechanisms that can discriminate this particular state of affairs we call 'light' against a large number of alternatives.

According to the IIT, it is all this added meaning, provided implicitly by how we discriminate pure light from all these alternatives that increases its level of consciousness. This central point may be appreciated either by 'subtraction' or by 'addition.' By subtraction, one may realize that our being conscious of 'light' would degrade more and more – would lose its noncoloredness, its nonshapedness, would even lose its visualness – as its meaning is progressively stripped down to just 'one of two ways,' as with the photodiode.

By addition, one may realize that we can only see 'light' as we see it, as progressively more and more meaning is added by specifying how it differs from countless alternatives. Either way, the theory says that, the more specific one's mechanisms discriminate between what pure light is and what it is not – the more they specify what light means – the more one is conscious of it.

Integration and the Camera

Information – the ability to discriminate among a large number of alternatives – may thus be essential for consciousness. However, information always implies a point of view, and we need to be careful about what that point of view might be. To see why, consider another thought experiment, this time involving a digital camera, say one whose sensor chip is a collection of a million binary photodiodes, each sporting a sensor and a detector. Clearly, taken as a whole, the camera's detectors could distinguish among 21 000 000 alternative states, an immense number, corresponding to 1 million bits of information. Indeed, the camera would easily respond differently to every frame from every movie that was ever produced. Yet few would argue that the camera is conscious. What is the key difference between you and the camera?

According to the IIT, the difference has to do with integrated information. From the point of view of an external observer, the camera may be considered as a single system with a repertoire of 21 000 000 states. In reality, however, the chip is not an integrated entity: since its 1 million photodiodes have no way to interact, each photodiode performs its own local discrimination between a low and a high current, completely independent of what every other photodiode might be doing. In reality, the chip is just a collection of 1 million independent photodiodes, each with a repertoire of two states. In other words, there is no intrinsic point of view associated with the camera chip as a whole. This is easy to see: if the sensor chip were cut into 1 million pieces each holding its individual photodiode, the performance of the camera would not change at all.

By contrast, you discriminate among a vast repertoire of states as an integrated system, one that cannot be broken down into independent components each with their own separate repertoire. Phenomenologically, every experience is an integrated whole, one that means what it means by virtue of being one, and that is experienced from a single point of view. For example, the experience of a red square cannot be decomposed into the separate experience of red and the separate experience of a square. Similarly, experiencing the full visual field cannot be decomposed into experiencing separately the left half and the right half: such a possibility does not even make sense to us, since experience is always whole. Indeed, the only way to split an experience into independent experiences seems to be to split the brain in two, as in patients who underwent the section of the corpus callosum to treat severe epilepsy. Such patients do indeed experience the left half of the visual field independently of the right side, but then the surgery has created two separate consciousnesses instead of one. Mechanistically then, underlying the unity of experience must be causal interactions among certain elements within the brain. This means that these elements work together as an integrated system, which is why their performance breaks down if they are disconnected, unlike that of the camera.

Mathematics: Measuring Integrated Information

This phenomenological analysis suggests that, to generate consciousness, a physical system must be able to discriminate among a large repertoire of states (information) and it must be unified, that is, it should be doing so as a single system, not be decomposable into a collection of causally independent parts (integration). But how can one measure integrated information? As we explain below, the central idea is to quantify the information generated by a system, above and beyond the information generated independently by its parts. A brief summary follows.

Information

First, we need to evaluate how much information is generated by the system when it enters a particular state. Consider a system of two binary units that can be thought of as an idealized version

of a photodiode composed of a sensor S and a detector D. The system is characterized by a state it has entered, say [11] (first digit for the sensor, second digit for the detector), and by a mechanism. This is implemented by the causal interactions among its elements and can be described by an input–output table. In this case, the elementary mechanism of the system is that the detector checks the state of the sensor and turns on if the sensor was on, and off otherwise.

Potential, the system could be in any of its four possible states [00 01 10 11] with equal likelihood. Formally, this potential repertoire is represented by the maximum entropy or uniform distribution of possible system states, which expresses complete uncertainty, that is, [¼ ¼ ¼ ¼]. However, as soon as the mechanism is engaged and the detector enters a particular state (in this case x_1 ¼ [11]), uncertainty is reduced: the fact that the mechanism has led to state [11] specifies that the previous system state x_0 must have been either [11 or 10], rather than [00 or 01] (in this system, there is no mechanism to specify the detector state, which remains uncertain). Formally, the mechanism leading to state 11 specifies an actual distribution or repertoire of system states that could have caused (led to) x_1, while ruling out (giving probability zero to) states that could not. In this way, the mechanism leading to state 01 generates information, in the classic sense of reducing uncertainty or ignorance. More precisely, it generates 1 bit of information by distinguishing between things being one way ([11 or 10], which remain indistinguishable to it) rather than another way ([00 or 01], which also remain indistinguishable to it). The corresponding probability distribution is [½ ½ 0 0].

In general, the information generated when a system characterized by a certain mechanism enters a particular state, can be measured by the relative entropy H between the actual and the potential repertoires ('relative to' is indicated by k), also known as effective information (ei):

$$ei(X_0 \;!\; x_1) = H(p(X_0 \;!\; x_1) \,\|\, p^{max}(X_0 \;!\; x_1))$$

Relative entropy is a difference between probability distributions: if the distributions are identical, relative entropy is zero; the more different they are, the higher the relative entropy. Figuratively, the mechanism leading to state [01] generates information by shaping the uniform distribution into a less uniform one – this is how much uncertainty is reduced. Though effective information is completely specified once the mechanism and output state are specified, to calculate it in practice one can perturb the system in all possible ways (i.e., try out all possible input states, corresponding to the maximum entropy distribution or potential repertoire) while keeping track of the resulting actual repertoire using Bayes' rule. Clearly, the amount of effective information generated by a system is high if it has a large potential repertoire and a small actual repertoire, since a large number of initial states are ruled out. By contrast, the information generated is little if the system's repertoire is small, or if many states lead to the current outcome, since few states are ruled out. For instance, if the mechanism is dominated by noise, any state could have led to the current state: no alternative is ruled out, and no information is generated.

Integration

Second, we need to find out how much of the information generated by a system is integrated information – that is, how much information is generated by a single entity, as opposed to a collection of independent parts. The key idea here is to consider the parts of the system independently, ask how much information they generate by themselves, and compare it with the information generated by the system as a whole.

This can be done by resorting again to relative entropy to measure the difference between the probability distribution generated by the system as a whole (actual of the system) with the probability distribution generated by the parts considered independently (the product of the actual repertoire of the parts). Integrated information is indicated with the symbol F (the vertical bar 'I' stands for information, the circle 'O' for integration:

$$F(x_1) = H\left[p(X_0 \;!\; x_1) \,\|\, \prod_{M^k \in p^{min}} p\left(M_0^k \;!\; \mu_1^k\right)\right]$$

That is, the actual repertoire for each part is specified by causal interactions internal to each part, considered as a system in its own right, while external inputs are treated as a source of extrinsic

noise. The comparison is made with the particular decomposition of the system into parts that leaves the least information unaccounted. This minimum information partition or p^{min} decomposes the system into its minimal parts.

To see how this works, consider two of the million photodiodes in the digital camera. By turning on or off depending on its input, each photodiode generates 1 bit of information, just as we saw before. Considered independently, then, two photodiodes generate 2 bits of information, and 1 million photodiodes generate 1 million bits of information. However, the product of the actual distributions generated independently by the parts is identical to the actual distribution for the system. For example, assuming the two photodiodes enter the same state, their independent distributions would both be [½ ½ 0 0]. The product of these two distributions would be [¼ ¼ 0 0 0 0 0 0 0 0 0 0 0 0], which in this case is of course identical to the actual distribution over the 16 states of the four units considered together. Therefore, the relative entropy between the two distributions is zero: the system generates no integrated information ($\Phi(x_1) = 0$) above and beyond what is generated by its parts.

Clearly, for integrated information to be high, a system must be connected in such a way that a lot of information is generated by causal interactions among its elements, rather than within its parts. Thus, a system can generate integrated information only to the extent that it cannot be decomposed into informationally independent parts. If the elements of a system are properly connected, the interaction between the minimal parts of the system can generate information above and beyond what is accounted for by the parts by themselves, and $\Phi(x_1) > 0$. This can be achieved, for instance, by ensuring that each unit is connected to other units in a different way, and yet all units can interact.

In short, integrated information captures the information generated by causal interactions in the whole, over and above the information generated independently by the parts.

Complexes

Finally, by measuring Φ values for all subsets of elements within a system, we can determine which subsets form complexes. Specifically, a complex X is a set of elements that generate integrated information ($\Phi > 0$) that is not contained in some larger set of higher Φ. A complex, then, can be properly considered to form a single entity having its own, intrinsic 'point of view' (as opposed to being treated as a single entity from an outside, extrinsic point of view). Since integrated information is generated within a complex and not outside its boundaries, experience is necessarily private and related to a single point of view or perspective. A given physical system, such as a brain, is likely to contain more than one complex, many small ones with low Φ values, and perhaps a few large ones. In fact, at any given time there may be a single main complex of comparatively much higher Φ that underlies the dominant experience (a main complex is such that its subsets have strictly lower Φ). In general, a main complex can be embedded into larger complexes of lower Φ: a complex can be casually connected, through ports-in and ports-out, to elements that are not part of it. According to the IIT, such elements can influence indirectly the state of the main complex without contributing directly to the conscious experience it generates.

Neurobiology: Accounting for Empirical Observations

Can this approach account, at least in principle, for some of the basic facts about consciousness that have emerged from decades of clinical and neurobiological observations? Measuring Φ and finding complexes is not easy for realistic systems, but it can be done for simple networks that bear some structural resemblance to different parts of the brain. For example, by using computer simulations, it is possible to show that high Φ requires networks that conjoin functional specialization (due to its specialized connectivity, each element has a unique functional role within the network) with functional integration (there are many pathways for interactions among the elements). In very rough terms, this kind of architecture is characteristic of the mammalian corticothalamic system: different parts of the cerebral cortex are specialized for different functions, yet a vast network of connections allows these parts to interact profusely.

And indeed, as much neurological evidence indicates, the corticothalamic system is precisely the part of the brain which cannot be severely impaired without loss of consciousness.

Conversely, Φ is low for systems that are made up of small, quasi-independent modules. This may be why the cerebellum, despite its large number of neurons, does not contribute much to consciousness: its synaptic organization is such that individual patches of cerebellar cortex tend to be activated independently of one another, with little interaction between distant patches.

Computer simulations also show that units along multiple, segregated incoming or outgoing pathways are not incorporated within the repertoire of the main complex. This may be why neural activity in afferent pathways (perhaps as far as V1), though crucial for triggering this or that conscious experience, does not contribute directly to conscious experience; nor does activity in efferent pathways (perhaps starting with primary motor cortex), though it is crucial for reporting each different experience.

The addition of many parallel cycles also generally does not change the composition of the main complex, although Φ values can be altered. Instead, cortical and subcortical cycles or loops implement specialized subroutines that are capable of influencing the states of the main corticothalamic complex without joining it. Such informationally insulated cortico-subcortical loops could constitute the neural substrates for many unconscious processes that can affect and be affected by conscious experience, such as those that enable object recognition, language parsing, or translating our vague intentions into the right words.

Other simulations show that the effects of cortical disconnections are readily captured in terms of integrated information: a 'callosal' cut produces, out of large complex corresponding to the connected corticothalamic system, two separate complexes, in line with many studies of split brain patients. However, because there is great redundancy between the two hemispheres, their Φ value is not greatly reduced compared to when they formed a single complex. Functional disconnections may also lead to a restriction of the neural substrate of consciousness, as is seen in neurological neglect phenomena, in psychiatric conversion and dissociative disorders, and possibly during dreaming and hypnosis. It is also likely that certain attentional phenomena may correspond to changes in the composition of the main complex underlying consciousness. The attentional blink, where a fixed sensory input may at times make it to consciousness and at times not, may also be due to changes in functional connectivity: access to the main corticothalamic complex may be enabled or not based on dynamics intrinsic to the complex. Computer simulations confirm that functional disconnection can reduce the size of a complex and reduce its capacity to integrate information. While it is not easy to determine, at present, whether a particular group of neurons is excluded from the main complex because of hard-wired anatomical constraints, or, is transiently disconnected due to functional changes, the set of elements underlying consciousness is not static, but form a 'dynamic complex' or 'dynamic core.'

The most common example of a marked change in the level of experience is the fading of consciousness that occurs during certain periods of sleep. Subjects awakened in deep nonrapid eye movement (NREM) sleep, especially early in the night, often report that they were not aware of themselves or of anything else, though cortical and thalamic neurons remain active. Awakened at other times, mainly during REM sleep or during lighter periods of NREM sleep later in the night, they report dreams characterized by vivid images. From the perspective of integrated information, a reduction of consciousness during early sleep would be consistent with the extreme bistability of cortical circuits that occurs during deep NREM sleep: due to changes in intrinsic and synaptic conductances triggered by neuromodulatory changes (e.g., low acetylcholine), cortical neurons cannot sustain firing for more than a few hundred milliseconds, and invariably enter a hyperpolarized down-state. Shortly afterwards, they inevitably return to a depolarized up-state. Indeed, computer simulations show that values of Φ are low in systems with such a bistable dynamics. Consistent with these observations, studies using TMS in conjunction with high-density EEG show that early NREM sleep is associated either with a breakdown of the effective connectivity among cortical areas, and thereby with a loss of integration, or with a

stereotypical global response suggestive of a loss of repertoire and thus of information. Similar changes are seen in animal studies of anesthesia.

Computer simulations also indicate that the capacity to integrate information is reduced if neural activity is extremely high and near-synchronous, due to a dramatic decrease in the repertoire of discriminable states. This reduction in degrees of freedom could be the reason why consciousness is reduced or eliminated in absence seizure and other conditions during which neural activity is both high and synchronous.

Finally, consciousness not only requires a neural substrate with appropriate anatomical structure and appropriate physiological parameters: it also needs time. The theory predicts that the time requirement for the generation of conscious experience in the brain emerge directly from the time requirements for the build-up of an integrated repertoire among the elements of the corticothalamic main complex so that discriminations can be highly informative. To give an obvious example, if one were to perturb half of the elements of the main complex for less than a millisecond, no perturbations would produce any effect on the other half within this time window, and Φ would be zero. After say 100 ms, however, there is enough time for differential effects to be manifested, and Φ should grow.

Specifying the Quality of Consciousness

If the amount of integrated information generated by different brain structures or by the same structure functioning in different ways can in principle account for changes in the level of consciousness, what is responsible for the quality of each particular experience? What determines that colors look the way they do, and different from the way music sounds? Once again, empirical evidence indicates that different qualities of consciousness must be contributed by different cortical areas. Thus, damage to certain parts of the cerebral cortex forever eliminates our ability to experience color (whether perceived, imagined, remembered or dreamt), whereas damage to other parts selectively eliminates our ability to experience visual shapes. There is obviously something about different parts of the cortex that can account for their different contribution to the quality of experience. What is this something?

The IIT claims that just like the quantity of consciousness generated by a complex of elements is determined by the amount of integrated information it generates above and beyond its parts, the quality of consciousness is determined by the set of all the informational relationships its mechanisms generate. That is, how integrated information is generated within a complex determines not only the amount of consciousness it has, but also what kind of consciousness.

Consider again the photodiode thought experiment. As we discussed before, when the photodiode reacts to light, it can only tell that things are one way rather than another way. On the other hand, when we see 'light,' we discriminate against many more states of affairs, and thus generate much more information. In fact, we argued that 'light' means what it means and becomes conscious 'light' by virtue of being not just the opposite of dark, but also different from any color, any shape, any combination of colors and shapes, any frame of every possible movie, any sound, smell, thought, and so on.

What needs to be emphasized at this point is that discriminating 'light' against all these alternatives implies not just picking one thing out of 'everything else' (an undifferentiated bunch), but distinguishing at once, in a specific way, between each and every alternative. Consider a very simple example: a binary counter capable of discriminating among the four numbers: [00 01 10 11]. When the counter says binary '3,' it is not just discriminating [11] from everything else as an undifferentiated bunch, otherwise it would not be a counter, but a [11] detector. To be a counter, the system must be able to tell [11] apart from [00] as well as from [10] as well as from [01] in different, specific ways. It does so, of course, by making choices through its mechanisms, for example: is this the first or the second digit? Is it a 0 or a 1? Each mechanism adds its specific contribution to the discrimination they perform together. Similarly, when we see light, mechanisms in our brain are not just specifying 'light' with respect to a bunch of undifferentiated alternatives. Rather, these mechanisms are specifying that light is what it is by

virtue of being different, in this and that specific way, from every other alternative, from dark to any color to any shape, movie frame, sound or smell, and so on.

In short, generating a large amount of integrated information entails having a highly structured set of mechanisms that allow us to make many nested choices as a single entity. According to the IIT, these mechanisms working together generate integrated information by specifying a set of informational relationships that univocally determine the quality of experience.

The Shape of Experience

To see how this intuition can be given a mathematical formulation, let us consider again a complex of n binary elements that enters a particular state and makes a discrimination that generates a certain amount of integrated information F. Let us now suppose that each possible state of the system constitutes an axis or dimension of a qualia space (Q-space) having 2^n dimensions. Each axis is labeled with the probability p for that state, going from 0 to 1, so that a repertoire, that is, a probability distribution on the possible states of the complex, corresponds to a point in Q-space.

Let us now examine how the connections among the elements of the complex specify probability distributions, that is, how a set of causal interactions specifies a set of informational relationships. First, consider the complex with all connections among its elements disengaged, so that no causal interactions can occur. In this case, the system could not discriminate among any states, which would all be equally likely, corresponding to the maximum entropy or uniform distribution (the potential repertoire). In Q-space, this probability distribution is a point projecting onto all axes at $p = 1/2^n$. Next, consider engaging a single connection: by itself, each connection specifies an actual repertoire – it generates some information by shaping the uniform distribution into a more specific distribution – for instance, it may create some peaks (increase the probability of some states) and valleys (reduce the probability of other states, perhaps down to zero if it rules them out). In Q-space, this corresponds to a point projecting onto higher p values on some axes and onto lower p-values (or zero) on other axes. That is, the mechanism implemented by this single connection specifies an informational relationship (a relationship between two probability distributions). This informational relationship can be represented as an arrow in Q-space (q-arrow) that goes from the point corresponding to the maximum entropy distribution to the point corresponding to the actual repertoire specified by that connection. The length (divergence) of the q-arrow expresses how much the connection specifies the distribution (the effective information it generates, i.e., the relative entropy between the two distributions); the direction in Q-space expresses the particular way in which the connection specifies the distribution. Similarly, if one considers all other connections taken in isolation, each will specify another q-arrow of a certain length, pointing in a different direction.

Next, consider all possible combinations of connections. For instance, consider adding the contribution of the second connection to that of the first. Together, the first and second connections specify another actual repertoire – another point in Q-space – and thereby generate more information than either connection alone as they shape the uniform distribution into a more specific distribution. To the tip of the q-arrow specified by the first connection, one can now add a q-arrow bent in the direction contributed by the second connection, forming a path of two q-arrows in Q-space (the same final point is reached by adding the q-arrow due to the first connection on top of the q-arrow specified by the second one). Each combination of connection therefore specifies a q-path made of concatenated q-arrows (component q-arrows). In general, the more connections one considers together, the more the actual repertoire will take shape and differ from the uniform (potential) distribution.

Finally, consider the joint contribution of all connections of the complex. As was discussed above, all connections together specify the actual repertoire of the whole. This is the point where all q-paths made of q-arrows converge (one for every ordered permutation of all connections). Together, these converging paths in Q-space delimit a quale, that is, a shape in Q-space (a kind of 'solid,' or rather the 'body' of a polytope in more than three dimensions) whose edges are q-arrows, whose bottom is

the maximum entropy distribution, and whose top is the actual repertoire of the complex as a whole. The shape of this solid specifies all informational relationships that are generated within the complex by the interactions among its elements – also known as the effective information matrix.

It is worth considering briefly two relevant properties of informational relationships or q-arrows. First, informational relationships are context-dependent, in the following sense. A context can be any point in Q-space corresponding to the actual repertoire generated by a particular subset of connections. It can be shown that the q-arrow generated by considering the additional effects of a connection (how it further changes the actual repertoire) can change in both magnitude and direction depending on the context in which it is considered.

Another important property of q-arrows is entanglement g. A q-arrow is entangled ($g > 0$) if the underlying connections generate more information jointly than considered independently, that is, if it specifies information above and beyond its component q-arrows. An entangled q-arrow specifies a concept in that it groups together certain states of affairs in a way that cannot be decomposed into the mere sum of simpler groupings. Thus, entanglement characterizes informational relationships (q-arrows) that are more than the sum of their component relationships (component q-arrows), just like F characterizes systems that are more than the sum of their parts. Geometrically, entanglement 'warps' the shape of the *quale* away from a simple hypercube (where q-arrows are orthogonal to each other). Entanglement is important in identifying modes, just as F is useful in identifying complexes. By analogy with complexes, modes are sets of informational relationships in Q-space that have denser internal structure (are more tangled) than their surroundings. Modes can be considered as 'subshapes' in Q-space.

Some Properties of *Qualia* Space

What is the relevance of these constructs to understanding the quality of consciousness? It is not easy to become familiar with a complicated multidimensional space nearly impossible to draw, so it may be useful to resort to some metaphors. Perhaps the most important notion emerging from this approach is that an experience (a *quale* in the broad sense) is a shape in Q-space. What gives each experience its particular shape are the informational relationships in Q-space (q-arrows between repertoires) generated by causal interactions among the elements of a complex. Only the informational relationships within a complex (those that give the *quale* its shape) contribute to experience. Conversely, the informational relationships that exist outside the main complex, for example, those involving sensory afferents – do not make it into the *quale*, and therefore do not contribute either to the quantity or to the quality of consciousness.

By the same token, different experiences are, literally, different shapes in Q-space. For example, when the same system enters a different state (firing pattern), it will typically generate a different shape or *quale* (even for the same value of F). Moreover, experiences are similar if their shape is similar, and different to the extent that their shapes are different.

Note that a *quale* can only be specified by a mechanism entering a particular state – it does not make sense to ask about the *quale* generated by a mechanism in isolation, or by a state (firing pattern) in isolation. On the other hand, it does make sense to ask what kind of shapes or *qualia* the same system (mechanism) can generate when it enters different states. The set of all shapes generated by the same system when it enters different states provides a geometrical depiction of all its possible experiences.

Another consequence is that two systems entering the same state can generate two different experiences (i.e., two different shapes). In other words, the notion of a system state, or firing pattern for a neural system, is meaningless without taking into account the mechanism that produced the state. As an extreme example, a system that were to copy one by one the state of the neurons in a human brain, but had no internal connections of its own, would generate no consciousness and no *quale*. Note also that informational relationships are specified both by elements that are firing and others that are not. That is, a complex includes silent elements that are just as necessary in specifying the shape of the *quale* as are active ones. An intriguing corollary is that, within the main

complex, disabling elements (neurons) that happen to be silent should change the quantity and quality of experience, though the system state (firing pattern) remains the same.

It also follows that two systems with different architectures can generate the same experience (i.e., the same shape). For example, consider again the photodiode, whose mechanism determines that if the current in the sensor exceed a threshold, the detector turns on. This simple causal interaction is all there is, and when the photodiode turns on it merely specifies an actual repertoire where states [00 01 10 11] have, respectively, probability [0 0 ½ ½]. This corresponds in Q-space to a single q-arrow, one bit long, going from the potential, maximum entropy repertoire [¼ ¼ ¼ ¼] to [0 0 ½ ½]. Now imagine the light sensor is substituted by a temperature sensor with the same threshold and dynamic range – we have a thermistor rather than a photodiode. While the physical device has changed, according to the IIT the experience, minimal as it is, has to be the same, since the informational relationship that is generated by the two devices is identical. Similarly, an AND gate when silent and an OR gate when firing also generate the same shape in Q-space, and therefore must generate the same minimal experience (the two shapes have the same symmetries, i.e., are isomorphic). Thus, different 'physical' systems generate the same experience if the shape of the informational relationships they specify is the same. On the other hand, more complex networks of causal interactions are likely to create highly idiosyncratic shapes, so systems of high Φ are unlikely to generate exactly identical experiences.

It is important to see what Φ corresponds to in this representation. The minimum information partition is just another point in Q-space: the one specified by the connections within the minimal parts only, leaving out the contribution of the connections among the parts. This point is the actual repertoire corresponding to the product of the actual repertoires of the parts taken independently. Φ corresponds then to an arrow linking this point to the top of the solid. In this view, the q-paths leading to the minimum information bipartition provide the natural 'base' upon which the solid rests – the informational relationships generated within the parts upon which are built the informational relationships among the parts. The Φ-arrow can then be thought of as the height of the solid – or rather, to employ another metaphor, as the highest pole holding up a tent. For example, if Φ is zero (say a system decomposes into two independent complexes), the tent corresponding to the system is flat – it has no shape. Conversely, the higher the Φ value of a complex, the higher the tent or solid, the more 'breathing room' there is for the various informational relationships within the complex (the edges of the solid or the seams of the tent) to express themselves.

In summary, and not very rigorously, the generation of an experience can be thought of as the erection of a tent with a very complex structure: the edges are the tension lines generated by adding each connection in turn (the respective q-arrow or informational relationship). The tent literally takes shape when the connections are engaged and specify actual repertoires. Perhaps an even more daring metaphor would be the following: whenever the mechanisms of a complex unfold and specify informational relationships, the flower of experience blooms.

From Phenomenology to Geometry

The notions just sketched aim at providing a framework for translating the seemingly ineffable qualitative properties of phenomenology into the language of mathematics, specifically, the language of informational relationships (q-arrows) in Q-space. Ideally, when sufficiently developed, such language should permit the geometric characterization of phenomenological properties generated by the human brain. In principle, it should also allow us to characterize the phenomenology of other systems. After all, in this framework the experience of a bat echo-locating in a cave must be just another shape in Q-space and, at least in principle, shapes can be compared objectively.

At present, due to the combinatorial problems posed by deriving the shape of the *quale* produced by systems of just a few elements, and to the additional difficulties posed by representing such high-dimensional objects, the best one can hope for is to show that the language of Q-space can capture, in principle, some of the basic distinctions

that can be made in our own phenomenology, as well as some key neuropsychological observations. A short list includes the following:

1. Experience is divided into modalities, like the classic five senses (sight, hearing, touch, smell, taste and several others), as well as submodalities, like visual color and visual shape. What do these broad distinctions correspond to in Q-space? According to the IIT, modalities and submodalities are sets of densely entangled q-arrows (modes) that form distinct subshapes in the quale. As a two-dimensional analogue, imagine a given multimodal experience as the shape of the three-continent complex constituted by Europe, Asia, and Africa. The three continents are distinct subshapes, yet they are all part of the same landmass, just like modalities are parts of the same consciousness. Moreover, within each continent there are peninsulas (sub-subshapes), like Italy in Europe, just like there are submodalities within modalities.
2. Some experiences are homogeneous and others are composite: for example, a full-field experience of blue, as when watching a cloudless sky, compared to that of a market street. In Q-space, though not unimodal, homogeneous experiences translate to a single homogeneous shape, and composite ones into a composite shape with many distinguishable subshapes (modes and submodes).
3. Some experiences appear to be 'primitives' that cannot be further decomposed. A typical example are what philosophers call a 'quale' in the narrow sense, say a pure color like red, or a pain, or an itch: it is difficult, if not impossible, to identify any further phenomenological 'structure' within the experience of red. According to the IIT, such primitives correspond to subshapes (modes) in Q-space that cannot be further decomposed into sub-subshapes (primitive modes).
4. Some experiences are hierarchically organized. Take seeing a face: we see at once that as a whole it is somebody's face, but we also see that it has parts such as hair, eyes nose and mouth, and that those are made in turn of specifically oriented segments. It can be shown that modes in Q-space lend themselves readily to hierarchical decompositions, where part–whole relationships are represented by entangled informational relationships.
5. We recognize intuitively that the way we perceive taste, smell, and maybe color, is organized phenomenologically in a 'categorical' manner, quite different from, say, the 'topographical' manner in which we perceive space in vision, audition, or touch. According to the IIT, these hard to articulate phenomenological differences correspond to different basic shapes in Q-space, such as 2^n-dimensional grid-like structures and pyramid-like structures, which emerge naturally from the underlying neuroanatomy.
6. Some experiences are more alike than others. Blue is certainly different from red (and irreducible to red), but clearly it seems even more different from middle C on the oboe. In the IIT framework, in Q-space colors correspond to different subshapes of the same kind (say pyramids pointing in different directions) and sounds to very different subshapes (say tetrahedra). In principle, such subjective similarities and differences can be investigated by employing objective measures of similarity between shapes (e.g., considering the number and kinds of symmetries involved in specifying shapes that are generated in Q-space by different neuroanatomical circuits).
7. Experiences can be refined through learning and changes in connectivity. Say one learns to distinguish wine from water, then reds from whites, then different varietals. Presumably, underlying this phenomenological refinement is a neurobiological refinement: neurons that initially were connected indiscriminately to the same afferents, become more specialized and split into subgroups with partially segregated afferents. This process has a straightforward equivalent in Q-space: the single q-arrow generated initially by those afferents splits into two or more q-arrows pointing in different directions, and the overall subshape of the quale is correspondingly refined.
8. Qualia in the narrow sense (experiential primitives) exist 'at the top of experience' and not at its bottom. Consider the experience of seeing a pure color, such as red. The evidence suggests that the 'neural correlate' of color, including

red, is probably a set of neurons and connections in the fusiform gyrus, maybe in area V8 (ideally, neurons in this area are activated whenever a subject sees red and not otherwise, if stimulated trigger the experience of red, and if lesioned abolish the capacity to see red). Certain achromatopsic subjects with dysfunctions in this general area seem to lack the feeling of what it is like to see color, its 'coloredness,' including the 'redness' of red. They cannot experience, imagine, remember and even dream of color, though they may talk about it, just as we could talk about echolocation, from a third person perspective. Contrast such subjects, who are otherwise perfectly conscious, with vegetative patients, who are, for all intents and purposes, unconscious. Some of these patients may show behavioral and neurophysiological evidence for residual function in an isolated brain area. Yet it seems highly unlikely that a vegetative patient with residual activity exclusively in V8 should enjoy the vivid perceptions of color just as we do, while being otherwise unconscious.

The IIT provides a straightforward account for this difference. To see how, call '*r*' the connections targeting the 'red' neurons in V8 that confer them their selectivity, and non-*r* ($\neg r$) all the other connections within the main corticothalamic complex. Considering *r* in isolation at the bottom of *Q*-space (null context), yields a small *q*-arrow pointing in a direction that represents how *r* by itself shapes the maximum entropy distribution into an actual repertoire. Schematically, this situation resembles that of a vegetative patient with V8 and its afferents intact but the rest of the corticothalamic system destroyed. The shape of the experience or *quale* reduces to this *q*-arrow, so its quantity is minimal (*F* for this *q*-arrow is obviously low) and its quality unspecified: as we have seen with the photodiode, *r* by itself cannot specify whether the experience is a color rather than something else, such as a shape, whether it is visual or not, sensory or not, and so on.

By contrast, subtract *r* from the top of the complex (the set of all connection). This 'lesion' collapses all *q*-arrows generated by *r* starting from any context, called *aq-fold*. In particular, the lesion collapses the *q*-arrow, called the upset of nonred, which starts from the full context provided by all other connections *r* and reaches the top of the *quale*. This *q*-arrow will typically be much longer and point in a different direction than the *q*-arrow generated by *r* at the bottom of the *quale* because, the fuller the context, the more *r* can shape the actual repertoire. Schematically, removing *r* from the top, a situation resembling that of an achromatopsic patient with a selective lesion of V8, leaves the bulk of the experience or *quale* intact (*F* remains high), but collapses a noticeable feature of its shape (the upset of nonred). According to the IIT, the upset of nonred captures precisely the quality or 'redness' of red.

It is worth remarking that the last example also shows why specific qualities of consciousness, such as the 'redness' of red, while generated by a mechanism, cannot be reduced to a mechanism. If an achromatopsic subject without the *r* connections lacks precisely the 'redness' of red, whereas a vegetative patient with just the *r* connections is essentially unconscious, then the redness of red cannot map directly to the mechanism implemented by the *r* connections. However, the redness of red can map nicely onto the informational relationships specified by *r*, as these change dramatically between the null context (vegetative patient) and the full context (achromatopsic subject).

To conclude, it is worth pointing out some outstanding issues that will need to be addressed in further developments of the theory. One of these is finding a principled way to determine the proper spatial and temporal scale to measure informational relationships and integrated information. What are the elements upon which probability distributions of states are to be evaluated? For example, are they neurons or minicolumns? Similarly, what is the 'clock' to use to identify system states? Does it run in milliseconds or hundreds of milliseconds? A working hypothesis is that the relevant spatial and temporal scales are those that jointly maximize *F* – different systems will generate maximal amount of integrated information at a particular spatiotemporal scale that is determined by their mechanism.

Another important issue has to do with the relationship between complexes and the outside world.

The mechanisms of a complex generate integrated information and informational relationships from within – as shown by the dreaming brain, an adult brain does not need the outside world to generate experience. However, the mechanisms inside the complex are what they are, and so is the quality of the experience they generate, by virtue of a long evolutionary history, individual development, and learning. In fact, it appears that as a system incorporates statistical regularities from its environment and learns, its capacity for integrated information may grow. It will thus be important to see how the informational relationships (q-arrows) inside a complex reflect and react to informational relationships existing in the world.

Conclusion

To recapitulate, the IIT claims that the quantity of consciousness is given by the integrated information (Φ) generated by a complex of elements, and its quality by the shape in Q-space specified by all the informational relationships they generate. As suggested here, this theoretical framework can account, at least in principle, for several basic neurobiological and neuropsychological observations. Moreover, the same theoretical framework can be extended to translate phenomenology into the language of mathematics.

At present, the very notion of a theoretical approach to consciousness may appear far-fetched, yet the nature of the problems posed by a science of consciousness seems to require a combination of experiments and theories: one could say that theories without experiments are lame, but experiments without theories are blind. Only a genuine theoretical framework can go beyond proposing a provisional list of candidate mechanisms (reentry, synchronization, broadcasting) or brain areas (frontoparietal networks, default system) without a principled explanation of why they may be relevant. And only a theory can account, in a coherent manner, for several key but puzzling facts about consciousness and the brain, such as the association of consciousness with the corticothalamic but not the cerebellar system, the 'unconscious' functioning of many cortico-subcortical circuits or the fading of consciousness during certain stages of sleep or epilepsy despite continuing neural activity.

A theory should also generate relevant corollaries. For example, the IIT: predicts that consciousness depends exclusively on the ability of a system to generate integrated information, whether or not it is immersed in an environment, it has language, capacity for reflection, attention, episodic memory, a sense of space, of the body, and of the self, contrary to some common intuitions, but consistent, as reviewed elsewhere, with the overall neurological evidence. Of course, the theory recognizes that these same factors are important historically because they favor the development of neural circuits forming a main complex of high Φ.

Finally, a theory should be able to help in cases that challenge our intuition or our standard ways to assess consciousness. For instance, the IIT says that the presence and extent of consciousness can be determined, in principle, also in cases in which we have no verbal report, such as infants or animals, or in neurological conditions such as minimally conscious states, akinetic mutism, psychomotor seizures, and sleepwalking. In practice, of course, measuring Φ accurately in such systems will not be easy, but approximations and informed estimates are certainly conceivable. The theory also implies that consciousness is not an all-or-none property, but is graded: specifically, it increases in proportion to a system's repertoire of discriminable states. In fact, any physical system with some capacity for integrated information would have some degree of experience, irrespective of the constituents of which it is made, and independent of its ability to report. Whether these and other predictions turn out to be compatible with future clinical and experimental evidence, a coherent theoretical framework should at least help to systematize a number of neuropsychological and neurobiological results that might otherwise seem disparate.

See also: Neurobiological Theories of Consciousness.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. New York, NY: Cambridge University Press.
Balduzzi D and Tononi G (2008) Integrated information in discrete dynamical systems: Motivation and theoretical

framework. PLoS Computational Biology 13;4(6): e1000091.
Cover TM and Thomas JA (2006) Elements of Information Theory, 2nd edn. Hoboken, NJ: Wiley-Interscience.
Crick F and Koch C (2003) A framework for consciousness. Nature Neuroscience 6(2): 119–126.
Dehaene S, Sergent C, and Changeux JP (2003) A neuronal network model linking subjective reports and objective physiological data during conscious perception. Proceedings of the National Academy of Sciences USA 100(14): 8520–8525.
Feldman J (2003) A catalog of Boolean concepts. Journal of Mathematical Psychology 47(1): 75–89.
Gazzaniga MS (2005) Forty-five years of split-brain research and still going strong. Nature Reviews Neuroscience 6(8): 653–659.
Hobson JA, Pace-Schott EF, and Stickgold R (2000) Dreaming and the brain: Toward a cognitive neuroscience of conscious states. Behavioral Brain Science 23(6): 793–842; discussion 904–1121.
Koch C and Tononi G (2008) Can machines be conscious? Spectrum IEEE 45(6): 55–59.
Massimini M, Ferrarelli F, Esser SK, et al. (2007) Triggering sleep slow waves by transcranial magnetic stimulation. Proceedings of the National Academy of Sciences USA 104(20): 8496–8501.
Massimini M, Ferrarelli F, Huber R, Esser SK, Singh H, and Tononi G (2005) Breakdown of cortical effective connectivity during sleep. Science 309(5744): 2228–2232.
Posner JB and Plum F (2007) Plum and Posner's Diagnosis of Stupor and Coma. 4th edn. Oxford; New York: Oxford University Press.
Steriade M, Timofeev I, and Grenier F (2001) Natural waking and sleep states: A view from inside neocortical neurons. Journal of Neurophysiology 85(5): 1969–1985.
Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5(1): 42.
Tononi G and Edelman GM (1998) Consciousness and complexity. Science 282(5395): 1846–1851.
Tononi G and Laureys S (2009) The neurology of consciousness: An overview. In: Laureys S and Tononi G (eds.) The Neurology of Consciousness. Elsevier.
Tononi G and Sporns O (2003) Measuring information integration. BMC Neuroscience 4(1): 31.
Tononi G, Sporns O, and Edelman GM (1996) A complexity measure for selective matching of signals by the brain. Proceedings of the National Academy of Sciences USA 93(8): 3422–3427.
van Zandvoort MJ, Nijboer TC, and de Haan E (2007) Developmental colour agnosia. Cortex 43(6): 750–757.

Biographical Sketch

Intentionality and Consciousness

R Menary, The University of Wollongong, Wollongong, NSW, Australia

Glossary

Consciousness – Can be usefully classified in three ways: (1) phenomenal consciousness, the way things feel to us; (2) access consciousness, a mental state is poised for use in reasoning and direct rational control of thought and action; and (3) reflexive consciousness, the ability to think about one's own thoughts.

Embodied mind – The view that consciousness and intentionality should be understood in terms of our bodily engagements with the world and not by bracketing off the external world.

Intentionality – The directedness of mental states, captured by the phrase what are you thinking about?

Naturalism – The attempt to give an explanation of intentionality, consciousness, and representation in terms that are consistent with the natural sciences.

Representation – States that present, or stand in for an object. Most theories take a representation to have a content that can be specified by a proposition such as ''the Earth is spherical.''

Introduction: Mind, Consciousness, and Intentionality

Conscious minds have both intentional and phenomenal properties. If a mental state is intentional then it is directed toward some object (it is about something). If a mental state is phenomenal then it feels a particular way to us. Cognitive states such as beliefs and desires are obviously intentional. My belief that there is a cold beer on the counter in front of me is directed at an object (the cold beer); my desire to drink that beer is also aimed at an object (the very same beer). My visual perception of the beer on the counter is also intentionally directed as is the tactile sensation of grasping the cold bottle in my hand. The tactile sensation also feels a certain way; the bottle feels cold. Hence some conscious experiences are both intentionally directed and feel a particular way.

One way to think of the relationship between consciousness and intentionality is that all conscious states are also intentional. When I am consciously aware of a sensation (a sound), then this experience is certainly phenomenal; but the experience is also intentional – because the sensation (the sound) is the object or content of my awareness. Many philosophers and cognitive scientists take consciousness to be the defining property of the mental, and they distinguish between consciousness as perceptual awareness (awareness of sights, sounds, smells, etc., which we share with some other animals) and consciousness as awareness of the contents of my experiences and thoughts, as 'my' experiences and thoughts.

Cognitive states such as beliefs and desires, and perceptual states such as looking at a blue mug or hearing a high C are intentional; they are directed at something other than themselves. Perceptions are intentional in so far as I am aware of them as having sensory objects (at which they are directed), light, sound, etc. Other phenomenal states such as itches and pains are less obviously intentional, when I am aware of my pain then the object of my awareness is the pain. It is not so obvious what the object or content of the pain itself is, consequently some thinkers deny that phenomenal feels (or *qualia*) such as pains and itches are intentional.

Conscious awareness of 'my' cognitive states, perceptual states, and phenomenal feels is intentional and this has been recognized for a long time. Take, for example, the following section of Descartes' second meditation:

> Finally it is I who has sensations, that is to say, who is aware of objects as though by the senses, since indeed

> I see light, I hear noise, I feel heat. But all these objects are unreal, since I am dreaming. Let it be so; certainly it seems to me that I see, I hear, and I feel heat. That cannot be false; that is what in me is properly called sensation; and in this precise sense, sensation is nothing but thought.

By contrast, Towser the dog may 'desire' the bone and may 'believe' that the bone is under the bush. But Towser does not think "I am thinking that there is a bone under the bush." Towser may have beliefs and desires, or equivalent intentional states, but is not conscious of them as beliefs and desires and certainly not as being his beliefs and desires.

We can, at this point, distinguish between intentional states, which are cognitive/conceptual such as beliefs and desires and intentional states, which are perceptual/nonconceptual such as sights and sounds. There are those such as McDowell who do not wish to make this distinction, claiming that there is only conceptual intentionality. Conceptual content can be thought of as the content of a thought that can be propositionally described. My thought that the beer in front of me is ice cold is just such a thought with conceptual content.

There are strong reasons to favor an explanation of mindedness as being both intentional and conscious (especially with the capacity for self-awareness). However, it is not so clear whether consciousness should be understood in terms of intentionality (some conscious states might not be intentional), nor whether intentionality should be understood in terms of consciousness (Towser's cognitive states are intentional, but not conscious).

What then is the relationship between intentionality and consciousness? Is intentionality a prerequisite for consciousness or is consciousness a prerequisite for intentionality? Opinions, unsurprisingly, differ on these questions. There are further questions about the relationship between intentionality and consciousness. If Brentano is right and intentionality is the hallmark of the mental, then presumably all conscious experiences must be intentional. If some conscious experiences are not intentional, it would appear to follow that they are also not mental. It would appear to be absurd to endorse the view that some of our conscious experiences are not mental phenomena.

However, as noted above, it is not clear that all conscious experiences are describable as having conceptual content. For example, pains are conscious experiences, in what sense might they be thought of as having conceptual contents? Some philosophers, such as Sellars, think that phenomenal feels are not intentional. If we followed Sellars in this then we would mark off conscious states that are intentional, such as beliefs, desires, and perceptions from those that are nonintentional, such as itches and tickles because they do not have content.

Furthermore, it may turn out that consciousness outstrips intentionality and conversely that some intentional states, such as representations, might not be accessible to consciousness (although both of these claims are controversial). Consequently the concepts of intentionality and consciousness may overlap at various points, but are not mutually defining: Not all conscious states are intentional; not all intentional states are conscious. These issues will be explored in sections 'Intentional inexistence,' 'Representational theories of consciousness,' and 'Is phenomenal consciousness intentional?.'

Finally there is the question of whether intentionality and consciousness can be talked about independently of, or bracketed off from, the body and its relationships with the world. Increasingly research in cognitive science is turning to an embodied and embedded approach to intentionality and consciousness, rather than an approach which solipsistically considers only states internal to the mind or brain. In the embodied approach both intentionality and consciousness become conceived of in terms of our embodied engagements with the world.

Therefore, it is still an open and important question to ask what the relationship between intentionality and consciousness is. Before considering this question I shall give more precise definitions of intentionality and consciousness in the next two sections entitled 'Intentionality' and 'Consciousness.' Then I shall go on to explore some of the puzzles that are produced by Brentano's theses about intentionality and then tackle the issue of whether phenomenal consciousness is intentional. Finally I shall look at the phenomenological tradition and the embodied approaches to intentionality that draw upon this tradition, as well as the externalism of the analytic tradition.

Intentionality

Intentionality is usually defined as the directedness of the mind toward something other than itself. I can be thinking about the ice cold beer sitting in front of me, I might want the beer to be ice cold, or hope that it is. The mental states of wanting and hoping, in this case, are about the beer sitting before me. Consequently, another definition of intentionality is the aboutness of mental states such as hopes, desires, and beliefs; they are about something other than themselves. The notion of intentionality can be neatly captured by the everyday question "what are you thinking 'about'?" The term intentionality should not be confused with the normal English word intention, I intend to do something. This is despite the fact that both uses are derived from the Latin *intentio*, which literally means to apply tension, but also to point at or extend toward. Scholastics of the middle ages, such as Thomas Aquinas, used the term in both the more familiar sense of intending to do something and in the restricted technical fashion (direction toward an object) that is still current today.

Intentionality becomes interesting when we look at thoughts that are not about objects which are within our perceptual range. For example, I might be sitting in the office thinking about the ice cold beer that I will quaff with gusto at the end of the day. My thoughts are directed at an object that is not sitting before me. My thoughts might also turn to objects that don't, nor ever will exist such as phlogiston, or Santa Claus or Hamlet. It is quite normal for us to think about nonexistent objects, but what then are my thoughts directed at? This is a puzzle that any account of intentionality must solve, as well as the, more straightforward, intentional phenomenon of directedness at existing objects.

The nineteenth century psychologist and philosopher Franz Brentano gave the formulation of intentionality, which underpins contemporary thinking in philosophy and the cognitive sciences today. Brentano claimed that intentionality was the defining feature of the mental, what he called the mark of the mental; all minds are intentional. Furthermore, he claimed that nothing physical exhibited intentionality; there are no physical objects which are intentional. He concluded that intentionality was the feature of mentality, which allows us to differentiate between minds and physical phenomena and that minds were not physical entities. It is one of the central tasks of contemporary philosophy, psychology, neuroscience, and artificial intelligence (what might be collectively described as the cognitive sciences) to explain how physical entities can have minds, or more precisely how physical systems can also be intentional systems.

Consciousness

Ned Block has distinguished three different senses of consciousness. The first is phenomenal consciousness (P-consciousness), the familiar sense of conscious experience as what it is like. Conscious experiences have a phenomenal character, or to put it another way, they have sensory qualities (otherwise known as *qualia*). There is something it is like for me to see the deep blue of the Pacific Ocean and to feel the warm water lapping over my feet, and to smell the briny breeze. The second is access consciousness (A-consciousness). A conscious state is A-conscious if it is poised for use in reasoning and direct rational control of thought and action. P- and A-consciousness although distinct will often interact: my becoming A-conscious of the shirt on the back of my neck changes my P-conscious state. Finally there is reflexive or monitoring consciousness. This amounts to awareness of ones own conscious experiences and thoughts. Classically this was known as introspection and was often thought of as a kind of inner sense directed at conscious experiences, but more recently reflexive consciousness has been thought of as a kind of representation of the contents of consciousness, especially by Rosenthal and Dennett.

A major point of disagreement is whether phenomenal consciousness is intentional. This leads to a question about explanatory priority: Is phenomenal experience necessary for intentionality or are all phenomenal experiences simply the way in which intentional states (representations) are experienced? Before proceeding to deal with these questions in sections entitled 'Is phenomenal consciousness intentional' and 'Embodied intentionality,' I will outline some of the standard approaches to intentionality in the analytic tradition. I shall

introduce the phenomenological tradition in the final two sections entitled 'The phenomenological tradition' and 'Embodied intentionality.'

Intentional Inexistence

Brentano bequeathed us three important claims concerning intentionality:

1. The directedness of mental states: All mental states are directed at something beyond themselves.
2. Intentional inexistence: All mental states are directed at an object or content that is internal to the mind.
3. Brentano's thesis: Only mental states exhibit the two above properties, this marks off the mental from the physical.

The first two claims are clearly related and jointly provide reason to believe in the truth of the third. While the first claim is easy enough to grasp, what Brentano meant by intentional inexistence is not so clear and the section entitled 'Analytic approaches to intentionality' will be devoted to giving an account. Section entitled 'Analytic approaches to intentionality' will look at how the first two claims have been dealt with in the analytic tradition, stretching from Frege and Russell to contemporary philosophy of language and mind. Sections entitled 'Brentano's thesis' and 'Naturalizing representations and content' will look at the final claim, often referred to as Brentano's thesis and recent attempts to show that the thesis is false and that physical systems can also be intentional systems.

As noted above we can think about things that don't exist as well as things that do and this makes intentionality a difficult phenomenon to explain. Brentano attempted to explain this by claiming that mental states intentionally contain an object in themselves and this is what he called intentional inexistence. Mental states such as wanting, believing, hoping, and fearing are directed at objects, but these objects don't always exist, for example I want world peace, but the world is lamentably war torn. To deal with this Brentano claimed that mental states are not directed at actual existing objects out there in the world, but directed at an object or content internal to themselves. Mental states can still be about actual existing objects in the world, but only by being directed at mental contents. Thus, we find Brentano (in Psychology from an Empirical Standpoint) saying that the intentional relation is a quasirelation, because it is not a relation between two existing things – only the thinker's mind need exist. Therefore, Brentano's account of intentionality is a form of internalism, the intentional relation is entirely contained within the boundaries of the mind (this should be contrasted with externalism, see section entitled 'Naturalizing representations and content').

It is not entirely clear what Brentano meant by a mental state intentionally containing an object in itself, an object at which it is directed. More recently Roderick Chisholm has attempted to clarify what he could have meant. When Brentano says that mental states intentionally contain an object within themselves, those states truly have an object even if that object does not exist. So Diogenes' looking for an honest man would still have the same object (an honest man) even if there are no such things. However, this kind of relation to an object never occurs in nonpsychological phenomena; for example, for Diogenes to sit in his tub there must be a tub for him to sit in. As Chisholm points out although there is a relation between Diogenes and an object in both examples above, the psychological relation is of a peculiar sort because it can hold, even though one of its terms (an honest man) does not exist. Minds can be intentionally related to things that exist within the mind exclusively. This reduces the initial plausibility of the definition of intentionality that it is the direction of the mind toward something external to itself, for it now appears that the mind is directed only at something internal to itself. An alternative view is that there are intentional objects at which the mind is directed, including nonexistent ones. This was famously argued by one of Brentano's pupils, Meinong, and has been subject to stringent criticism, especially by Russell and Quine (see section entitled 'Analytic approaches to intentionality').

There are three puzzles that Brentano's account of intentionality give rise to

1. How can two different beliefs be about the same thing?
2. What are thoughts about nonexistent things about?
3. How can Intentionality be reconciled with a physicalist ontology?

Before looking at contemporary attempts to answer these questions (especially question three) we shall look at the historical antecedents in the analytic tradition.

Analytic Approaches to Intentionality

Early analytic philosophy has provided an answer to the first two questions from the end of the previous section entitled 'Intentional inexistence' and has provided the impetus for contemporary accounts of intentionality. Frege and Russell were primarily concerned with the logical structure of language, but since most analytic philosophers take language to be the expression of thought, it is also a method for studying the intentional structure of thought. The intentional structure of the propositional attitudes – hope that, fear that, believe that, etc. – can be revealed in an analysis of the sentences used to express them. One revealing claim is that sentences express, or have as their contents, the very same propositional contents of mental states (such as the propositional attitudes). Therefore my belief that Paris is the capital of France has as its content the very same proposition as my utterance: "I believe that Paris is the capital of France." They have the 'same' content (this is what I earlier called conceptual content).

We should, therefore, think of the contents of mental states and sentences as independent of those states and sentences. If we share the same belief – that the world is round – then this is not because our beliefs are numerically identical (one and the same); the identity is one of content, we believe the same proposition (that the world is round). Furthermore, the contents of singular thoughts such as beliefs are propositions, not sentences of a language. To play this role, propositions must be both abstract and objective. If I were to translate the belief sentence, (1) "I believe he is here," into German, I translate it as "ich glaube er ist hier." If beliefs merely have sentences as their contents (and not propositions) then I am saying (2) "I believe the sentence, 'he is here,'" which would be translated as "Ich glaube den satz 'he is here.'" But we would not translate in this way, (3) it should be translated as, "Ich glaube den satz 'er ist hier.'" The sentence is not what I believe (the sentence token "he is here"), but the meaning of the sentence, the proposition.

Frege explains this by introducing the distinction between sense and reference. The reference of a name is the object to which that name refers, for example: plant, chair, quark, etc. A classical example to illustrate the difference between sense and reference is the planet Venus. The Greeks believed that there were two different stars that rose in the morning and the evening – not a single planet, Venus. The reference of the words Hesperus and Phosphorus is the same object – Venus. However, the senses of Hesperus, the evening star, and Phosphorus, the morning star, are not the same. Words or expressions which have the same meaning, that is, are synonymous, are called 'intensionally equivalent;' and words or expressions which have the same reference or extension are called 'extensionally equivalent.' Take the following two statements: 'creature with a heart' and 'creature with a kidney,' both of these statements are extensionally equivalent, any creature with a heart is also a creature with a kidney. They are not, however, intensionally equivalent, they do not mean the same thing. Our examples show that words or expressions may be extensionally equivalent without being intensionally equivalent. The first puzzle now comes into play. Meno believes that Phosphorous rises in the morning and he believes that Hesperus rises in the evening. These are two different beliefs, they are not intensionally equivalent, yet they are extensionally equivalent – they are both about the planet Venus, how can this be the case?

For Frege the sense of a sentence is the 'proposition' it expresses, however, the reference of sentences is not simply a spatiotemporal object. This is because Frege held that the reference of a sentence was the truth – value of that sentence. Frege held that a thought is what is expressed in a proposition, and a proposition is a function with a value, which is always a truth value. Frege took his inspiration for this conclusion from mathematics. The reference of a whole expression, such as $X^2 + 2$, is its value. $X^2 + 2$ has as its value the number 6, when we fill in the variable X with the argument 2. The expressions $X^2 + 2 = 6$ and X is bald, have as their values 'the true,' when we fill in their variables with the arguments 2 and Socrates.

The sense of a sentence or proposition is the thought expressed by it. The reference of the proposition 'Paris is the capital of France' is its truth value, its sense is the thought that 'Paris is the capital of France.' Thoughts refer to truth values. Thoughts are not subjective. Thoughts and their relations to truth are objective and mind independent. They must be publicly accessible and objective, they are also abstract and non-physical, much like numbers, according to Frege.

The meaning of names and propositions is made up of two elements: their reference and their sense. Returning to the first puzzle (how can two different beliefs be about the same thing?) Frege presents a solution because Meno's beliefs have different senses, but the same reference. Or, to put it another way, each belief presents its referent differently (senses are modes of presentation) to the thinker (Meno).

In considering the second puzzle (what are thoughts about nonexistent things about?) we must turn to Frege's contemporary Bertrand Russell. Russell made a distinction between two ways of knowing: Knowing by acquaintance and knowing by description. When I know by acquaintance, I, the knowing subject, stands in a direct relation to some object of awareness. ''I say that I am 'acquainted' with an object when I have a direct cognitive relation to that object, for example, when I am directly aware of the object itself.'' When Russell says 'directly aware,' we should not think in terms of a direct theory of perception. This is because, the direct objects of acquaintance are things such as: Sense – data, memories (which are particulars); and concepts of redness or roundness (which are universals). Particulars and universals are, therefore, the only things that we can know by acquaintance. When we know something by description we are not directly acquainted with the object of knowledge, and descriptions are introduced by sentences of the form, ''a so-and-so'' or ''the so-and-so.'' Physical objects and other minds are not objects of acquaintance, so they must be known by description. The sentence form, ''a so-and-so,'' is an ambiguous description; and the sentence form, ''the so-and-so,'' is a definite description. As a consequence of Russell's theory, nouns and proper names are considered to be descriptions. This is because the logical form of proper names can only be revealed as a description.

In sentences that contain names and nouns that have no denotation (i.e., they are fictional), such as 'the present King of France is bald,' or 'the Unicorn is white,' the subject expression does not denote, yet the sentences are meaningful. Here our second puzzle enters the stage, what does the sentence (or thought) 'the present King of France is bald' denote?

Russell had earlier accepted Meinong's solution to this problem (Meinong was a pupil of Brentano's). Meinong held a realist theory of 'objects,' objects both exist and subsist. Those objects that exist form only a small class of objects, whereas, the class of the objects of knowledge, which is often nonexistent, is large and subsists; in the sense that they are the objects of thought and talk. This position requires that all possible negative facts subsist, such as England did not win the Soccer World Cup in 2006, and all impossible objects, such as, round squares. Consequently Meinong held a view that for any thought, it had an intentional object that either existed or subsisted – this was Meinong's way of explaining intentional inexistence. Russell came to believe that this Meinongian realism offended against his sense of reality. Russell devised the theory of descriptions to avoid Meinong's ontological profligacy, yet maintained the denotative theory of meaning. Therefore, most proper names and nouns are concealed descriptions.

Only logically proper names denote, so how can descriptions be meaningful if one accepts a denotative theory of meaning? Russell's answer was to apply a logical analysis to sentences that appeared to denote objects that do not in fact exist.

The two most famous examples that Russell gave of sentences containing definite descriptions were

1. The present King of France is bald.

Russell's analysis of this phrase will give us:

1a. There is a King of France.
1b. There is not more than one King of France, and
1c. Anything which is the King of France is bald.

$$\delta\exists x\flat\delta Fx \,\&\, \tfrac{1}{2}\delta\square y\flat\, \delta Fy \;!\; y \tfrac{1}{4} x\flat\square \,\&\, Gx\flat$$

1a. is false, therefore 1. is false. 1a. is defined in terms of the existential generalization (there is an x); 1b. gives the uniqueness condition as determined by the use of the definite article 'the' (i.e.,

'the King of France'); 1c. defines the existent x as having a particular property, baldness.

Sentence 1 is an example of what Russell meant when he claimed that the grammar of language can be misleading. It appears as if the description has a denotation, however, in the formal paraphrase there are no denoting singular terms, only variables bound by quantifiers (along with predicates and identity). Furthermore because 1a is false there is no need to suppose that there is an object to which 1 refers, it only appears that there is, because we expect there to be an extension of the subject of predication uniquely picked out by the definite article. In his definitive paper 'On Denoting,' Russell says:

> 'the present King of France is bald' is certainly false; and 'the present King of France is not bald' is false if it means 'There is an entity which is now King of France and is not bald' but it is true if it means 'It is false that there is an entity which is now King of France and is bald.'

However, Russell's analysis of definite descriptions has been criticized by Strawson and Donellan and recent theories of direct reference have superseded the account to some degree. Frege and Russell produced analyses that were supposed to solve some of the puzzles produced by Brentano's analysis of intentionality. They also began a century long investigation of intentionality and reference, which has taken a naturalistic turn in more recent philosophical work and this, leads us directly to the next puzzle.

Brentano's Thesis

Brentano's thesis is that intentionality marks off the mental from the physical. The property of intentional inexistence is a property to be found only in mental states and never in physical objects or processes. Chisholm brought this thesis to the attention of the late twentieth century philosophers in arguing that intentional states such as belief could not be accounted for simply in terms of behavior or behavioral dispositions. This is because such behavioral analyses are circular; they depend upon other intentional states. To be able to explain my behavior in terms of my desire for something I also have to explain it in terms of my beliefs about it, therefore if I desire a beer, my behavior is also explained in terms of my beliefs concerning what a beer is and where I can procure one. Chisholm concluded, in line with Brentano's thesis, that behavioral, reductive, or physicalist explanations of intentionality were doomed to failure.

Contemporary philosophers of mind and psychology have not agreed with Brentano and Chisholm's conclusion. They have taken up the challenge to provide a naturalistic or physicalist explanation of intentionality. One way to answer Brentano's challenge is simply to demonstrate that there are physical or nonmental entities that exhibit intentionality. One obvious contender is language, sentences of natural language can exhibit intentionality: they can be directed at something other than themselves. A reason for rejecting a linguistic explanation of intentionality has been proposed by John Searle and that is the distinction between original and derived intentionality. Sentences of language do not have any intrinsic meaning or content, they have meaning and content conferred upon them by people who use sentences to express their thoughts and it is these thoughts which intrinsically have meaning or content. Consequently the mental states expressed by sentences have original intentionality, whereas the sentences themselves have merely derived intentionality.

Another response has been proposed by Daniel Dennett who takes it that intentional idioms do not describe any actual phenomena, but they do have instrumental value for predicting the behavior of complex physical systems such as human beings. Although there are no actual intentional states or objects on Dennett's account, the intentional idiom of beliefs and desires and other intentional states is required, to be able to predict and explain the behavior of others, in place of a complex physical story about how our brains and bodies work.

An even more ambitious answer to the challenge is to demonstrate that there are physical systems which are nevertheless intentional systems. Fred Dretske has given a detailed information theoretical account, which shows that there are many natural states which indicate features of their local environment. This brings us to the representational account of intentionality.

Naturalizing Representations and Content

The Representational Theory of Mind (RTM) is an attempt to explain intentional mental states in terms of the concept of representation. As such,

intentional states such as beliefs, fears, hopes, and perceptions are inner states of people, perhaps brain states, which have representational content. The hope is that intentionality can be explained by the more familiar notion of representation, and that representations are better candidates for being physical states.

The RTM posits semantically evaluable representations that enter into causal connections with other representations and behavior. Beliefs and desires are semantically evaluable, beliefs can be true or false, desires can be satisfied or frustrated, etc. It is assumed that what makes beliefs true or false is their relation to the external world. So, we give beliefs and desires semantic evaluations just because we evaluate them in terms of their relation to the world. However, beliefs and desires are attitudes to 'something' and we can call this 'something' the belief 's content (what was above called conceptual content). It is the content, or proposition, (hence propositional attitude) that is semantically evaluable (here we see the continuity with the early analytic tradition exemplified by Frege and Russell). An example: Hamlet believes that his uncle killed his father. The belief has a semantic value, it is a true belief. The content of Hamlet's belief is that "Hamlet's uncle killed his father" (the content of the belief is the proposition, "Hamlet's uncle killed his father"). Knowing the content of a belief allows you to know what it is about the world that determines the semantic evaluation of the belief. Hence, intentional states are to be understood in terms of representation. I now turn to two attempts to give a naturalistic explanation of representation.

Dretske's indicator semantics

Dretske wants to build his account of representation from the bottom-up. We derive an account of representation and intentionality, from a natural, nonintentional, account of indicators and their causal relation to what they indicate.

Dretske begins with an analysis of natural signs or indicators. Tracks in the snow indicate the previous presence of an animal; the width of tree rings indicates the amount of rainfall there was in a year, etc. The indicator is dependent on the presence of the indicated, in other words, there would be no tracks in the snow (indicator) if the animal (indicated) had not been present. The dependency of indicator on indicated is causal, the indicator covaries with the indicated. Natural signs derive their indicative powers from the way in which they are objectively related to the conditions that they signify. The 24 rings of a tree stump indicate that the tree is 24 years old and the fact that the 24 rings indicate that the tree is 24 years old could not be the case if the tree was not 24 years old. The power of indication is dependent upon the relation between the indicator and what it indicates. The relation should be lawful or causal in character. Therefore, an indicator C indicates E, iff C indicates the presence of E only when E is present in the environment.

Hence, for Dretske there is no such thing as misindication. Indicators become representations, which can misrepresent when a representation is biologically supposed to indicate an environmental stimulus. The indicational content of a representation is explained in terms of the adaptive function of that representation.

A neural state N has the function of representing food when it indicates the presence of food in the environment, and it is recruited in the service of the function of the organism's moving toward and consuming the food. How might an internal state I acquire its function of indicating? Dretske is clear on this matter: I is recruited as a cause of some behavior B, because of what it indicates about a state of the environment E. Once I becomes a cause of B it acquires the function of indicating E – it comes to represent E. I becomes a representation when its natural meaning acquires an explanatory relevance. Dretske uses the example of northern hemisphere marine bacteria to illustrate the idea. These bacteria contain magnetosomes, which align the bacteria parallel to the Earth's magnetic field (the indicators). The bacteria are propelled downward, to the geomagnetic north, away from the oxygen-rich surface water, which is toxic for the bacteria (the indicators are recruited to help perform this function).

We must be clear about the notion of recruitment at work here. The recruitment involves a process of selection or reinforcement, which establishes a link between I and B, by virtue of the consequences of producing B in certain circumstances. In the bacteria example, recruitment of the

indicator is selected for to produce movement toward the geomagnetic north. However, for cases where there is a genuine belief, learning is the recruitment process. It is during this period of recruitment that I comes to represent E, I can now misrepresent E even though before the period of recruitment, when it was a mere indicator of E, I could not misrepresent E.

Millikan's biosemantics

Millikan's understanding of proper biological functions allows us to understand how there could be representation in the biological world. Millikan shows that the production and consumption of representational vehicles, what she calls intentional icons, are biological functions and the normativity of representations is derived from the normativity of biological functions, which she calls proper functions. Proper functions are normative, in the sense that a device might have a proper function even though it fails to perform it. Here the possibility of misrepresentation might be made clear.

What allows for the continuance of a proper function throughout generations? For a device/organism to have a proper function, it must share this function in common with its ancestors. You and I both have hearts that pump blood, because we share a common ancestor whose heart had the proper function of pumping blood. Proper functions are copied and reproduced through generations. However, we know (according to the best neo-Darwinian accounts) that no heart is directly copied from any other heart, rather, it is the genes which are directly copied and it is these genes that have the proper function of producing hearts. The normal explanation of the performance of a proper function makes reference to the normal conditions under which, historically, the proper function has been performed and selected for. This is illustrated by Millikan's bee dance example.

There are mechanisms in bees that have the proper function of producing a bee dance. There are also mechanisms in bees that have the proper function of consuming bee dances. The proper function of the bee dance producer (more strictly the relational proper function, because the function is related to something in the organism's environment) is to produce the consequence that consumer bees fly off in the direction of the orientation of the bee dance. The relational proper function of the bee dance producer is selected for iff in normal conditions it has, historically, led bees to find flowers, pollen, nectar, food – that which optimizes survival value. The consumer mechanism gets selected for iff, under normal conditions it has, historically, produced behavior leading to flowers, nectar, etc. on the basis of the consumed bee dances.

The producer mechanism has the function of producing intentional icons for consumer mechanisms and consumer mechanisms have the function of consuming the intentional icons produced by producer mechanisms for some further end. This requires that the producer and consumer mechanisms can only function properly if they are both present and coordinating; this is the normal condition for the mechanisms to function properly.

A normal condition of the environment, the location of nectar, has the effect of producing bee dances. These have the effect of sending consumer bees to the location of the nectar. This in turn produces two normal conditions in the environment, the nectar being located in the hive and flowers being pollinated at the first location of the nectar.

For the mechanisms to function properly the normal conditions must be in place. It is quite easy to see how contingent factors could interfere with the normal conditions of proper functioning, but this is why biological functions are normative; it is the proper function of mechanisms in normal conditions that is selected for. The next issue concerns why the bee dance is a representation.

Is this really representation? Firstly the distinction between proper functioning and malfunctioning looks secure. This could, in principle, underwrite the normative notion of content demanded by the possibility of misrepresentation. Secondly, the relationality of some proper functions gives them an intentional aspect, in Brentano's sense, they are directed at something beyond themselves. The intentional icons have three properties:

1. They are relationally adapted to some feature of the world.
2. The relation can be seen in terms of a 'mapping.'
3. The icons have the proper function of guiding a consumer mechanism in performing its proper function.

If the relational conditions for this bee dance (qua intentional icon) are normal, then it will successfully map the location of flowers, etc., call this indicative mapping. If this is successful, then the icon directs the consumer bee to the location of the nectar, call this imperative mapping. It is in the consumption of an icon that the representational function is established. The direct proper function of an icon is the effect it ought to produce (sending consumer bees in the direction of nectar), not what it statistically does produce.

Millikan provides an account of biological norms and how representation can arise according to these norms. This is quite different from Dretske's indicator function theory. The indicator relation is a causal relation that is recruited for some end. On Millikan's approach intentional icons represent because they are consumed by another mechanism and this relationship constitutes the proper function (hence the norm). Both these naturalistic accounts of representation (hence intentionality) are externalist in character. The content fixing relations and norms are based in the external environment of the organism, hence externalists seek to explain intentionality as a relation between representations and the world and not as a relation between representations and inner mental objects.

Unsurprisingly the naturalistic turn to analyzing intentionality in terms of representation has led to representationalist theories of consciousness. I turn now to considering the relationship between consciousness and intentionality.

Representational Theories of Consciousness

While Block's tripartite definition of consciousness provides a useful classification of different states of consciousness there are those who have denied the existence, or at least importance, of one of the types of conscious states. For example, Dennett has, famously, denied that there is a distinct sense of phenomenal consciousness at all. This is because phenomenal consciousness is really just a form of reflexive consciousness. Higher order thought (HOT) theories of consciousness are based on just such a thought, a state is conscious if we have a HOT about that state. Therefore, for a mental state M to be conscious there must be a HOT M*, which is about M. For a state to be conscious is for us to be aware of that state. To be consciously aware of pain is to have a thought that one is in pain. Thoughts are all potentially unconscious, but are made conscious by our becoming aware of them, by our having a further thought about them. This view of consciousness is not so very far away from the view presented by Descartes in the meditations (quoted in section titled 'Introduction: Mind, consciousness, and intentionality'):

> certainly it seems to me that I see, I hear, and I feel heat. That cannot be false; that is what in me is properly called sensation; and in this precise sense, sensation is nothing but thought.

If I am reflexively aware of a sensation of pain, if I have a thought about it, then it is a conscious thought. There is of course the question of whether animals and infants have the capability to form thoughts involving concepts of their sensory states – what they see, hear, smell, and feel. If they do not possess concepts of red, or pain, then they cannot form HOTs about those sensations and cannot, therefore, have conscious experiences.

Rosenthal counters this objection by claiming that the conditions for sensory concept possession are so minimal that most animals and infants are likely to possess them. Nevertheless some critics think that HOT theories involve an unnecessary additional layer of mental complexity to direct phenomenal experience, which is unmotivated. If I am in pain or seeing a red object why do I need to posit a further mental state to make the experience a conscious one?

A different approach is to accept that conscious experiences are representational, but not make the further move of modeling all of conscious experience on reflexive consciousness. Dretske disagrees with the HOT theorists on this matter, conscious mental states are not objects that we are conscious of having. Mental states make us conscious of something – my experience of a red balloon makes me conscious of the external object, the balloon. To be consciously aware of the balloon (to experience it), I do not have to be aware that I am conscious of the balloon. Dretske's approach is to apply his account of representation in terms of function of indication to sensory consciousness.

Dretske claims that all mental facts are representational facts and that experiences do not have phenomenal features of their own. The subjective quality of an experience is just the way the experience represents things to be. Dretske's representational theory of consciousness is compatible with Brentano's thesis that intentionality is the mark of the mental, but is incompatible with Brentano's thesis that the mental and the physical are separate categories, because nothing physical is intentional. However, Dretske's view is in tension with those views that hold that phenomenology is more fundamental to consciousness than representation.

Is Phenomenal Consciousness Intentional?

Phenomenal consciousness is a matter of the way things are experienced by us. To understand this aspect of consciousness philosophers have long posited a category of raw inner feels, or the qualitative character of experiences, that they refer to as *qualia*. *Qualia* are supposed to be the properties that constitute the phenomenal character of conscious experience, and what it is like for us to experience them. When you see a red object, or feel a stinging pain in your finger, or taste the smokiness of a heavily peated whisky, these experiences all have a particular phenomenal character, or *qualia*. Two questions arise at this point: Are phenomenal experiences intentional? Does intentionality itself depend upon the phenomenal character of consciousness?

One way of answering both questions is Sellars's view that sensations and thoughts are distinct. The realm of raw sensory feels is nonintentional and not to be confused with the space of reasons where intentional mental states such as beliefs are governed by norms of rationality. Davidson and Rorty also endorse this separation of the phenomenal realm of sensations and the intentional realm of thought. This position separates the phenomenal from the intentional, the answer to both questions is a resounding no. However, as we have already seen with Dretske's representational theory of consciousness there is an alternative to the separation move, which is to make all phenomenal experience at root intentional, because phenomenal experiences just are the way that things are represented in consciousness. Horgan and Tienson have recently argued that phenomenal experience is thoroughly intentional. my experience of a red balloon is intentionally directed at the redness of the balloon. The unity of our phenomenal consciousness amounts for them to the what-it's-like of being in the world. However, they also argue for a thesis of phenomenal intentionality: there is a kind of intentionality that is determined solely by phenomenology alone and not by bodily and sensory contact with the environment.

If they are right then externalist theories of content, such as Dretske's and Millikan's will turn out to be mistaken for these cases of phenomenal intentionality. Therefore, it would show that the strategy of explaining consciousness by giving a naturalistic account of intentionality was doomed to failure, because some of our phenomenology could not be explained in terms of representational content.

One might accept that there is a category of phenomenal intentionality where phenomenal experiences are intentionally directed, while rejecting the standard view of phenomenal experiences as being internal 'raw feels' or that the phenomenal character of experiences is independent of the way the world is (see the section titled 'Embodied intentionality'). This would be to reject Horgan and Tienson's position that phenomenology, and hence phenomenal intentionality, are independent of the world (what exists beyond the skin of the individual). Horgan and Tienson's position depends upon thought experiments such as brains in vats or deception by an evil demon. Normally our phenomenal experience is dependent upon our sensory modalities and our bodily progress around an environment. However, Horgan and Tienson claim that if you were a brain in a vat, with no sensory or bodily contact with the world, you could have the same phenomenal experiences as if you were really embodied. To argue for the narrowness of phenomenal experience and intentionality on the grounds that such scenarios are imaginable is empirically weak. Our phenomenology is due to our bodily contact with the world and it is in these terms that we will construct empirical theories. We shall not construct such theories by considering imaginable scenarios such as brains in vats. However that may be, Horgan and Tienson have

drawn our attention to the close relationship between phenomenal experience and intentionality. These issues are taken up further in the final two sections entitled 'The phenomenological tradition' and 'Embodied intentionality.'

The Phenomenological Tradition

Husserl followed Brentano in attempting to analyze conscious thought in terms of intentionality. In his earlier writings (*The Logical Investigations*) Husserl rejected Brentano's notion of intentional (or mental) inexistence – the object of thought is always internal (immanent to) the mind. By contrast Husserl thought that the objects of intentional states such as beliefs and desires, often transcend the mind, in the sense that they exist externally to it.

However, Husserl also proposed a 'phenomenological reduction' of mental phenomena, a method for investigating consciousness that 'brackets off' judgments about the ontological existence or inexistence of the world. Phenomenological investigations were, thereby, restricted to understanding the structure of mental acts. The exact interpretation of the phenomenological reduction is a matter of dispute, and is contested by Heidegger who views intentionality as the ontological structure of 'being-in-the-world,' and by Merleau-Ponty, who emphasizes the embodied nature of intentionality. Merleau-Ponty rejects the idea that there is a purely mental experiencing subject that can be considered apart from the body and the rest of the world. Rather, a conscious subject already presupposes a bodily and temporal existence situated in an environment. The conscious subject can only be understood in so far as he is a bodily subject of experience situated in and related to a preexisting environment. Consequently, for Merleau-Ponty, there can be no 'bracketing off' of the world when performing phenomenological investigations of the intentional structure of consciousness. For Merleau-Ponty an embodied subject is action-oriented to the world and has an intentional relation to that world, even before it starts to reflect on it. We can, therefore, derive from Merleau-Ponty a kind of prereflective embodied intentionality, one that does not yet involve conceptual content.

Embodied Intentionality

The move away from Brentano's account of intentionality in terms of intentional or mental inexistence is completed in the emerging approach to the mind and cognition, variously labeled: the embodied mind; the extended mind; distributed cognition; and cognitive integration. We saw how an externalist account of intentionality rejected the idea that intentional relations were understandable purely in terms of a pure conscious/mental subject, intentionally related to mental/intentional objects wholly encapsulated by the mind. Furthermore, Merleau-Ponty gave us the notion of a prereflective conscious subject as essentially embodied and situated in an environment. From this starting point the conscious subject and intentionality look very different from the Cartesian/Brentanian starting point in which intentionality is an entirely internal affair.

Combining a Merleau-Pontian position on embodied intentionality and an externalist position on representation would give us a relationship between intentionality and consciousness that would deal with Brentano's thesis and relieve the tension between the phenomenological and intentional aspects of consciousness. We would begin by thinking of our primary conscious engagement with the world as an intentionally directed bodily engagement, which also has a phenomenal character in Merleau-Ponty's prereflective sense. We would then add a layer of cognitive sophistication involving contentful (hence intentionally directed) mental states such as beliefs and desires. This fusion of embodied phenomenology and externalist theories of content is a promising line of inquiry into the relationship between intentionality and consciousness.

See also: Brain Basis of Voluntary Control; Concepts and Definitions of Consciousness; Free Will; Language and Consciousness; Mental Representation and Consciousness; Phenomenology of Consciousness.

Suggested Readings

Block N, Flanagan O, and Güzeldere G (eds.) (1997) The Nature of Consciousness. Massachusetts: MIT Press.
Brentano F (1874) Psychology from an Empirical Standpoint, Rancurello AC, Terrell DB, and McAlister L (trans.).

London: Routledge, 1973 [2nd edn., intr. by Peter Simons, 1995].
Crane T (2001) Elements of Mind: An Introduction to the Philosophy of Mind. Oxford: Oxford University Press.
Dennett D (1990) Consciousness Explained. London: Penguin.
Dretske F (1988) Explaining Behavior: Reasons in a World of Causes. Massachusetts: MIT Press.
Dretske F (1997) Naturalizing the Mind. Massachusetts: MIT Press.
Gallagher S (2005) How the Body Shapes the Mind. Oxford: Oxford University Press.
Gallagher S and Zahavi D (2008) The Phenomenological Mind: An Introduction to the Philosophy of Mind and Cognitive Science. Oxford: Routledge.
Husserl E (1900/1) Logical Investigations, Findlay JN (trans.). London: Routledge.
Menary R (2007) Cognitive Integration: Mind and Cognition Unbounded. Basingstoke: Palgrave Macmillan.
Merleau-Ponty M (1945) Phenomenology of Perception, Smith C (trans.). New York: Humanities Press, 1962 and London: Routledge & Kegan Paul, 1962 [translation revised by Forrest Williams, 1981; reprinted, 2002].
Millikan R (1993) White Queen Psychology and Other Essays For Alice. Bradford Books/MIT Press.
Moore AW (1993) Meaning and Reference. Oxford: Oxford University Press. (Collects papers by Frege and Russell with later developments in the field).
Rosenthal D (2005) Consciousness and Mind. Oxford: Clarendon press.
Tye M (1997) Ten Problems of Consciousness. Massachusetts: MIT Press.

Biographical Sketch

Richard Menary read for a BA in philosophy at the University of Ulster, an MSc in cognitive science at the University of Birmingham, and then a PhD in philosophy at King's College London. He has been a senior lecturer at the University of Hertfordshire in the UK and is currently a senior lecturer at the University of Wollongong in Australia. Richard has written articles on cognition and consciousness for journals such as *Philosophical Psychology*, *the Journal of Consciousness Studies*, and *Language Sciences*. He has also edited two books *Radical Enactivism* and *The Extended Mind*, and is the author of *Cognitive Integration*. He is currently writing another book on the philosophy of cognition.

Intuition, Creativity, and Unconscious Aspects of Problem Solving

J I Fleck, The Richard Stockton College of New Jersey, Pomona, NJ, USA
J Kounios, Drexel University, Philadelphia, PA, USA

Glossary

Impasse – A mental sticking point, usually during problem solving, after which no further mental progress seems possible.
Incubation – A period of time away from an unsolvable problem when the individual makes progress on the problem, likely via unconscious processes, while consciously engaged in other thoughts.
Insight – A moment of sudden understanding that accompanies the solution of a problem. This understanding is usually preceded by misdirected thought.
Metacognition – The understanding or impression of one's own thought processes.
Restructuring – A change in one's representation or understanding of a problem. Most commonly associated with the insight experience, restructuring is thought to occur after initial misdirection.
Semantic priming – Responses to stimuli are facilitated when items similar in meaning have been previously presented.
Synchrony – Large clusters of neurons begin firing together in the completion of a particular task. This synchrony can appear as greater EEG power.

Introduction

Creativity, insight, and intuition are just some forms of thought believed to rely heavily on the unconscious, giving these areas special status in psychology. Growing interest in creativity and problem solving has led to an influx of novel approaches in the empirical exploration of topics previously believed to be beyond our reach. This article reviews existing research and theory concerning the cognitive, behavioral, and neural components of the unconscious in creativity and problem solving. Of importance is understanding the field's past and conceptualizing its future.

An Historical Account

When Archimedes first stepped in the public baths of Syracuse he had no conscious awareness of what was to come, he merely intended to take a bath. But there came the answer to a problem which to that point had baffled him. Archimedes, according to written accounts, was attempting to determine if the crown of Hiero was made of pure gold, or if the crown instead contained a cheaper silver alloy. Though the finished crown was identical in weight to the gold initially provided for the crown's construction, it had been suggested to Hiero that the crown's constructor had tricked him. Archimedes was puzzled as to how to determine if the suspicion was true.

We cannot be sure of the mechanisms that led to Archimedes' Eureka moment – an eureka moment is one in which a solution to a problem that previously eluded the individual suddenly pops into mind, often so suddenly that it surprises the person resulting in an emotional reaction. For Archimedes, water splashing over the edge as he stepped into the bath facilitated the sudden and unexpected solution to his problem – the crown and an equal weight of pure gold should exhibit the same volume and, therefore, should displace the same amount of liquid if submerged in water. If the crown contained some amount of silver, it would be larger in volume (because silver is less dense than gold) and would thus displace more water than an equal weight of pure gold. Archimedes is said to have found just that – the crown

displaced more water than the pure gold providing proof that Hiero had indeed been duped.

The experience portrayed above of Archimedes's Eureka moment is similar to popular anecdotes of spontaneous discovery relayed in the creativity literature that detail the innovations of Newton, Poincaré, Kekulé, and others. Though what occurs in the minds of persons who produce such discoveries is often unclear, the phenomenological experiences are more widely recognized. Many researchers and laypersons alike have held the belief that creativity and related processes (e.g., intuition and insight) must differ in some fundamental way from other forms of thought (e.g., memory and logic) partly because they are associated with different experiences. For the individual, creative insights seem to come when least expected – after an unsuccessful and at times lengthy period of directed cognition, the insight or answer to one's problem appears in what seems a spontaneous manner when the person is otherwise engaged. Experimental definitions of insight are diverse; some of the most common include the moment of understanding after a period of mental confusion, generating a solution outside awareness, and generating a solution by overcoming a mental sticking point after changing one's thinking about the problem at hand. Historically, the mystery surrounding creative insights was so great that creativity was believed to be a gift from the gods. How else could such unprompted events be explained?

Graham Wallas first drew our attention to the complex interplay between conscious and unconscious forms of thought as it pertains to creativity. Wallas was an English psychologist and political scientist whose book *The Art of Thought*, published in 1926, significantly impacted how researchers would study creative thought and its unconscious components. Wallas suggested that creativity was composed of four stages: preparation, a period of directed knowledge acquisition; incubation, a period of unconscious task-related activity occurring while the conscious mind is otherwise engaged; illumination, the moment of insight when the answer emerges into conscious awareness; and finally, verification, when the now-conscious insight is subjected to a period of directed fact checking. Of these stages, incubation and subsequent illumination have held the most potential and interest for researchers attempting to reveal the thoughts responsible for creative insights. Almost a century of research and theory in this area has provided merely a glimpse into the cognitive and neural components – conscious and unconscious – preceding that final Eureka or Aha! moment.

In this article, we address several themes surrounding the study of the unconscious in creativity, problem solving, and related processes. First, we present existing data on the role of the unconscious in problem solving, specifically problem solving by insight, believed by many to elicit unique and unconscious components of thought. Second, we explore the process of incubation, the stage of creative thought proposed by Wallas with the most significant tie to the unconscious. Third, we consider contributions to our understanding of the unconscious in problem solving and creativity that have emerged from advances in neuroscience. Finally, we present a glimpse of the up-and-coming topics sure to influence our future conceptualization of the unconscious in problem solving, creativity, and related processes.

The Study of Creative Insights

A Search for Definitions

There are few fields in psychology that have been more stymied by the discrepancy over a definition than the field of creativity. A general definition, though unsatisfying to many, is that creativity is a process that results in the generation of something that is both novel and useful (the terms novel and useful have also been much debated). A series of factors have been proposed as part of a complex state preparing the individual for the experience of creative insight; personality characteristics, environment, knowledge acquisition, intelligence, and motivation are just some of these factors. Even mental illness (e.g., bipolar disorder and schizophrenia) has been associated with enhanced creative potential. Added to the mix is the potential relevance of the magnitude of the creative product as well as the eminence of its creator. Should our focus be the creative

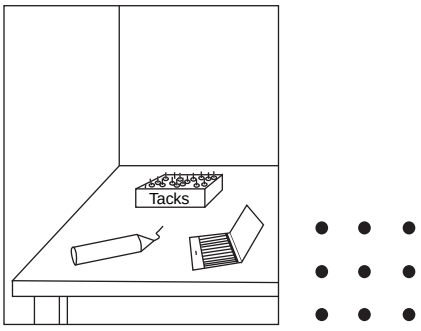
process in eminent individuals, in normal adults, or both; and, how do the creative processes in these two groups relate to each other? Significant in all fields but perhaps more relevant to the study of creativity, is how researchers are able to manufacture experimental situations in a controlled environment that are valid reflections of the natural creative process. Finally, and of greatest relevance here, are the mechanisms that lead to creative thought the result of conscious or unconscious processes?

Studying the Unconscious through Insight

Insights, most often explored in conjunction with problem solving, are analogous to the illumination stage of Wallas' creativity stages. Insight problem solving has been used by researchers to explore creative thought in a controlled laboratory environment; however, little research exists that directly explores the relation between insight and creativity. As with illumination, there is strong debate about the components (conscious or unconscious) that precede the insight experience. Insight has been described as the end product of a process in which the individual is initially misdirected by the problem components, with this misdirection leading the individual to a mental sticking point or impasse. Further progress is only possible when the individual experiences a sudden restructuring or change in mental representation of the problem – in fact, many argue that restructuring leads the individual directly to the problem's solution. The final stage of the process, after the solution enters awareness, is often met by an Aha! experience that occurs when the verification process reveals that the solution in hindsight could have been easily and directly obtained had the solver correctly applied the problem information. Metacognitively, insights have been described as the proverbial lightbulb suddenly going on in one's head.

The Gestalt psychologists initiated the study of insight as an alternative to the trial-and-error approach to learning proposed by Edward Thorndike. While Thorndike described learning as the gradual acquisition or building of knowledge, premier psychologists of the time including Karl Duncker and Wolfgang Köhler suggested that more intelligent forms of thought such as insight were possible. The school of Gestalt psychology originating in early nineteenth century Germany is best known for its theories of perceptual organization, most popular the tendency of humans to organize the visual field as a whole rather than by its constituent parts. For the Gestalt psychologists, the change that led to the insight experience was sudden in nature and was preceded by the misapplication of prior knowledge in situations where the individual's existing knowledge base was not relevant – in other words, the conscious application of information obtained via long-term memory search impeded progress. Examples of this misapplication or fixation are apparent in two classic insight problems from the time: the candle problem and the nine-dot problem (see Figure 1). Classic insight problems are those originally used in the time-period of the Gestalt psychologists but have remained key problem stimuli in more recent studies of creativity and insight. In the candle problem, it is common for problem solvers to become fixated on the tack box as serving the function of holding the tacks (i.e., box as container) and for individuals not to extend the box to other roles such as a shelf or ledge to support the candle. In the nine-dot problem, perceptual fixation stemming from the inferred perimeter formed by the square configuration of the nine dots reduces the likelihood that participants will extend lines beyond the square when attempting the problem's solution.

Preliminary research into the mechanisms of insight problem solving originally focused on if, and in what manner, solutions to problems solved via insight were somehow the result of processes that were comparable to the processes supporting other forms of problem solving in which strategies were methodically and consciously applied by problem solvers (e.g., analysis-based problem solving). Though many researchers argued for unique cognitive processes in insight, others supported the idea that little more than speed of processing differentiated insight from analysis-based problem solving. Research conducted in the late 1980s and throughout the 1990s revealed several possible differences between the two problem types, especially with regard to the role of conscious processing.



(a) (b)

(c)

Figure 1 Classic insight problems used in the study of insight and creativity. (a) The candle problem. Using the items provided, attach the candle to a door so it can burn properly. Reprinted from Cognitive Psychology, Vol 4, Robert W. Weisberg and Jerry M. Suls, An information-processing model of Duncker's candle problem, pp. 255–276, Copyright (1973), with permission from Elsevier. (b) The nine-dot problem. Without lifting your pencil, connect the nine dots using four straight lines. (c) The two-string problem. Two strings are hanging from the ceiling. The distance between the strings is such that you cannot reach one of the strings while holding on to the other. Using the items provided, tie the two strings together. Reprinted from Cognitive Psychology and Its Implication, J. R. Anderson, p. 247, Copyright (1980), with permission from W. H. Freeman. Problem solutions are available in Figure 4.

First, it appeared to be the case that the metacognitive awareness that existed during analysis-based problem solving was lacking in insight. In an interesting series of studies, Janet Metcalfe and colleagues explored participants' awareness of their problem-solving potential and their online perceptions of their solving progress for a set of insight problems and a set of noninsight problems. After participants' brief viewing of the to-be-solved problems, they were unsuccessful in predicting their abilities to generate solutions to insight problems, although they were quite successful in anticipating their likelihood of generating the correct responses to a series of trivia questions. Differences also emerged during the solving process itself. During analysis-based problem solving, participants correctly perceived and reported that they were inching ever closer to the problems' solutions. Not so with insightful problem solving; solvers instead were unclear as to their progress until the interval immediately preceding the solution. Comparable findings have also been observed in physiological recordings of heart rate during problem solving, with a gradual increase in heart rate accompanying the incremental solution of analysis-based problems and a sudden increase in heart rate immediately prior to the solution of insight problems.

Also of interest, it did not appear that forcing thoughts linked to problem solving into consciousness was advantageous when attempting insight problems. Instructing participants to generate verbal protocols (i.e., to verbalize their thoughts during problem solving by thinking out loud) as a window into their thought processes negatively influenced participants' solving performance for insight problems but had little effect, or in some cases facilitating effects, for participants providing verbal protocols while attempting to solve analysis-based problems. It should be noted that the interfering effects of concurrent verbalization on the solution of insight problems has been viewed as tenuous with subsequent studies having difficulty replicating these effects.

Researchers have extensively examined the restructuring component of insight, thought to result in the sudden awareness or illumination of the problem's solution. Explorations of the nature of restructuring have been limited primarily to behavioral approaches and have focused on the potential occurrence and characteristics of restructuring itself rather than the factors that facilitate the restructuring. Several classification systems have been proposed detailing the possible

mechanisms through which restructuring is possible, the most notable of which was proposed by Stellan Ohlsson. According to Ohlsson, forms of restructuring include, but are not limited to, a change in one's representation of the problem state (the initial problem configuration) or the goal state (the desired ending effect), a change in the operators or problem components available for application, the decomposition of problem components, and perhaps receiving most attention, the relaxation of previously imposed problem constraints (the problem solver's beliefs about how different components relate to each other or the mechanisms acceptable in the problem's solution). A change in any one of these things could lead to a problem's solution, though it has also been suggested that multiple restructurings may be needed to solve a problem depending on problem difficulty.

Behavioral studies have been divided on the role of the unconscious in restructuring, with researchers suggesting that both conscious and unconscious components can elicit restructuring and subsequently insight. Proponents of the conscious view suggest that long-term memory search and retrieval initiated by bottom-up (external) or top-down (internal) sources can lead to restructuring experiences. During the problem-solving experience when impasse is reached, the solver consciously begins to search memory for relevant information. This conscious application of new information is only possible, however, if the individual possesses in memory relevant knowledge for the problem's solution. Some support for this possibility has been found in research employing classic insight problems such as the nine-dot problem. Researchers observed limited restructuring in participants attempting to solve the nine-dot problem even when solution hints, such as the instruction to extend solution lines outside the dots' configuration, were provided. Solution rates improved, however, when participants were provided with training in activities such as line extension prior to attempting the nine-dot problem. The beneficial effects of prior training suggest that relevant knowledge must be in place prior to problem solving and recent enough in memory to be retrievable if it is to be applied during problem solving.

In contrast, others have proposed that restructuring occurs via the unconscious spreading of activation stemming from problem components. Spreading of activation occurs when a node or representation is activated in the brain, resulting in the spread of activation to related nodes or concepts. Support for this theory appeared first in research reporting the detrimental effects on insightful problem solving when individuals' thought processes are forced into consciousness. As noted above, some researchers have observed that verbalization resulted in a significant decrease in solution rates for insight problems when compared to participants attempting the same problems but during silent conditions. These results led the researchers to suggest that certain components needed for insight could not be verbalized.

Some of the most recent research in the debate on the conscious versus unconscious nature of restructuring has explored the role of working memory in the solution of insight problems. Working memory is most often defined as the online storage and processing, including directed attention, of information during conscious thought. In order to specifically isolate the restructuring component of insight, researchers manipulated the amount of processing required in the initial phases of problem solving. With many problems, solvers must explore a number of possible solution approaches before reaching the impasse state thought to precede restructuring. When participants attempted to solve problems that isolated the restructuring component of insight (i.e., the initial solving space was small so that participants quickly exhausted their initial ideas) working memory was not a significant predictor of success. In contrast, when participants were presented with problems that included a much larger initial solving space (i.e., multiple solution paths requiring consideration prior to restructuring), working memory was a significant predictor of solving success. Similar research exploring the role of working memory in the solution of insight versus analysis-based problems revealed that working memory was a significant predictor of success for analysis-based problems, whereas short-term memory capacity (storage only) was a significant predictor of success on insight problems. Together, this line of research suggests that the directed

attention and processing present in working memory may only serve a role in the initial stages of insight problem solving when the individual is attempting to rule out possible paths to the solution; of course, the solver must have the impression that all reasonable possibilities have been attempted in order to achieve impasse. After this mental sticking point is achieved, working memory is not a factor that influences whether or not the individual will experience a restructuring of the problem's components – that instead may be directed by processes outside awareness.

Incubation – The Unconscious at Work?

Though the unconscious has been the cornerstone of many theories of creativity and insight problem solving, evidence of its influence has been difficult to obtain. One of the most fundamental questions that researchers faced was whether incubation, the unconscious period in Wallas' stages of creativity, was passive, advancing thought as a mere byproduct of the passage of time or was an active process requiring additional mental processing such as the search of long-term memory or the integration of new information from the environment. Passive theories suggest that time away from the primary task allows original memory traces to weaken in favor of more fruitful ideas. Empirical support has been found for this theory through the early work of Steven Smith and Steven Blankenship who demonstrated that time away from a problem-solving task resulted in higher solution rates when participants later returned to the task of interest. Solving rates during this second solving window after a break were greater than those achieved when the second problem-solving window followed immediately after the initial solving window. This effect was later extended in research demonstrating that both short and long breaks were equally effective in increasing problem-solving success.

A similar theory focusing instead on cortical arousal suggests that a decrease in cortical arousal during incubation could effectively alleviate mental sticking points by generating an environment where more diffuse attention is possible. Focused attention could cause problem solvers to become fixated on their initial solution ideas, therefore resulting in impasse. Undirected cognition or cognition directed toward the completion of another task would cause a decrease in arousal, thereby reducing the focus on one's conscious solution ideas allowing novel solution possibilities, perhaps outside awareness, to be considered.

Alternative theories of incubation suggest more active forms of processing. Most active theories to date imply that the application of external hints could be sufficient to facilitate novel solution ideas. One such idea is the opportunistic assimilation hypothesis proposed by Colleen Seifert, David E. Meyer, and colleagues. Opportunistic assimilation proposes that when an impasse or sticking point has been reached, a mental flag or failure index is linked in the brain to existing task-related information. As the individual then moves through his or her daily events, external stimuli encountered initiate new activation that spreads to associated areas in the brain. New solution ideas may enter into consciousness if this novel activation reaches the stored failure indices, indicating the relevance of the external stimuli in the problem's solution. Research evaluating the role of external stimuli in incubation has generated mixed results.

Some of the first research exploring the value of external hints in the unconscious facilitation of problem solving was conducted by Norman Maier, using the two-string problem (see Figure 1). During the solving period, a confederate entered the room where the participant was working and, within the participant's view and in an apparently unintentional manner, brushed one of the strings causing it to swing from side to side. The participant group exposed to the confederate exhibited higher solution rates than those attempting to solve the problem without the confederate's presence, though most participants did not report the actions of the confederate as influencing their solution ideas. In this instance, external cues influenced unconscious but not conscious thought. Although participants in Maier's research were able to apply the external hints in generating the solution, other researchers have had less success in replicating this effect. For example, participants were not successful in applying solution hints to problems from Sarnoff Mednick's remote associates task (RAT), a popular test of creativity

developed in the 1960s. In the RAT, participants are presented with three words and must identify a fourth word that can link all three (e.g., problem words: letter, puppy, true – solution: love). During the incubation phase, participants were asked to perform a word generation task in which they needed to recombine the letters of a stimulus word to form new words; the stimuli for the task contained the solution words for the RAT problems. Participants were not able to apply these hints to the RAT problems during the second solving period unless they were explicitly told of the hints' value in solving the problems. They instead seemed to compartmentalize the initial problem-solving phase and the word-generation task or incubation phase that followed.

Recent research has suggested that a key factor in the effectiveness of the incubation period is the solver's expertise in the topic area of the problem. For example, experts were less likely to solve misleading problems during the initial solving period than novices presumably due to fixation resulting from their application of irrelevant prior knowledge; however, after an incubation period, they exhibited a reduction in this fixation and improved solution success. Interestingly, an incubation period did not facilitate the solution of problems for which their expertise was relevant or when they were asked to solve neutral problems outside their areas of expertise. Further, no incubation effects were present in novices who were not experts in fields related to the problems and, therefore, were not fixated during the initial problem presentation. These results suggest that mental fixation by the solver during the initial solving phase may be a necessary prerequisite for the occurrence of the incubation effect.

Evidence for an Active Unconscious

Some of the first behavioral evidence for an active unconscious came from research conducted by Kenneth Bowers and colleagues. In an interesting series of studies, researchers were able to demonstrate that lack of awareness of the solution on the solver's part did not coincide with lack of processing. Bowers presented participants with two three-word sets in a task he called The Dyads of Triads Task. In one of the sets, all three words were associates of a fourth (not presented) target word.

A	B	
Playing	Still	A
Credit	Pages	Card
Report	Music	

Figure 2 Sample dyads of triads task. Only triad A is coherent. Each word in A can be combined with the target word card. Reprinted from Cognitive Psychology, Vol 22, Kenneth S. Bowers, Glenn Regehr, Claude Balthazard, and Kevin Parker, Intuition in the context of discovery, pp. 72–110, Copyright (1990), with permission from Elsevier.

In the other set, only two of the words were associates of the same word; the third was not (see Figure 2). When possible, participants were asked to provide the target word for the coherent triad. If a solution was not possible, they were then asked to choose which of the two sets was the coherent triad and to rate their confidence in this decision. For unsolved items participants were able to choose the coherent dyad at rates above chance. Even more interesting was the observation that participants demonstrated increased accuracy during the forced-choice decision when the solution word for the coherent dyad performed a semantically similar role for all three words than in instances where the dyad words were all linked to the same solution word, but that word carried different meanings as an associate.

These and other results led Bowers and associates to suggest that the summation of activation to the target word on the unconscious level led to the correct choice for solvers during the forced-choice decision even though this information was not yet available in consciousness.

Neural Underpinnings of Creative Thought

The Creative Right Hemisphere

Prior to the use of modern imaging techniques such as functional magnetic resonance imaging (fMRI) and high-density electroencephalography (EEG) which will be discussed below, much was learned about the brain's involvement in creative thought through research using priming paradigms previously developed for language research. In priming, judgments regarding a target word are faster if that target word is preceded by a

semantically related source word (i.e., the prime) than when judgments are made about the same target word without the preceding prime (e.g., people are faster in making judgments about the target word dog when it is preceded by the prime word cat). In semantic priming tasks, primes presented to the right visual field (i.e., going to the left hemisphere) have been associated with priming effects for primary word meanings and direct semantic associates. Primes presented in the left visual field (going to the right hemisphere) are instead associated with priming effects for more diffuse semantic connections, including secondary word meanings and remotes associates. It has been further suggested that the structure of the neural networks in the hemispheres may be uniquely constructed to support these two forms of priming.

Researchers utilizing priming paradigms have observed that priming effects in the right hemisphere are linked to the creative solution of compound remote associate (CRA) problems. CRAs, similar in nature to the problems from Mednick's RAT, are three-word problems in which a solution word can be combined with each of the problem words to generate a compound word form (e.g., problem: pine, crab, sauce – solution: apple – pineapple, crabapple, and applesauce). Participants were given 15 s to generate the solution for the three problem words presented in the center of the computer screen. If participants were unsuccessful in generating the problem's solution, then a target word was briefly presented (180 ms) in either the right visual field or the left visual field; this target word was either the problem's solution or an unrelated word. Participants were required to determine as quickly as possible if the target word was the correct or incorrect solution for the problem. After making their solution/nonsolution judgment, participants were asked to report the nature of the thought processes surrounding the judgment; was their choice the result of a deliberate, conscious strategy (low-insight/low-creativity) or instead made without conscious awareness of the processes behind their decision (high-insight/high-creativity). A right-hemisphere priming effect (i.e., faster judgments) occurred in the solution/nonsolution identification task for decisions that were reported by participants as high insight. Similar effects have been observed in research using classic insight problems. When participants were initially unsuccessful in generating the solutions to classic insight problems, solution hints presented in the left visual field facilitated the creative solution of unsolved problems. Together these results suggest that the activation of right-hemisphere semantic networks may be more likely to produce an environment where the sudden solution of verbal problems outside awareness is possible.

While the above priming research began to suggest hemispheric differences between creative and noncreative thought, most behavioral research was limited in its ability to definitively demonstrate that creative thought was reliant upon unique, perhaps unconscious, mental processes. At the rise of the cognitive neuroscience revolution in the early 1990s, many researchers held the theory that creativity stemmed from unique processes but had no way to substantiate that claim. Supporters of the business-as-usual approach to insight and creativity were also restricted in their ability to demonstrate that the processes in creative and noncreative thought did not differ.

Creativity and the Unconscious – The Right Hemisphere and Beyond

Since expanding to a more diverse set of tasks and techniques, researchers have suggested that a variety of different regions of the brain support creativity and related unconscious processes; the frontal lobes, the anterior cingulate, posterior regions of the cortex, and the cerebellum are just some of these regions. A number of researchers have used EEG to examine neural differences between convergent and divergent thought. With EEG, electrodes placed on the scalp record changes in electrical brain activity for significant numbers of neurons. These changes stem from large numbers of neurons firing in synchrony reflecting mental processing in the brain. Divergent thinking has traditionally been viewed as a form of creative thought. Typical divergent thinking tasks are similar in nature to the tasks originally proposed by Joy Paul Guilford. Guilford's assessments of divergent thinking required participants to generate as many unusual uses for

common objects (e.g., shoe, matchstick) as possible. Responses were scored based on the number of responses generated (fluency), and their originality from other responses from the same individual (flexibility), as well as from the responses of other participants in the study (originality). In contrast, convergent thinking tasks often included activities such a mental arithmetic, the completion of verbal math problems, and sequence completion tasks, all tasks requiring a specific response.

EEGs recorded from participants completing divergent and convergent tasks revealed greater EEG complexity over frontal, central, and parietal regions when participants were engaged in divergent thought. Greater EEG complexity often suggests that there is less synchrony in activity between different regions of the cortex or that multiple neural assemblies (large groups of neurons) are necessary to support the task of interest. Interestingly, when participants were subdivided into high and low performance groups based upon divergent thinking ability, high performers exhibited less complexity over frontal/central regions than did low performers. A similar effect was observed when participants were provided with training in divergent thinking (nine 30 min sessions). In fact, following training, participants exhibited greater low-alpha power (an EEG frequency indicating cortical inactivity) over frontal regions when compared to the control group. Together these results suggest that as individuals are more skilled in divergent thinking, fewer mental resources, especially in frontal regions of the brain, are needed to search for relevant strategies to apply during divergent-thinking tasks.

Researchers exploring alpha power throughout the cortex during divergent thinking have found general support for theories proposing that reduced cortical arousal facilitates creative thought. In general, greater alpha power has been observed over central and parietal regions during divergent thinking; this increase was more enhanced during segments of the EEG associated with ideas rated by participants and external raters as highly original. Patterns of alpha activity observed during creative thought generally support theories highlighting the critical role of decreased cortical arousal in the generation of novel ideas. Therefore, it would appear that high arousal strengthens prepotent (initial) response ideas in consciousness, whereas lower arousal adjusts the playing field of ideas allowing ideas in the unconscious to enter consciousness. A common component of several theories of creative thought is that the individual must be able to generate a substantial number of solution ideas linked to the task at hand – some have proposed that posterior regions of the cortex are responsible for this prolific generation. Enhanced alpha power over posterior regions of the cortex may reflect a decrease in conscious processing in the region, allowing the combination of problem components to occur on the unconscious level. After the unconscious combination of problem components a solution or idea could then be selected from among this collection and subsequently implemented.

New Tasks, New Methods: Exploring the Moment of Insight

One of the most critical facets in understanding the unconscious and its role in creativity is understanding the moment of insight – when the work of the unconscious enters awareness. Before considering research in this area, it is important to note that by no means is there universal agreement regarding the neural substrates of unconscious processing supporting creativity and problem solving. For example, some researchers have proposed that the basic process of plasticity (neural changes occurring in response to learning or experience) may be enough to produce the unconscious incubation effects that lead to the moment of insight. In these theories, the strong activation of problem components facilitates additional spreading of activation and modification of connections on the unconscious level while the problem solver is otherwise engaged. Then, when returning to the original problem, the new neural networks constructed via the brain's own plasticity may be more effective than during the initial attempt.

Studying the exact moment of insight, when unconscious thought becomes conscious, has proved difficult. One key problem noted above is that unconscious processes are not directly reportable by participants reducing the utility of verbal and behavioral measures of thought. Further, the classic problems that shaped the initial study of creativity,

chiefly the comparison of solutions in insight versus analysis, posed several problems. First, there was concern that problems labeled as insight-based and those as analysis-based may differ in ways other than their mechanisms of solution: concerns over differences in task difficulty, the ambiguity of the goal state in insight versus analytic problems, and the general lack of control that stems from using different problems are just some of the criticisms lodged. An even greater concern, however, has been the reliance by researchers on the assumption that insight problems are indeed solved via insight and that analytic problems are solved independent of insight. In response to these criticisms new stimuli have been introduced in the study of creativity and unconscious problem solving.

A change in task stimuli was also a requirement for neuroscience applications, where a substantial number of standardized trials are necessary in order to identify activation patterns or components unique to creativity. We will explore just some of the tasks that have gained popularity. In recent years, Jing Juo, Xiao-Qin Mai, and colleagues have explored neural correlates associated with the solution of riddles and brain teasers using fMRI and EEG techniques (e.g., ''The thing that can move heavy logs, but cannot move a small nail'' Luo and Knoblich, 2007, p. 79 – solution – ''a river''). The use of these techniques together can provide the researcher with a more complete picture of the neural underpinnings of creativity and related processes; fMRI provides spatial data indicating where in the brain differences are present, whereas EEG provides temporal data signifying when differences are occurring. To compare brain mechanisms associated with creative versus noncreative processing, Luo manipulated the type of hint presented to participants for unsolved riddles; hints were either those intended to elicit restructuring or those that added more detail to the problem without changing solvers' mental representations. Unique activation following restructuring cues was observed in the anterior cingulate, a region previously linked to processing cognitive and emotional conflicts.

Though noteworthy, the above research could be criticized for its reliance on the assumption that insight processes (e.g., restructurings) occur following the presentation of restructuring hints and not following the other hint types. To address this concern, Mai, Luo, and colleagues presented participants with verbal riddles, half of which were difficult to solve increasing the likelihood of impasse and subsequent Aha experiences, and half which were easy by comparison and not likely to elicit Aha experiences. It should be noted that riddles of both difficulty types could be solved with or without insight. During the 8 s solving interval, participants were to generate a solution if possible. Two seconds after the solving period elapsed, the solution to the riddle was presented. Participants then self-reported whether the solution presented matched the one they had generated during the solving interval (no-Aha trials) or if the solution, while clearly correct, was different from the solution they had generated or made sense even though they were unable to generate a solution (Aha trials). It should be noted, however, that many researchers question the classification methods used in the Aha/No-Aha categorizations noted above. For example, it seems possible that a participant's idea that matches the presented solution could have been generated via insight; it also seems possible that a nonmatching idea generated during the solving window could have been the result of an analysis-based strategy.

The researchers' analysis of segments of interest recorded using high-density EEGs during the problem-solving task revealed early visual event-related potential (ERP) components for both Aha and no-Aha trials. However, a later negative component unique to Aha trials was observed over midline electrodes. Subsequent analysis localized the source of this component to regions in and around the anterior cingulate. These data corroborate the results of their prior fMRI study in identifying the anterior cingulate as a structure linked to the restructuring component of insight. If we can accept the assumption that restructuring hints facilitated insight/creative solutions, the question becomes how activation in the anterior cingulate is related to processing during these experiences. Because the anterior cingulate is involved in conflict processing, it may be that activation following restructuring hints stems from the individual considering multiple solution ideas – the initial prepotent response, as well as other ideas that are being considered but are still at the unconscious level.

A second set of experiments conducted by Mark Jung-Beeman, John Kounios, and colleagues has been instrumental in expanding our understanding of differences in the brain during creativity and related processes. These researchers have used fMRI and high-density EEG to isolate neural differences when participants solve CRA problems (described above) using different strategies. Participants were asked to attempt 180 CRA problems, and were given up to 30 s to generate each problem's solution. If successful, participants were asked to report the type of strategy they used in determining the problem's solution, either insight or noninsight. A noninsight solution was one in which a participant could identify the strategy used in generating the problem's solution – a conscious application. In contrast, an insight solution was one in which the solution suddenly popped into the participant's awareness without conscious access to the strategy yielding the solution.

Both spatial and temporal information distinguishing the two solution types was obtained. Results of the analysis of fMRI data for the brief time interval immediately prior to a solution response revealed a significant difference in right-hemisphere brain activity for insight and noninsight problem types – greater activity was present in the anterior superior temporal gyrus (aSTG) for CRA problems solved via insight than for those solved via noninsight strategies. EEG data, more sensitive to temporal information, were highly valuable in this paradigm due to the sudden nature of insight solutions. Isolation of the window immediately preceding the solution response corroborated and extended the fMRI results. A surge of high-frequency (gamma-band) activity was observed over a region of the STG in the right hemisphere for insight solutions about 0.3 s prior to the solution response (see Figure 3); the gamma frequency band has been associated in prior research with perceptual and semantic integration. Activity in earlier time intervals preceding the solution provides a clearer picture of the insight/noninsight distinction. When comparing the two problem types for the second preceding the insight burst (i.e., 1.5 to 0.3 s preceding the solution response), gamma activity did not significantly differ between insight and noninsight solutions.

Finally, the analysis of EEG data in the above research includes a finding in the alpha frequency band that further supports Martindale's theory of hypothesis selection as a critical component to the creative process. Greater alpha activity was present prior to solution over posterior regions of the cortex for solutions reported to result from insight solution strategies; this activity occurred immediately prior to the right-hemisphere gamma burst. Posterior alpha is linked to sensory gating (the reduction of visual inputs) and could reflect the inhibition of inputs from the environment that might otherwise interfere with the successful retrieval of a weakly activated (unconscious) solution idea.

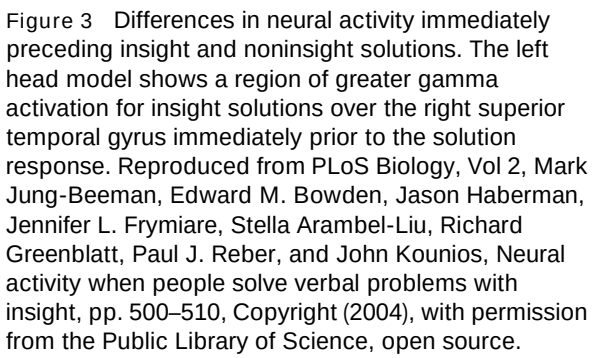


Figure 3 Differences in neural activity immediately preceding insight and noninsight solutions. The left head model shows a region of greater gamma activation for insight solutions over the right superior temporal gyrus immediately prior to the solution response. Reproduced from PLoS Biology, Vol 2, Mark Jung-Beeman, Edward M. Bowden, Jason Haberman, Jennifer L. Frymiare, Stella Arambel-Liu, Richard Greenblatt, Paul J. Reber, and John Kounios, Neural activity when people solve verbal problems with insight, pp. 500–510, Copyright (2004), with permission from the Public Library of Science, open source.

Special Topics in the Study of Intuition, Creativity, and Insight

Is the Unconscious Better Suited for Creativity?

Recent publications in the field of creativity suggest that creativity can result from either conscious or unconscious processes. Beginning with Mednick, the idea that creative thoughts are more likely to be achieved via flatter, diffuse associations throughout the brain has been a popular theory. Central to this theory is the idea that each concept is associated to a large number of

other concepts with relatively equal strength. This is in contrast to a steep gradient of association in which each concept is strongly associated to only a small number of other concepts. The need for flatter associations makes sense if we consider that unique combinations of problem subcomponents, whether verbal or structural in nature, often spring from distantly rather than directly related items. For example, the solution to tasks such as remote associate problems (e.g., CRAs) often relies on the solver to generate unique and relatively distant associates of target words.

Another popular idea in creativity theories is the importance of the unconscious in generating an extensive list of novel ideas for consideration. For example, Dean Keith Simonton suggested that creative products may be the end result of the recombining of problem components on the unconscious level. According to this theory, the prolific unconscious generates combinations of problem components in search of a solution. Because the unconscious is not directed in the combinations composed, chance plays a considerable role in the ideas produced. Cognitive theories addressing the general mechanisms of unconscious thought suggest that the unconscious may be uniquely suited for this role. The unconscious is believed to have more available resources than the conscious which is limited to processing items within the scope of awareness. In addition, the unconscious is capable of processing information in parallel making sufficient resources available for idea generation, less likely to occur in the serial processing that limits conscious thought.

Further support for the role of the unconscious in creative thinking comes from recent studies using neuroimaging techniques which have suggested that gating of external stimuli may be necessary to facilitate creative thought. As noted above, the research of Jung-Beeman, Kounios, and colleagues demonstrated a period of increased alpha power over posterior regions of the cortex immediately preceding insights. This alpha power, linked in prior research to sensory gating, may reflect the need to devote conscious resources to the selection of the best alternative from among those formed by the unconscious. It should be noted, however, that in several theories of creative thought, such as opportunistic assimilation, broader perception (i.e., attention to a wider variety of external stimuli) can stimulate insight experiences.

Few studies have directly explored whether unconscious or conscious thinking is more likely to produce creative products. Preliminary research in the area suggests greater potential for creative output by unconscious processing. In a series of experiments conducted by Ap Dijksterhuis and Teun Meurs, participants were asked to perform generation tasks (e.g., name as many cities as possible that begin with the letter 'A') immediately after the instruction was given, after 3 min of directed thinking about the generation task (conscious condition), or after 3 min of performing a distractor task (unconscious condition). Participants in the unconscious condition listed more atypical cities than did participants completing the task under the other two conditions. A similar pattern of results was observed when the alternate uses task (e.g., generate as many uses as possible for a brick) was performed under the same three conditions. Though these results seem promising, the observed effect in the unconscious condition could simply be the result of the passage of time allowing prepotent thoughts to weaken in favor of more remote alternatives. If so, a passive period of incubation may be sufficient to produce these effects rather than the generation of solution ideas by an active unconscious.

When we consider whether the unconscious is sufficient for creativity or insight, the answer is no. No matter how prolific the unconscious, creative thought demands that the individual select an idea for implementation. In insight, conscious thought reduces the initial problem space such that the individual is able to reach impasse, setting the stage for restructuring. In creativity, an extensive period of knowledge acquisition, often in excess of 10 years for significant creations, is needed as preparation for creativity. Finally, it is only through illumination that ideas formed by the unconscious can be evaluated as to their merit in Wallas' final stage of creativity, verification.

Equal Preparation for Equal Results? Predictors of Creativity

There is a long history in cognitive psychology suggesting that individual differences in attention, working memory, fluid intelligence, and other

processes translate into differences in task performance. This is no different in creative processing, where factors such as diffuse attention and diffuse semantic networks have been highlighted as key areas contributing to enhanced creative potential. For example, normal adults scoring high on psychometric measures of creativity exhibited weaker performance on measures of selective attention than their less creative counterparts. Also of interest is the finding that normal adults who exhibited strong performance on the RAT were more likely to apply hints presented outside the scope of focused attention when solving subsequently presented creative problems. Thus, possessing reduced selective attention (or increased diffuse attention) may ultimately facilitate the occurrence of creative thought by allowing the individual to consider ideas outside the scope of focused attention, even those ideas generated at the level of the unconscious.

Several affective and clinical factors have been examined as to their effects on selective attention. Research has demonstrated that positive affect results in the more diffuse allocation of attentional resources, facilitating creative problem solving, as well as problem solving via insight. A predisposition toward psychopathologies, such as schizophrenia, has been found to coexist with altered cognitive processing in both attention and language and may facilitate insight and creativity. For example, normal adults who exhibit high levels of traits often present in schizophrenia such as sensitivity to light and belief in the paranormal exhibit more diffuse allocation of attention resources. These individuals also present more diffusely organized language networks than do normal adults with low levels of these traits. Finally, a number of correlational studies report that higher levels of schizophrenic-like traits are associated with improved performance on psychometric measures of creativity.

A novel area of research has explored brain states that may be linked to a predisposition for creative forms of thought. Researchers have used EEG to examine the link between baseline (resting) activity in the brain and the strategies utilized most often by participants during the solution of anagrams. Anagrams are brief puzzles in which participants must unscramble the letters provided to form a word (e.g., problem: vleo – solution: love). Participants were asked to solve a series of anagrams and, for each solution, to indicate whether that solution was the result of an insight (i.e., sudden) or a conscious search strategy. Interesting differences were observed between participants whose self-reports indicated high levels of insight-based solutions versus those with lower rates of insight-based solutions. Less high alpha and low beta activity was observed over occipital regions in participants exhibiting high rates of insight than in their low-insight counterparts; these differences support prior theories of creativity suggesting that diffuse attention is a key factor in the occurrence of insight-based solutions. Further, high-insight subjects exhibited more right-hemisphere activity over frontal and temporal regions and low-insight subjects exhibited more left-hemisphere activity over frontal and temporal regions. These differences suggest that brain activity during a resting state may influence the form of strategy implemented during problem solving. Researchers have also observed links between the neural activity immediately preceding the presentation of a problem (the pretrial interval) and the type of strategy (insight or search) applied during that problem's solution. In sum, these results suggest that cognitive, affective, and neural factors may be predictive of creative thought. It should be noted, however, that studies exploring individual differences in this area have focused primarily on the divergent-thinking components of creativity and insight, leaving many aspects of creativity and related processes unexplored.

New Thoughts and Directions

Conventional thought concerning the unconscious and its role in creativity and related processes has at times restricted progress in this exciting area. Researchers have been criticized for being narrow-minded when considering the cognitive (diffuse attention) and neural (right-hemisphere) components linked to creative insights. For a substantial period of the field's history there was little variety in the types of stimuli used in research or, until recently, the techniques at our disposal to study the unconscious and its role in creativity. A re-interest in creativity and related fields such as insight, has coincided with a change in thinking

about the subject area leading to promising research in new areas.

It is difficult to review all of the interesting new developments in the study of creativity so only a selection will be provided here. One area of interest is the recent exploration of the purported link between creativity and psychopathology, previously supported almost exclusively via anecdotal data. Although the results of several correlational studies were promising, suggesting a link between creativity and psychopathology, there was little to indicate the exact source of the relationship. Research has begun to examine the connection – studies exploring sensory gating in the first-degree relative of schizophrenics and in normal adults high in creativity have shown that both groups exhibit reduced levels of sensory gating when compared to normal adults who are low in creativity. Also of interest is the finding that normal adults high in traits associated with schizophrenia and those high in creativity both exhibit a more diffuse network connecting semantic concepts than is present in normal adults lower in creativity, thus providing support for the diffuse association network proposed in Mednick's theory of creativity.

Another area gaining momentum is the study of neurotransmitters (chemicals that allow communication between neurons) and how altering neurotransmitter levels affects cognitive processes, such as creativity. Researchers have explored the effect of propranolol, a beta-adrenergic antagonist that blocks the effects of the neurotransmitter norepinephrine, on the ability to creatively solve problems. Over-activity by the noradrenergic system which supplies norepinephrine throughout the brain has been linked to an increase in anxiety and a decrease in cognitive flexibility. The administration of propranolol which blocks the effects of norepinephrine decreased solution times in individuals solving anagrams. In contrast, administration of a beta-adrenergic agonist increased effects in the brain linked to norepinephrine, thus decreasing cognitive flexibility during problem solving. Similar research has been conducted manipulating the neurotransmitter dopamine and exploring the effects of these manipulations on problem-solving ability.

A final area to note is research exploring how one's mindset, altered through the use of counterfactuals, affects creative thought. Counterfactuals are situations in which individuals reflect on how past events could have had a different outcome with only a slight alteration to the actual events. Counterfactuals can be those in which the thinker alters the situation by adding an event (i.e., if only this would have happened) or by subtracting an event (i.e., if only this had not happened). Counterfactual thinking has been shown to influence the nature of an individual's thought processes subsequent to the counterfactual, thereby changing one's mindset or approach to upcoming stimuli. Research conducted by Laura Kray, Keith Markman, and others has demonstrated that counterfactual thinking that is additive in nature facilitates the solution of creative problems, such as RAT problems, whereas counterfactual thinking that is subtractive in nature facilitates analytic thought. These results add promise to the theory that it is possible to alter our abilities for creative thought.

Though many new lines of research now exist in the study of the unconscious in creativity and problem solving, this remains a field in which

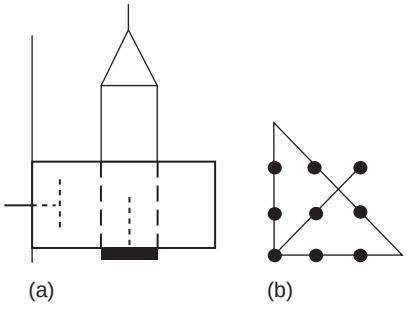


Figure 4 Classic insight problems – solutions. (a) The candle problem. Empty the tacks from the box and use the box as a ledge to support the candle. Reprinted from Memory & Cognition, Vol 32, Jessica I. Fleck and Robert W. Weisberg, The use of verbal protocols as data: An analysis of insight in the candle problem, pp. 990–1006, Copyright (2004), with permission from The Psychonomic Society. (b) The nine-dot problem. Reprinted from the Journal of Experimental Psychology: General, Vol 110, Robert W. Weisberg, and Joseph W. Alba, An examination of the alleged role of 'fixation' in the solution of several 'insight' problems, pp. 169–192, Copyright (1981), with permission from the American Psychological Association. (c) The two-string problem. Tie the pliers to one of the strings and get the string to swing back and forth like a pendulum. You would then be able to pull the second string over as far as possible and wait for the first string to swing within your reach.

there are more unknowns than knowns. It is, therefore, an exciting time to be part of this growing research area. An expanding view of creativity and insight suggests that both conscious and unconscious aspects of thought may result in creativity. Further, it appears that both forms are necessary in some capacity. But is there a fundamental difference between creative achievements that rely primarily on the unconscious and those that result predominantly from conscious thought? Another question of considerable interest is how can we increase the likelihood of creative achievements? Regardless of the domain, business, education, the arts, or the military, the need to think creatively and flexibly is a necessity for successful performance. How can we alter our abilities to think creatively if it is the unconscious and not the conscious that affects our output? These are just some of many the questions waiting to be answered as we move forward.

See also: Attention: Selective Attention and Consciousness; Mental Representation and Consciousness; Unconscious Cognition.

Suggested Readings

Bowden EM, Jung-Beeman M, Fleck J, and Kounios J (2005) New approaches to demystifying insight. Trends in Cognitive Sciences 9: 322–328.

Bowers KS, Regehr G, Balthazard C, and Parker K (1990) Intuition in the context of discovery. Cognitive Psychology 22: 72–110.

Dijksterhuis A and Meurs T (2006) Where creativity resides: The generative power of unconscious thought. Consciousness and Cognition 15: 135–146.

Fink A, Benedek M, Grabner RH, Staudt B, and Neubauer AC (2007) Creativity meets neuroscience: Experimental tasks for the neuroscientific study of creativity. Methods 42: 68–76.

Jung-Beeman M, Bowden EM, Haberman J, et al. (2004) Neural activity when people solve verbal problems with insight. PLoS Biology 2: 500–510.

Kounios J, Fleck JI, Green DL, et al. (2008) The origins of insight in resting-state brain activity. Neuropsychologia 46: 281–291.

Luo J and Knoblich G (2007) Studying insight with neuroscientific methods. Methods 42: 77–86.

Metcalfe J and Wiebe D (1987) Intuition in insight and noninsight problem solving. Memory & Cognition 15: 238–246.

Ohlsson S (1992) Information-processing explanations of insight and related phenomena. In: Keane MT and Gilhooly KJ (eds.) Advances in the Psychology of Thinking, pp. 1–44. New York: Harvester Wheatsheaf.

Seifert CM, Meyer DE, Davidson N, Patalano AL, and Yaniv I (1995) Demystification of cognitive insight: Opportunistic assimilation and the prepared-mind perspective. In: Sternberg RJ and Davidson JE (eds.) The Nature of Insight, pp. 65–124. Cambridge, MA: MIT Press.

Smith SM and Blankenship SE (1989) Incubation effects. Bulletin of the Psychonomic Society 27: 311–314.

Wallas G (1926) The Art of Thought. London: Cape.

Weisberg RW (2006) Creativity: Understanding Innovation in Problem Solving, Science, Invention, and the Arts, pp. 386–446. Hoboken, NJ: John Wiley & Sons, Inc.

Biographical Sketch

Jessica Fleck is currently an assistant professor of psychology at the Richard Stockton College of New Jersey. She completed her undergraduate education at Shippensburg University, majoring in psychology and criminal justice. She then obtained her PhD in brain, behavior, and cognition from Temple University, studying under Robert Weisberg. Prior to joining the faculty at Stockton, Fleck was a postdoctoral research associate in Drexel University's Electrophysiological Laboratory, working with John Kounios.

Her published research includes topics such as the cognitive and neural components of problem solving, creativity, and working memory, as well as the relationship between creativity and personality. At present, Fleck is researching cognitive and neural factors in the link between creativity and schizophrenia.

John Kounios attended Haverford College, where he received his BA with a double major in psychology and music theory/composition. He then enrolled in the University of Michigan's graduate program in experimental psychology, where he received his PhD under David E. Meyer. Kounios has held research and faculty positions at Princeton University, Tufts University, the Boston Veterans Affairs Medical Center, the University of Pennsylvania, and Drexel University. He has published research on a variety of topics in cognitive psychology and cognitive neuroscience, including knowledge representation, insight in problem solving, creativity, and episodic memory. He is currently a professor of psychology at Drexel University where his research focuses on the neural and cognitive bases of semantic information processing, problem solving, and creativity. His research has been funded by grants from the National Institute of Mental Health, the National Institute of Deafness and Other Communication Disorders, and the National Institute of Aging. His research has been reported by National Public Radio, *US News & World Report*, *Scientific American*, *The Times* (London), *The Wall Street Journal*, and other print and electronic media.

Language and Consciousness

P Gordon, University of Columbia, New York, NY, USA

Introduction

Although language is often considered to be the very medium by which our thoughts are made conscious, the extent of conscious control of language may be more illusion than reality. In thinking about the role of conscious and unconscious processes in language, one is reminded of Freud s metaphor of the mind being like an iceberg: Consciousness is represented as the relatively small exposed surface of the iceberg that is supported by vast structures below the surface representing the unconscious mind. Although Freud s metaphor refers to the dynamic processes driving our psychosocial interactions, the metaphor is nevertheless apt in thinking of the vast hidden structures that support the ability to produce and comprehend utterances in one s native tongue. To explore the relationships between language and consciousness, we must first understand what kinds of categories of conscious experience are linguistic in nature, which experiences are made available or highlighted by language, and which are left under the waters of the unconscious mind.

Conscious Linguistic Processes

Linguistic processes that impinge on our conscious experience require a medium through which the phenomena or qualia of consciousness can be experienced. Such media include the familiar modalities of speech, hearing, reading, and writing. In nontraditional groups or cultures, members might use other means of communication such as signing in the deaf, Braille in the blind, and Tadoma in deaf-blind persons. In some languages, the familiar speech segments of consonants and vowels are augmented by meaningful tonal structures, where the same syllable spoken with low, mid, high, rising, or falling intonation might each have distinct meanings. There is a classic sentence in Thai that means: Who sells chicken eggs? but to most nontonal speakers the sentence sounds like the same syllable repeated four times. Consciousness of language therefore extends to consciousness of distinctions in one s native language and is affected by the meanings we assign to different sounds and qualities of sound.

In other cultures, such as the Piraha of Lowland Amazonia, communication with children may occur through the use of hum speech that finesses a cultural convention against directly speaking to children. In hum speech, speakers hum the tonal structure of the intended utterance without use of vowels or consonants. Whistle speech in the Piraha is a similar mechanism for communication between adults when hunting in the jungle or communicating over long distances where the discrete syllabic elements of speech are deleted and the overall tonal contour of the utterance is communicated instead. Lest we think that such communicative modes are exotically restricted to the jungles of Brazil, consider the communicative act that consists of: uh uh uh (low, high, rising), where the message is clearly understood as I don t know, often uttered by bored, put-upon teenagers. In such cases, the conscious experience of such signals is completely different when they are interpreted as linguistic than when they are received by someone unfamiliar with the language and its conventions.

In addition to the modes of communication that transmit the actual interpretable symbolic structures of the language, we also communicate with paralinguistic gestures that do not contain discrete linguistic elements such as words, affixes, and sentences, but nonetheless modulate the meaning of a message. These paralinguistic elements include manual, facial, and intonational gestures that may accompany or replace spoken language. An utterance can have the completely opposite meaning depending on whether one uses a neutral tone or a sarcastic one (compare That s really nice of you

spoken in an enthusiastic/neutral tone versus a sarcastic tone). Emphasizing one word in a sentence can also radically alter the intention of the message (note the different intentions of a hat sales person saying: You want that hat? and You want THAT hat!). In a similar manner, facial gestures, smiling, showing disgust, or doubt, can alter the intended meaning of an utterance as well. We can infer that experience of messages that have distinct meanings in this way are distinct conscious experiences to the extent that meaning modulates or scaffolds the conscious experience of linguistic qualia.

Manual gestures are also known to be a rich source of information accompanying speech. While such gestures normally augment the spoken utterances, research has shown that the two modalities are sometimes mismatched within the same communicative exchange: People can say one thing while gesturing the opposite. In a remarkable experiment by Susan Goldin-Meadow and her colleagues, children were asked to explain how quantities are affected by changing the configuration of the items in Piaget s famous conservation task. Children who were going through the beginnings of a cognitive reorganization employed manual gestures that indicated knowledge of the correct, more advanced, solution, even though their simultaneous, spoken explanations were more immature reflecting earlier stages of cognitive development. This mismatch between language and gesture suggests that unconscious thought processes can bleed through into modalities that are not being utilized by the more conscious process of communication.

The idea that there can be uncoordinated parallel processing of information where only one channel can be verified through conscious linguistic report is reminiscent of earlier studies by Roger Sperry and Michael Gazzaniga and their colleagues on split brain patients. In these studies, patients had undergone commissurotomy a surgical procedure that severs the corpus callosum, a bundle of nerve fibers connecting the two cortical hemispheres of the brain. This procedure, which is carried out to relieve massive neural cross firing during epileptic seizures, made it impossible for the two sides of the brain to talk to each other. Experiments with these patients examined how the two hemispheres processed information by presenting stimuli separately using a split screen presentation so that information presented in one half of the visual field could not be seen by the other side of the brain. Information presented in the right visual field was sent only to the left hemisphere, which is the hemisphere that is primarily responsible for language. Similarly, information presented in the left visual field was processed only by the right hemisphere, which is more responsible for analyzing spatial and other holistic information about the world. When words were presented to the left visual field, split-brain patients were unable to name them. However, because the right hemisphere is able to read words, but not speak them, the patients were able to identify the meaning of a word like spoon by reaching onto a table and picking up the appropriate object. Like the case of the asynchrony between language and gesture, the case of split brain processing of language suggests a similar split in the consciousness of reading processes.

Other phenomena exist where language lurks beneath conscious experience and can suddenly pop up into consciousness without deliberate control. The tip-of-the-tongue phenomenon involves the familiar experience of searching for a word to fit a particular context that you just cannot find in your mental lexicon. In an early study of this phenomenon, the psychologist, Roger Brown, showed that the word can be accessed and brought to consciousness by providing clues such as the initial sounds of the word. Sometimes, trying to find a word seems to actually inhibit success, and it is not until we stop trying that the word itself pops into consciousness. This phenomenon suggests that when one actively searches for a word whose semantic properties (meaning) fit in a particular context, one is reactivating the semantic features that are already activated by the context. It is possible that reactivating the semantic features in this manner inhibits activation of the phonological features that allow you to access the form of the word. The strategy of searching for the word by going through the alphabet can often be more fruitful in such cases because it focuses on phonologically rather than semantic properties. In the tip

of the tongue phenomenon, we see a particularly striking experience that highlights the border between unconscious and conscious processing of language.

In summary, consciousness of language requires a sensory medium through which the linguistic signals can be experienced at the level of qualia. These signals can take on many forms and are as varied as the extent of human ingenuity. Conscious awareness of language is not just of the experience of shapes, movements, and sounds that are used to transmit the signals, but it is the meaningful interpretation of those shapes, movements, and sounds as words, sentences, and speech acts. Meaning can be like an attentional spotlight that focuses the mind on certain aspects of the signal, albeit unconsciously, but the resulting conscious experience can be very different depending on how the behaviors are interpreted. Awareness of the intent of the message often means interpreting subtle cues from multiple sources. Gesture, expression, and tone of voice conspire to modulate the meaning of the message as a whole. We might say, in such cases that there is an effect on the level of the speech act. For the most part, unless a gesture or expression is completely obvious or highly conventionalized such as the profane cultural forms (middle fingers, V-shaped fingers, pumping the bicep, etc.), the processing of information from these paralinguistic channels is often not done consciously and requires expert scientific study in order to fully understand the meanings behind them. Sometimes, a totally artificial form of encoding language might come along, and not only are the signals experienced within the sensory domain of their transmission, but higher order properties are also picked up through the analysis of invariant properties across situations of use. For example, radio operators who are experts in using Morse code a sequence of short and long beeps are able to recognize the identity of the person sending the message based on their signature rate and rhythm in hitting the telegraph key much in the same way as if they were recognizing a familiar voice. In fact, when one recognizes a familiar voice, there is a special conscious experience of this voice in which the conceptual properties of the whole person are experienced at some level.

Unconscious Aspects of Language

Below the level of communication at which the conscious signals are processed, there are multiple levels of unconscious processing of linguistic inputs and outputs. These include levels of phonology (sound structure), morphology (word structure), syntax (sentence/utterance structure), semantics (meaning), and pragmatics (appropriate and conventional social use of language). When we attempt to bring these levels of language processing to conscious introspection, we refer to this as metalinguistic knowledge (knowledge about our implicit knowledge of language). Because of the unconscious and inaccessible nature of linguistic structures and processes, attempts at conscious reflection of language are often fraught with possibilities for error. More often than not, those who claim to be language experts the so-called language mavens are not trained in formal linguistic analysis show confusions about many aspects of grammar and thereby fail to grasp the reason why certain constructions might or might not sound acceptable in the language.

One of the major advances in our understanding of language came in the latter half of the twentieth century from Noam Chomsky s approach to linguistics, which became known as Generative Grammar. Chomsky s proposals represented the beginnings of modern cognitive science in which cognition as mental representation is characterized as a set of algorithms or rules in which symbols are formally manipulated (moved, rearranged, rewritten, etc.). Implicit in the theory of generative grammar is the idea that languages have a mathematical-like structure that generates the sentences and utterances that we use in everyday life, and that facts about the underlying grammar can be discovered by using a formal analysis of linguistic structure. Grammar is said to consist of category symbols like N(oun) and V(erb), which are like variables in algebra (x, y). Rewriting category symbols as words $(N \rightarrow \mathit{dog})$ is like substituting a variable with a numerical value $(x = 5)$. The method for discovering the systematic principles of grammar involves formulating rules and seeing if they generate grammatical or ungrammatical sentences when words are substituted as values for the category symbols. Thus, grammaticality

judgments became the metric for judging whether a particular grammatical hypothesis was a good fit to the language.

Grammaticality, the conscious perception of a sentence as either sounding good or bad, became the data for the science of linguistics. Sometimes the judgments were extremely subtle and led to remarkable insights into the nature of grammatical competence. For example, changing the wording every so slightly in the following two sentences leads to strikingly different perceptions of grammaticality:

Who did Mary tell the man that Bill hit?
*Who did Mary tell the story that Bill hit?

The asterisk indicates the common judgment that the second sentence sounds bad compared to the first. Yet both would appear to be derived from questioning of perfectly grammatical sentences:

Mary told the man that Bill hit Tom
Mary told the story that Bill hit Tom

In both cases, the questioned element (Tom) is substituted with a wh-question form (who) and is moved to the front of the sentence. Yet, for some reason, it is only the first sentence that leads to the experience that it is part of the English language. In this instance, we have a consciously differentiated experience of two sentences that are remarkably similar in their actual wording. In one case we experience a regular sentence that is part of our language, and in the second case, we experience a string of words that is jarring and outside of the permissible word combinations of the English language. In such cases, speakers of a language generally agree about their experiences of whether something sounds good or sounds bad, but we are completely unable to articulate the reasons that we share those common judgments, since their origins lie within the hidden depths of our unconscious language processing systems.

As mentioned previously, when people attempt to formulate rules of language through their own conscious introspection and intuition, they do not access the level of analysis explained by generative grammar, and hence tend to invoke prescriptive rules that are made up by language mavens and other guardians of the mother tongue. Sometimes, language experts make disastrous pronouncements about what is, or is not, proper grammar. For example, African American English was called illogical because it failed to show verb agreement with the subject of the sentence (e.g., I be going); whereas a deeper analysis of such vernaculars revealed a language that was every bit as complex and logically consistent as the received dialect of Standard English; it was just a different set of rules. In cases like this, formally trained linguists generally agree that the explanation of why language sounds good or sounds bad, should not be left to the language mavens, but should be explored using the formal tools of linguistic analysis that can gain insight into the unconscious representations of grammar that undergird our conscious intuitions about grammaticality.

The rise of generative grammar in the late 1950s and early 1960s stood in stark contrast to the then dominant behaviorist approaches within psychology that eschewed any study of mental processes as mediating stimulus and response. Along with the concomitant rise of the digital computer in society, there was a trend for researchers in the cognitive sciences to begin to think about the mind as a kind of computer. Unconscious thought processes were like the hidden software codes that generate the programs that we interface with on the computer screen. The computer metaphor for the human mind is apt in many ways for thinking about how consciousness interfaces with the rest of cognition. What appears on the monitor screen of a computer is a useful, if not complete, analogy for what makes it to consciousness in the human mind. In a computer, there are whole unseen worlds of complex computation slaving away inside the memory and processing chips that allow us to interface at only a superficial level with what appears on the screen as we breeze through writing a document using a text editor interface. If all of the computational codes were to appear on the screen, we would not be able to function effectively. So it is with language, that mental processes are churning away to allow us to use language with little or no effort, yet we are not consciously aware of the machinery that permits us to speak in grammatical sentences and to present thoughts in a conversation.

When Chomsky began to characterize the extremely complex nature of grammatical computation, he noted the unconscious nature of the

knowledge underlying grammar, and likened it to the idea of tacit knowledge explored by the philosopher, Michael Polanyi. In his writings, Polanyi explored ways in which knowledge can guide our behavior in a manner that is not open to conscious introspection. So it is, that the Freudian iceberg is appropriate for describing the unconscious computation that goes into producing the conscious experience of language that is quite marginal in many ways.

Modularity and Consciousness

The relationship between conscious and unconscious aspects of language can be usefully addressed by employing the concept of modularity, which was proposed by the psychologist and philosopher, Jerry Fodor in the 1980s. Fodor proposed that the mind is composed of a set of horizontal and vertical faculties. The horizontal faculties are general-purpose processes that cut across domains. These might include mental abilities such as memory, reasoning, and judgment. For all intents and purposes, such faculties can potentially draw on information of any kind to form associations and to make inferences. Vertical faculties, on the other hand, are restricted to a specific domain of computation. In Fodor s original conception, these vertical faculties, which he called modules, were small in number, consisting of a few core processes such as perceptual processing in the sensory domains (vision, hearing, etc.) and language. Modules are characterized as being dumb but fast. They are automatic and epistemologically encapsulated. In other words, the mind accomplishes rapid processing within the module by processes that specialize in formal computation over the elements in the system and are immune to external influences arising from general knowledge.

A classic example of modularity in the domain of vision is the Muller Lyer illusion shown in Figure 1. In this illusion, we perceive two lines as being unequal in length even though we know them to be of equal length (this can be easily demonstrated by removing the biasing end pieces). Even the most seasoned vision scientists, who have seen the illusion thousands of times cannot persuade themselves to consciously perceive the lines

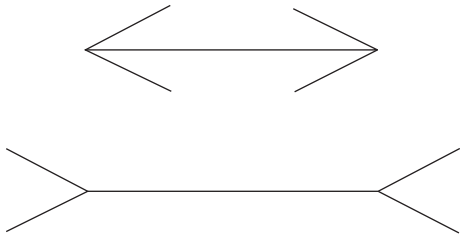


Figure 1 The Muller–Lyer illusion.

as equal. Instead, the visual system is bound by automatic processes that cannot help but produce a conscious perception of unequal length. This is because the vision module can only follow the unconscious computational rules of seeing and is destined to ignore any conscious knowledge we may possess of the actual state of world (i.e., consciously knowing that the lines are equal in length). In this case, we see that dual consciousness being conscious of the state of the world, and conscious perception of a stimulus does not guarantee their mutual influence and coherence. Rather, the conscious output of the modular process takes precedence and is immutable, and multiple mutually incompatible conscious states can coexist.

The claim that language processing is similarly modular in nature suggests that the outputs of linguistic processes are also automatically computed and are isolated from general reasoning processes. Some researchers have suggested that parsing biases in language show characteristics of modularity. For example, there is an extensive literature on a phenomenon known as minimal attachment. This is a principle that concerns a certain bias in the human grammatical parser for speakers of English. For example, in certain constructions that contain relative clauses, we may find sentences where the meaning is ambiguous depending on what part of the sentence the clause is thought to modify. For example, in the sentence:

The spy saw the cop with the binoculars, there is an ambiguity as to whether the spy or the cop had the binoculars. Experiments with such sentences suggest that there is a preference to attach the subordinate clause, with the binoculars, so as to produce the least amount of additional structure. In this case, it means that the preference is to interpret the sentence as the cop having

the binoculars despite the possible bias in meaning that would associate the spy seeing with binoculars.

Sometimes the effect of modularity on the conscious perception of language is more obvious at the level of phonology the perceived sounds of language. One example of how language channels conscious perception of language is in the difference between perceiving the sounds of one s own native language and that of an unfamiliar foreign language. Fodor gives the example that native speakers of English generally know what Chinese sounds like and what French sounds like, but it is harder to have a sense of what English sounds like and similarly for whatever native language one might speak. The inability to perceive a general sound of one s native language based on its generalized tonal and rhythmic properties suggests that, for competent speakers of a language, there is an obligatory processing of that language whereby sounds are converted into meanings and there is no conscious access to generalized phonological properties that would constitute the sound of the language. Such experiences are afforded only when the sounds are left uninterpreted. When listening to one s own native language, processing of that language into meaning is obligatory. We cannot help but convert linguistic signals into meanings in the same way that we cannot help but see an object as an object when it is presented to our visual system.

The output of a language comprehension module is a set of structured representations of meanings of the propositions encoded by the sentences being received. Conscious perception of the abstract sound characteristics of the speech in which this message is encoded is minimal. This effect is rather like the inability of the artistically untrained eye to perceive the formal structure of an object and its relation to the perspective of a scene. The awkward drawings of the untrained amateur are notoriously unsuccessful because they are guided by the attempt to capture the meaning of an object in its environment rather than capturing its actual formal linear properties, which require training to divorce oneself from the scene and see the lines and the structure of objects in space rather than to obey one s intuition. The old trick of copying a drawing upside down serves to excise meaning from a scene and allows the budding artist to concentrate on form. In like manner, a copy editor might read a text backwards to detect typos like repeated words that might otherwise go undetected because the language processor focuses on meaning rather than form.

Modularity makes language processing an obligatory process. It is only when we hear an unfamiliar language that we consciously perceive the detailed contour of the sounds unaffected by meaning. We cannot choose to perceive English or some other native language as simply a sequence of sounds, but rather are required to interpret those sounds as propositions and speech acts, first and foremost. In the written domain, there is a similar effect in the Stroop task, where words naming colors are presented in colored fonts that are distinct from the name of the color encoded in the word. Suppressing the color word being read and naming only the color of its font is one of the most difficult tasks in the psychological test arsenal. Thus, in the modality of reading, the processing of language as meaningful is obligatory and suppresses conscious attention to the physical characteristics of the stimulus, in the case of the Stroop task, the color of the font.

The Role of Consciousness in Symbolic Representation

While we have explored the idea that unconscious mental computation of linguistic forms sends up to consciousness a mental representation of the meaning of a word or utterance, there is almost no scientific understanding of the notion of meaningfulness itself. Language researchers who study meaning or semantics often analyze how meanings relate to each other and how we can draw inferences from linguistic utterances. For example, researchers might develop a theoretical framework that explains how we represent that a dog is a kind of animal, and that telling someone to do something and persuading them to do so have different implications regarding whether or not they actually did it only the latter entails that the action was completed. However, such approaches only analyze the formal relations between meanings, not the nature of meaning itself. Similarly, the

analysis of grammatical structure and the claims about modularity of such processes rely on a formal analysis of the syntactic structure of the language. Formal analyses of grammar can tell us how we put strings of words together to form sentences, but the analyses do not really tell us anything about how those sentences allow us to refer to objects, events, feelings, and so on that relate language to the real world. Nobody really understands how to capture what it means for a brain to possess meanings for words such as dog, persuade, or virtue. Such puzzles have plagued philosophers since the time of Plato and do not have recommendable solutions. Imagine trying to build a robot that was able to recognize instances of virtue in the world. What would its software program look for and how would it interact with the world to be able to do so?

Perhaps the strongest critique of the purely formal analysis of language comes from the philosopher, John Searle, who illustrates the emptiness of formal linguistic analysis with the thought experiment known as the Chinese room. In this thought experiment, we imagine a person who does not speak Chinese, and is placed in a room where they receive pieces of paper with Chinese writing on them. When these pieces of paper are received, the person in the room consult a book containing a set of rules that allows them to respond appropriately, again using Chinese symbols. The moral of this thought experiment is that, while the inhabitant of the Chinese room functions in a manner that is indistinguishable from an actual speaker of Chinese, we cannot say that this person understands Chinese, or that there is anything about the room itself that understands Chinese. This case extends as an analogy to the approach in cognitive science that characterizes linguistic comprehension and production as the manipulation of formal symbolic structures.

According to Searle, understanding a language requires a conscious individual that stands in a particular relationship to the formal symbolic structures in the language. He notes that a symbol is a formal element that represents something (the word dog is a symbol that represents instances of dogs in the world, and the word virtue represents instances of virtue in the world). But symbols cannot function to link language and the world without some sentient individual who is able to understand the relationship between the symbol and what it represents. In other words, a symbol is not a symbol without some individual who can recognize it as such. Otherwise, it is just some meaningless shape or sound. In this respect, Searle suggests that language understanding cannot occur in the absence of consciousness. Otherwise, we are dealing with a version of the Chinese room inside our heads. The argument is a powerful indictment of the claim that we can understand language and cognition in general as a process of symbol manipulation alone.

How Language Affects Conscious Perception of the World

In previous sections, we have illustrated how language can shape the forms of consciousness that we experience when we hear language spoken or see language written or gestured. First, the very fact that we perceive sounds as meaningful language means that we hear those sounds in a very different way than if they were uninterpreted. Since the very early study of speech sound processing, researchers have been interested in how language affects our perception of speech sounds, as opposed to other sounds. The perception of distinctions between speech sounds is shaped by whether those sounds are used in language or not. We are all familiar with the common alterations in the phonological systems that are made by foreign speakers attempting to speak English, and, of course, the same would be true of English speakers attempting to speak in a foreign language. Speakers of a particular language tend to show common changes to a foreign tongue. French speakers might say ze instead of the, Japanese speakers might say psychorogy instead of psychology. The commonality of such transformations by foreign speakers highlights the differences between the inventory of sounds employed in the native tongue versus the phonology of the language in which they are attempting to speak. Since the th sound is not available in French, speakers fall back on the closest equivalent, which is produced with the tongue slightly further back in the mouth resulting in the z sound.

In addition to being unable to produce the target sound of a foreign language, such attempts are often confounded by the inability to hear sound distinctions required in the foreign language. For example, when Japanese speakers substitute r for l (and vice versa) in English words, it is partly because they do not hear this distinction. Instead, the perception is slotted into the existing Japanese phoneme (speech sound) that consists of a sound somewhere in between r and l. In this case, we see that speaking a particular language focuses our conscious perception of speech sounds into categories of sound that are used in our own language and this makes us unaware of sound categories and distinctions that are used in other languages that employ different sound distinctions.

Research on speech perception in infants has tracked the developmental trajectory of speech perception in exquisite detail. In many cases, the ability to distinguish between speech sounds is not acquired when a particular language is learned but is lost when the native language does not make use of a particular distinction. For example, Japanese infants would have no problem differentiating between river and liver, but that ability would be lost sometime between 6 and 12 months of age, as the language they are exposed to continually fails to make use of such a distinction. On the other hand, if the same babies were adopted into English-speaking families, then they would not lose their ability to discriminate these sounds, and they would be incorporated into the English sound system they would be acquiring. Of course, the same is true of infants born into English-speaking homes, who can distinguish sounds made in various exotic languages that their parents cannot perceive and they too would lose within the second half of the first year of life.

One other aspect of speech perception that has been of particular interest is the notion of categorical perception. Perception of speech sounds, in some respects, is similar to the perception of colors. In the case of color, we know that electromagnetic radiation reflected from objects impinge on our eyes as continuous differences in wave forms, the whole range of which can be seen in a rainbow or when a prism is used to disperse white light into its component wavelengths across the spectrum. Our perception of these wavelengths, however, is not as a continuously changing phenomenon, but as consciously distinct qualia, in the form of discrete colors. Our visual system cuts the continuous spectrum into discrete colors seen in the rainbow or the refraction of a prism. By analogy, there are also variations in speech signals that can be presented as continuously changing along a temporal or frequency spectral dimension, yet we perceive the changes not as continuously changing but as discrete categories much like we perceive colors as discrete. Such categorical perception occurs when people are presented with sounds that vary in the timing between the onset of a consonant and the vibration of the vocal cords signaling the following vowel. Hard, or voiced consonants like /b/ /d/ /g/ show almost no interval between the consonant and following vowel, when we hear a syllable like bah. The softer, unvoiced consonants like /p/, /t/, and /k/ have a period of silence between the vibration burst of the initial consonant and a following vowel, as in the syllable pah. When the timing of the consonant and vowel, know as the voice onset time (VOT), is artificially manipulated, humans perceive the change, not as a series of gradually changing sounds, but as one of two distinct speech sounds, either /b/ or /p/ depending on whether the VOT is greater than about 30 ms. Therefore, much in the same way that a continuously changing stimulus is perceived as a discrete set of categories, speech perception is like color perception in this respect, and the qualia of conscious perception in speech are those categories that are differentiated by the auditory system.

Perceiving the Structure in Language: Critical Periods in Language Learning

Effects of exposure to language found in speech perception can presumably be found in other modalities in which language is communicated. However, the visual modality of sign language presents special opportunities to study the effects of being exposed to language at different ages. Few children learn sign language at the age or with the amount of natural exposure that one would learning a native spoken language. Some learn sign later in

life when exposed to a signing community. Hearing signers usually choose to learn sign at a relatively late age to communicate with family members or friends or for professional reasons.

Extensive research on the acquisition of sign language by Elissa Newport, Ted Supalla, and their colleagues reveals differences in the way that early learners and late learners perceive and produce sign language. Research shows that a fully articulated sign language like American Sign Language (ASL) has the richness and discrete structure of any spoken language. Newport and Supalla have demonstrated that mime-like sign sequences that appear iconic rather than arbitrary in nature contain discrete elements that are often imperceptible to the hearing parents of deaf children who learn sign language later in life to communicate with their children. Young children who learn sign language as a native language are able to pick out the discrete structure in the sign system that is imperceptible to their parents and other late learners. For example, when a sign such as a flat hand shape is used as a sign for a vehicle, it can undergo motions that include paths in straight lines, zigzags, ups and downs, left and right turns, and so on. To the late learner, such motions appear iconic in nature. That is, they appear to refer by resemblance to their message. To the native signer, such sequences appear as discrete finite elements in a systematic use of limited hand motions. These motions are used to indicate linguistic elements of path, manner and direction, much in the same way that we use prepositions like at, across, to, toward, and so on, in English.

A similarly striking phenomenon has been found in Nicaraguan Sign Language (NSL), which only emerged as a conventional sign language in the 1980s after the Sandinista government attempted to develop centralized comprehensive education for deaf children and adults, who had previously been isolated from each other around the country. Initial attempts by educators to teach the students how to communicate were not successful. Instead, NSL spontaneously grew out of the attempts of young children to communicate with other deaf children. Because of this unique situation, it was possible to compare those who came to the deaf community early in life with those who came later in their development. The striking finding was that only the early learners managed to fully incorporate the grammatical structure of the new language into their communications. Later learners, whilst still communicating with others, were unable to internalize the grammatical functions in the same way.

Deaf signers who are isolated from a natural sign language develop their own home sign systems. These are iconic gesture systems that each individual deaf child invents to communicate with other (hearing) people around them (signs for objects will either involve pointing at the object if present or mimicking its shape or function; signs for actions generally involved mimicking the actions). Such gesture systems lack the subtle and complex grammatical structure of a conventional language. Consequently, those who relied on this deficient grammatical system and never learned to communicate fluently in NSL missed out on the subtle functions encoded by the grammatical structure.

Annie Senghas, a researcher who studied the emergence of NSL in the Nicaraguan deaf community, developed a simple task in which Nicaraguan deaf children and adults were tested for their ability to sign spatial positions of people and objects. Participants viewed a sequence of pictures of a man and a tree in nine different combinations of orientation and relative position (e.g., man to the left of the tree facing forward). Their task was to describe to another participant which of the nine configurations was present in each picture. Late learners of NSL were confused and could not signal relative configuration, nor could they see the directional components in the messages of other signers. On the other hand, the children who learned NSL with native-like competence developed a fast and efficient way to uniquely specify relative position and direction that was easily transmitted to other early signers of NSL. This ability to communicate rapidly and in agreement with others is not something that was explicitly articulated within the community as a prescribed form. In other words, there was no set of individuals who decided what the gesture forms were going to be that would encode the spatial functions within this system. It simply emerged from the natural process of young children being put into a communicative environment that employed gesture as a medium. As a result, a real language forms

that is capable of fully encoding discrete and complex meanings that are only interpretable by others who acquired the NSL in this manner. Those who came late to the language could not break the code.

In addition to the spontaneous nature of how this system emerged within the deaf community, native signers cannot easily explain the structure of the system. That is, they are not consciously aware of what they have created. For example, they cannot explain how the system of signs works that allows them to specify the relative positions of the elements in the Senghas task. This is rather like speakers of English who often fail to articulate the basic principles of grammar underlying their spoken competence, where their consciousness of language at the metalinguistic level fails. It is as if native signers of NSL come to a common consensus merely by dint of being exposed to the language early as a first language. It appears then that conscious perception of language differs depending on the circumstances under which that language is acquired.

To summarize, the perception of linguistic messages can be highly influenced by developmental experience with the language in question. Regarding effects of age of exposure, it appears that being exposed to a language early in life leads to induction of the structure underlying the formal structure of the language. This, in turn, leads to the perception of functions that are encoded by the structure of the grammar that are often not encoded by those exposed to language later in life. In other words, if language exposure occurs late in the developmental history of an individual, the grammar does not develop fully, and so the messages conveyed by the language heard or seen are deficient. Hence, the conscious interpretation of language appears to depend on the very subtle grammatical clues that are conveyed perhaps at a non-conscious level, but affect the conscious experience of the message.

Language, Thought, and Consciousness

The way that language itself is consciously experienced in its various forms is affected by the way that language shapes the medium within which it is transmitted. Whether it be the categorization of sounds or the ability to perceive discrete structure in the gestures of sign language, as humans we are highly suited (perhaps evolutionarily adapted) to pick up and create the regularities of linguistic structure. If experience with language can shape the way that we consciously perceive linguistic signals in their own medium, can experience with a particular language also shape the manner in which we experience reality outside of the context of speaking, signing, and listening to language? It is interesting to note, in the context of the Nicaraguan experience, that children and adults who learned the conventional sign systems of NSL also appeared to experience a change in the way that they thought about the world in general. In one telling comment from a BBC documentary about the phenomenon, a deaf Nicaraguan woman who had learned to sign commented that for the first time, she knew what it was to think. Is there a sense in which language causes us to think or at least shapes the way in which we think? Although researchers generally reject the idea that we think using language as a general rule, there remains the question of whether the structures of language create possibilities for thought that would not exist without the existence of language.

Before generative grammar became the predominant theory in linguistics, a highly influential approach to the study of language and thought in the middle of the twentieth century was the Sapir-Whorf hypothesis, named after the linguists Edward Sapir and his student, Benjamin Lee Whorf. This hypothesis proposed the principles of linguistic relativity and linguistic determinism. Linguistic relativity proposed that languages could differ radically in the manner by which they encode reality, and linguistic determinism proposed that language can causally determine how reality is perceived in nonlinguistic domains. Although highly influential for many years, the ideas developed by these linguists fell out of favor amongst linguists and psychologists for various reasons. In Whorf s writing about this topic, he often assumed what he set out to prove. That is, if a language encoded a particular mode of expression in a manner that seemed radically different from English, it was assumed that this shaped the way

that speakers thought about the world. Sometimes this involved taking grammatical elements too literally without knowing how they are actually interpreted in the minds of people in the culture. For example, when we say: I will go , there is no implication that there is any will (i.e., intentionality) involved in how we interpret this expression. Instead, it is just a neutral construction for future tense. Ironically, Whorf took the existence of a similar construction in the Navajo language to claim that Navajos did not have a conception of future time. Unfortunately, he failed to notice that the English future tense construction was entirely parallel and his reasoning would imply the English speakers also cannot conceive of future time. The general problem with the approach was a lack of independent evidence that differences in linguistic expression were accompanied by perceptual or conceptual differences that contrasted between speakers of radically different languages.

Studies that used empirical methods to test whether independent measures of perception and cognition were correlated with differences in linguistic structure focused initially on the domain of color names. It is well known that languages in some cultures use only sparse naming systems for colors and do not cover the color categories of the spectrum found in English. A language might have a word for dark and light and, perhaps, red and green. Obviously, in naming other colors, such terms would bleed over into yellow, blue, and so on, and so the direct translation into the English terms should not be taken literally. In the 1970s, Eleanor Rosch studied color memory in the Dani tribe of Melanesia, who possess only two color terms. In reporting the results, Rosch focused on the similarities between Dani speakers and English speakers. In particular, she found that Dani speakers showed better memory for focal, saturated colors when compared to nonfocal, unsaturated colors, even though they had no names for these colors. These results were used to argue that language was not necessary for organizing the color spectrum around the focal colors of the color categories found in English. More recently these results have been questioned and it has been pointed out that, when memory for colors is tested, there is a clear deficit found in cultures that have few color names. In addition, Rosch original results had design problems. When corrections were made for order effects in the tasks, there was a failure to replicate the similarity between the Dani and English speakers with respect to the advantage in memory for focal over nonfocal colors.

Debi Robertson studied color naming and memory in the Berinmo tribe of Papua, New Guinea. As in the case of the Dani, she similarly found very poor memory for colors and no effect for focal colors. Instead she found that performance was determined by how the language formed boundaries between colors. Unlike English, the Berinmo language showed no blue-green boundary, but instead had a boundary between names for colors noi and wor. Although inhabiting roughly the same area of the color spectrum as blue and green, these two color names covered very different ranges of colors. The category boundary between these two color names was what determined their memorability and naming properties. In other words, the way that words in the language carve up perceptual spaces can have a strong effect on how those spaces are perceived and committed to memory. In a clever study, Jonathan Winawer showed that Russian speakers are faster than English speakers in a discrimination task where one is required to distinguish between shades of blue. He attributed this difference to the fact that Russians have distinct words for light and dark blue, which are not found in English.

Language also has effects on the perception of number. Several cultures do not have the fully elaborated counting systems that we take for granted in developed western societies. This difference allows us to ask whether the lack of number words affects perception of number. Cultures that live in small communities with limited need for trading do not require the ability to keep track of numerical values for objects and individuals. One can keep track of individuals without actually counting them, but generally accurate exact tracking of individuals is restricted to small numbers up to about 3 or 4 depending on the demands of the task. The ability to keep track of small numbers is often called subitizing, which is thought to involve an immediate apprehension of the numerosity of a set in the absence of counting. In experiments with people who can count, researchers have designed tasks to prevent them from counting

using very rapid presentation of stimuli and/or having participants repeat the word the to prevent sub-vocal counting. In a culture that does not count, such measures are not necessary, and arrays of objects can be inspected for up to several seconds before performing a task.

Cross-cultural studies of numerical cognition have been performed with cultures that lack elaborated counting systems, including the Piraha and Munduruku tribes of Amazonia. The Piraha have two words for numbers, hoi and ho· (falling and rising tone respectively), which are relatively imprecise, but are conventionally described as a 1-2-many system. The Munduruku have number words up to 5, but again the values are also imprecise. In tasks that require exact appreciation of numerosity, speakers of these languages look much like undergraduates in experiments where counting is suppressed. They tend to show accurate appreciation for numerosities for quantities up to 3 or 4, and then show increasing error in their estimations with increasing set sizes of stimuli. In fact, the relation between set size and error, known as the coefficient of variation, is exactly the same in the Piraha who cannot count under any circumstances and undergraduates in New Jersey when they are prevented from counting. On the other hand, if the undergrads had been given the ample opportunity to inspect the arrays without counting suppression, they would have been flawless.

In one of the Piraha studies, participants viewed an array of 4 objects for an extended period, but then failed to pick that array when later presented with a choice of 4 versus 5 objects to obtain a reward. Such studies suggest that conscious perception of numerical quantity differs radically from that of counting cultures. It is almost impossible for us to fail to perceive 4 or 5 as a quantity given enough time to encode the numerosity of the set (see Figure 2). For a culture that does not count, the distinction between 4 and 5 is not at all distinct. The failure to distinguish quantities of 4 and 5 when the language does not encode such a distinction is quite parallel to the case of failing to distinguish between phonemes such as /r/ and /l/ when the language does not make use of such a distinction.

Other effects of language on concepts have been found in studies by Lera Boroditsky who

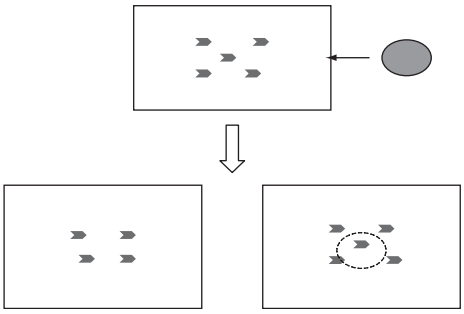


Figure 2 Number stimuli used with the Pirahã. A candy is placed in a box with five fish, and then presented alongside a box with four fish. Participants were very poor at picking the correct box.

compared languages that differentially assign masculine or feminine gender to arbitrary nouns. For example, if one language assigns feminine gender to the word for table, and another language assigns masculine gender to this word, speakers of the two languages will show differential responses when asked to choose adjectives that are associated with the word. When table is feminine, speakers will associate more feminine descriptors with tables, such as soft, gentle, and warm. When the same word is assigned to masculine gender by the grammar, speakers associate the word with masculine descriptors such as strong, aggressive, and dominant. In English, since we refer to boats as she we often find similar gender effects that have nothing to do with the biological gender of the objects themselves. However, whether this reflects a difference in the conscious experience of the objects, as opposed to merely producing covert conceptual associations is unclear.

General Conclusions

Language emerges into consciousness primarily through the medium in which it is transmitted. Such media are richly varied reflecting the ingenuity of the human mind to develop channels for communication. The structure of the communication system that we refer to as natural language requires a vast supporting structure of symbolic structures, rules, and principles that allow us to communicate in conventional ways. Much of this structure is unconscious in nature. One way to understand the interface between conscious and

unconscious linguistic processes is within the framework of modularity, which suggests that the linguistic processor acts in a primarily unconscious mode of symbol manipulation much like the inner workings of a digital computer. Where language does play a role in affecting conscious experiences is in focusing attention on aspects of sound, structure, and meaning. Experiments comparing languages that differ in how they use the available sounds for language, how they structure the elements of language, and how they denote conceptual spaces of color, number, and other domains suggests that having a language can have profound effects on how speakers consciously perceive the sights, sounds and concepts in their mental construction of conscious reality.

See also: Inner Speech and Consciousness; Intentionality and Consciousness; Meta-Awareness.

Memory: Errors, Constructive Processes, and Conscious Retrieval

W Hirstein, Elmhurst College, Elmhurst, IL, USA

This article focuses on aspects of the neuroscience and psychology of memory related to memory errors, the processes by which we form and reconstruct memories, as well as the conscious retrieval of memories. We examine two neurological syndromes in which patients report false memories, Korsakoff's syndrome and aneurysm of the anterior communicating artery, as well as false memory syndrome, which can affect normal children and adults. We are most interested in those aspects of memory that relate to consciousness, the place where memory and perception come together, an idea captured by Edelman's apt phrase, 'the remembered present.' Thoughts held in consciousness naturally attract relevant memories, so that they also enter into consciousness. Once the mnemonic material is retrieved, however, it must be carefully checked for accuracy and actual relevance to the current situation.

Implicit and Explicit Memory

The brain's many memory systems can be divided into two main types: implicit memory systems and explicit memory systems. The primary difference between the two is their relation to consciousness: explicit memory systems deliver information to consciousness in the form of thoughts or images, whereas implicit memory largely bypasses consciousness. Procedural memory is an example of a type of implicit memory. Procedural memory allows us to acquire skills, such as playing the piano or playing a sport. It functions largely without consciousness, and in fact conscious awareness can interfere with its workings, as evidenced by the infamous nasty trick sometimes played on fellow golfers by asking them whether they inhale or exhale when they swing. When the victim attempts to discern the answer it can cause the intricate pattern of muscle activations to disintegrate. Classical conditioning, of the type discovered by Pavlov, is also a form of implicit memory. We will focus here on a type of explicit memory known as episodic or autobiographical memory.

Can you remember having breakfast this morning? You need to employ your 'autobiographical memory' in order to do this. Autobiographical memory is a rough record of our personal experiences, usually from our point of view. Most researchers agree that autobiographical memory is either the same as, or a subset of, episodic memory. The autobiographical record is fragmentary, whole hours, then days, weeks, and even years can fall into the abyss of forgetting. It is an especially personal form of memory, not only because it records most indelibly those things of importance to us but also because losing it means losing a sense of ourselves, as anyone who has ever watched someone succumb to Alzheimer's can testify. The Alzheimer's patient eventually forgets you and claims you are someone else, or a stranger, and this flags another function of the autobiographical memory system: it records information about other people, places, and things that are significant to us.

With some of the things you know, the knowledge of when and where you first acquired that information is long gone. You know that cats have claws, but you most likely have no idea when you learned this. Other information does bring with it what researchers call a source memory: a type of episodic memory about when and where a memory was acquired. Source memory is a fragile thing, and we are prone to characteristic errors in source memory tasks. In one study, normal people and hospital patients with frontal cortical lesions learned the answers to a set of trivia questions. When they were tested a week later, the frontal patients had normal memory of the answers

themselves, but showed poor source memory, often claiming they had learned the answer at some earlier point in life.

The other type of explicit memory which does not necessarily come with a source tag is called semantic memory. It includes knowledge of facts, such as that the Eiffel Tower is in Paris, that Truman was a US president, and so on. Notice that these facts are represented from an impersonal perspective, whereas autobiographical memory is essentially personal. So far, researchers have been unable to cleanly separate the neural loci of semantic memory and episodic memory, and perhaps for good reason. The two memory systems interact in several ways, and some have suggested that they are merely different levels of categorization in the same memory store.

The Neuropsychology of Autobiographical Memory

The medial temporal lobe memory system includes the hippocampus formation and the adjacent parahippocampal and perirhinal cortices. The hippocampus is not the place where the content of memories is stored, but rather appears to contain a set of neural links to the content, which is distributed widely throughout the cortex. Memories of an episode in one's life typically contain information from more than one modality: vision, hearing, and even taste, touch, and smell. Each of these components is stored in a unimodal sensory area, for example, the visual components of an episodic memory are stored in the visual cortex in the occipital lobe, while the auditory components are stored in the auditory cortex in the temporal lobe. These distributed representations are linked to a central index in the hippocampus. When recent episodes are retrieved, the index is reactivated, causing activation to spread to each of the associated unimodal areas. This is more correct of recent episodes, however. Once a representation of an episode has been fully consolidated, activation can spread between the separate features themselves, so that hippocampal activation is no longer needed.

We are also beginning to gain an understanding of the brain areas that comprise the frontal components of the medial temporal lobe memory system. Medial temporal and hippocampal regions tend to be more involved in spatial context memory, while the frontocortical region, the diencephalon, and the temporal lobes are involved in temporal context memory. Much has also been learned about the neural bases of short-term memory systems located in the frontal lobes. Psychologists have had trouble determining whether there is one type of short-term memory, or several. The time span involved – exactly what 'short' means – is also not widely agreed upon. In the 1980s, however, neuroscientists began exploring a large area in the dorsolateral portion of the prefrontal lobes. This area seems to be responsible for a sort of memory that has been called 'working memory' – a concept that at least overlaps with the psychologist's concept of short-term memory.

The recent history of autobiographical memory research begins in the 1950s with the study of H.M., a man who developed severe amnesia after bilateral surgical removal of the medial temporal lobes (including most of the hippocampus, the parahippocampal gyrus, and the amygdala) in an effort to lessen the severity of his epilepsy. H.M. retained his basic intelligence and his personality but lost the ability to remember anything that happened to him after the operation. Researchers noticed, however, that H.M. could retain information for a short time, and could also acquire new motor skills such as mirror writing, solving puzzles, or tracing mazes, without knowing that he was doing so, a form of procedural memory.

Korsakoff's Syndrome

In 1887, Sergei Korsakoff observed mental problems in a group of alcoholic patients he was treating, including memory loss, anxiety, fear, depression, irritability, and confabulation. We have since learned that the syndrome is caused by a lack of vitamin B1, or thiamine, and not directly by alcohol itself. Korsakoff's syndrome can come on quickly, after an alcoholic coma, or it can progress slowly over many years. The syndrome occurs primarily in alcoholics, but may also occur in other cases where the patient's digestive system fails to absorb B1, in conditions known as the malabsorbtion syndrome and regional enteritis, as well as with cancer of the stomach. The means by

which alcoholism causes Korsakoff's is not fully understood, but alcohol is known to interfere with transport of thiamine in the gastrointestinal tract, and chronic liver disease can also affect the liver's ability to store thiamine. Thiamine itself is important for proper brain metabolism, because chemicals derived from thiamine play a role in the synthesis of neurotransmitters, particularly acetylcholine, as well as gamma-amino-butyric-acid (GABA).

The memory loss in Korsakoff's is anterograde – the patients are unable to form new memories. Procedural memory is intact; for instance, the patients can still drive. Korsakoff's patients tend to underestimate the time they have spent in the hospital, as well as their own ages. Korsakoff himself successfully traced the memory reports of his patients to actual memories but found that the memories had been displaced in time by the patients. In the early phase of their illness the confabulations of Korsakoff's patients is typically an internally consistent account concerning the patient. The contest of this account is drawn fully or principally from the patient's recollection of his actual experiences, including his thoughts in the past. Korsakoff said of one of his patients,

> Telling of a trip she had made to Finland before her illness. . . [the patient] mixed into her story her recollections of the Crimea, and so it turned out that in Finland people always eat lamb and the inhabitants are Tartars.

Anterior Communicating Artery Aneurysms

Aneurysms of the anterior communicating artery (ACoA) – which distributes blood to portions of the ventromedial frontal lobes (including parts of the orbitofrontal lobes) and related structures, including the basal forebrain, the fornix, the septum, the anterior cingulate gyrus, and the corpus callosum – can also cause confabulation. The ACoA is a common site for brain aneurysms, which occur when the walls of a blood vessel are weakened by some insult, such as infection or degenerative illness. Often the vessel will rupture, causing a hemorrhage and destruction of the surrounding brain tissue. The ACoA forms the anterior link of the Circle of Willis, which is a ring of arteries that spans the inner portions of the two hemispheres, and interconnects the two large anterior cerebral arteries. While the ACoA is small, it feeds a variety of brain areas and organs, and damage to it may also seriously affect blood flow in one or both of the anterior cerebral arteries.

The important cognitive features of the classical ACoA syndrome are

1. *Memory loss.* Patients show anterograde amnesia as well as retrograde amnesia, often for a few years preceding the aneurysm. As in Korsakoff's, short-term memory appears to be intact. In tests of recognition memory, they can often correctly recognize, for example, famous people, at a normal level, but they can exhibit something called 'pathological false recognition,' that is, cases where they claim to recognize a stimulus they are actually seeing for the first time.
2. *Changes in personality.* Just as do Korsakoff's patients, ACoA patients have social interaction problems. Impulsivity, impatience, disinhibition, emotional lability, depression, problems in decision-making, and poor judgment in social situations have also been observed.
3. *Executive deficits.* These include perseveration, poor concept formation, problems with set shifting, reduced verbal fluency, and impairments in cognitive estimation.
4. *Confabulation.* Confabulation appears in the acute phase right after the aneurysm, and very often remains in the chronic phase. It can change from implausible 'spontaneous' confabulation in the acute phase to more plausible 'provoked' confabulation in the chronic phase. De Luca reports that questions such as "What did you do last night?" will usually produce confabulations from the patient. One reported, "We went to dinner in New York City with my brother who came in to visit us"; something that had happened, but years before. When a patient was asked how he would respond if he was told his report was wrong, he said, "I'd be surprised 'cause my experience, what I learn from my eyes and ears tell me differently. . . .I'd want some evidence."

The memory deficits caused by anterior communicating artery aneurysms and by lifelong drinking

in Korsakoff's syndrome hold a special interest for memory researchers because such deficits indicate that there are also important frontal components to the memory system. The anatomists confirm the idea that the frontal and temporal lobes work together to achieve memory, by finding that the areas that constitute the medial temporal lobe memory system have strong, reciprocal connections to at least two frontal areas. The sites of lesion in Korsakoff's and anterior communicating artery syndrome are clearly different from those involved in medial temporal lobe amnesia. There are corresponding differences between the temporal and frontal memory patients, the most important being that medial temporal lobe patients do not confabulate, and will admit their poor memories and pursue compensatory strategies. Medial temporal lobe patients have been found to be less likely than normal people to produce false memories on tasks, specifically designed to elicit them (see below). Medial temporal lobe amnesics show much higher latencies in giving their answers and make many more self-corrections than confabulating frontal memory patients in memory tasks, probably indicating intact executive processes struggling to correct degraded memories.

False Memories

We all experience the memory malfunctions one sees in different types of amnesics. We sometimes remember what we intended to say or do, rather than what we actually said or did. We frequently displace events in time upon recalling them. And then there is the interesting error of mistaking events that were merely dreamed about for real events, and the rarer case of mistaking memory of real events for memory of dreamt events. Recent trends in memory research have strongly confirmed what memory researchers have always known, that is, memorizing something is not at all like recording it, and the act of remembering is not the mere replaying of this recording. Memory is a selective and reconstructive process, which can go wrong in any of several ways.

The phrase 'false memory' is a bit of a contradiction in terms, of course, given that it is reasonable to hold that something is not a memory of any sort if it is not correct, but its meaning is clear enough. False memories can easily be produced in children by asking them leading questions, and by a procedure such as the following: Children were presented with a deck of cards, each of which described an event. Some of the events had actually happened to the children, while others had not. When they were repeatedly asked whether the false events had happened to them, a majority of the children eventually agreed that they had, and many of them embellished the event with confabulated details. Apparently the memory systems themselves have a basic accuracy level; then we use different frontal checking procedures to increase this level. As we noted, normal correct memories are rational reconstructions, in that the reconstruction process is guided by what seems rational to the person. One can see this in certain patterns of error in false memories, where something odd in an event is misremembered as something more normal or rational.

Young children exhibit some of the same patterns of memory problems that frontal patients show. This may be due to the fact that the frontal lobes are among the last cortical areas to mature. A large part of the development of the frontal lobes occurs between ages 5 and 10, and they do not fully mature until the teenage years. Perhaps nature's plan is that the checking processes described above will be instilled after birth, during the long training period we humans require, principally by our parents. What begins as an external loop is made internal: The child confabulates, the parent corrects, the child changes what he said. As we mature, we internalize these corrections so that the loop runs completely within our brains, although it shares some of the same dynamics: there is still a candidate claim, and there is still a check that has the power to inhibit the claim from actually being made.

Adults are also prone to false memories, however, given the right circumstances. The 'misinformation effect' is a way to induce false memories in adults in laboratory settings. In a typical such experiment, the subject will be shown a video depicting a staged crime, and then exposed to false information designed to interfere with his or her memories of the event. Subjects show a strong tendency to incorporate

this information when asked later to recount the event. Loftus and her colleagues have also shown that exposure to prejudicial information after having witnessed an event can influence the subject's later recall of that event. Garry's research group has shown that both imagining events that never happened and paraphrasing descriptions of such events can make one more likely to later report that those events actually happened. In another type of experiment, normal subjects are presented with a list of words related to sleep, excluding the word 'sleep' itself: 'bed,' 'rest,' 'awake,' 'tired,' 'dream,' 'wake,' 'snooze,' etc. When they were later tested, between 30% and 40% of the subjects claimed that they had seen the word 'sleep.' Researchers observed the brains of normal subjects using positron emission tomography (PET) as they performed a task in which subjects first heard a list of related words, and then were tested for memory of the words. The researchers were able to successfully differentiate correct from incorrect memories by their different patterns of activation. Subjects of hypnosis may also confabulate when they are asked to recall information associated with crimes, causing researchers to warn criminologists about the dangers of obtaining information from hypnotized subjects. There are also anecdotal reports of hypnotized subjects confabulating when asked why they did something in accord with their hypnotic suggestion. For instance, a stage hypnotist gives his subject the suggestion that he will wave his hands whenever he hears the word 'money.' When asked later why is he is waving his hands, he replies ''Oh, I just felt like stretching.''

Theories of Confabulation

Confabulation caused by frontal brain injury is the best-studied type of confabulation, and there are hints that some sort of frontal damage is essential for confabulation, so this is the best place to start our review of theories of confabulation. This damage produces confabulations about past events in the patient's life, which either did not happen, did not happen to him, or did not happen to him when he believes they did. A man with Korsakoff 's syndrome might claim that he was at work at his supermarket, finishing up the year-end inventory, for example, when he had been in bed at the hospital the entire time. Stated in term of individually testable criteria, the recommended definition of 'confabulation' is as follows: S confabulates (in claiming that p) if and only if: (1) S claims that p; (2) S believes that p; (3) S's thought that p is ill-grounded; (4) S does not know that her thought is ill-grounded; (5) S should know that her thought is ill-grounded; (6) S is confident that p. The concept of 'claiming' (rather than, for instance, 'saying' or 'asserting') is broad enough to cover a wide variety of responses by subjects and patients, including nonverbal responses, such as drawing and pointing. The second criterion captures the sincerity of confabulators. If explicitly asked, ''Do you believe that p?,'' they invariably answer yes. The third criterion refers to the problem that caused the flawed response to be generated: Processes within the relevant knowledge domain were not acting optimally. Criterion number four refers to a cognitive failure at a second phase, the failure to check and reject the flawed response. The fifth criterion captures a normative element in our concept of confabulation: If the confabulator's brain were functioning properly, she would know that the claim is ill-grounded, and not make it. The claims made are about things any normal person would easily get right. The sixth and last criterion refers to another important characteristic of confabulators observed in the clinic, that is, the serene certainty they have in their claims, even in the face of obvious disbelief by their listeners. This epistemic approach eliminates a problem endemic to the falsity criterion in the traditional definition, according to which confabulations are false memory reports: A patient might answer correctly out of luck. The problem is not so much the falsity of the patients' claims but rather their ill-groundedness and consequent unreliability, at least in the affected domain. In short then, in this epistemic view, to confabulate is to confidently make an ill-grounded claim that one should, but does not, know is ill-grounded.

With the increasing information available about how our memory systems work, the discussion of memory-based confabulation has grown increasingly sophisticated. One theme of great interest that comes up frequently in the literature is the idea that these types of confabulations might be caused by two separate malfunctions. First, the

patients have a memory problem, which they share with medial temporal lobe patients. But second, the patients have what is typically referred to as an executive problem, which is responsible for the failure to realize that the memories they are reporting are fictitious. In a particular case of confabulation, the two problems manifest as two 'phases': first a false memory is produced, but then frontal areas fail to perform functions that would allow the person to realize the falsity of the memory. Notice that this implies that the thoughts that give rise to confabulations exist as genuine beliefs in the patient's mind (as opposed to the patient merely finding a certain claim coming out of his mouth, without his actually believing it). This implies that the patient's confabulations are accurately reporting his (disordered or ill-grounded) conscious experience. Theories of the nature of the problem in memory confabulation are divided into two categories, depending on which of the two problems they emphasize:

1. *Retrieval theories*: According to these theorists, confabulation is caused by a deficit in the 'strategic retrieval' of memories. This causes a loss of the sense of the temporal order of one's memories, and of their 'sources' – the places and times they represent. Theories of this type trace all the way back to Korsakoff.
2. *Executive theories*: These theories typically acknowledge that there is an amnesia present, but add that confabulators are to be differentiated by their additional frontal damage. According to these theories, two different processes are damaged in confabulation: (1) A memory process and (2) An executive or 'monitoring' process. The executive processes fail to correct the false memory.

Cognition requires both representations and processes for manipulating those representations – these are executive processes. Executive processes perform many different operations on representations. Your memory itself is just a huge collection of representations; executive processes must control the search and reconstruction process that take place when we remember. Marcia Johnson's view, according to which confabulation is attributed to a deficit in a more general executive function she calls 'reality monitoring,' – the ability to distinguish real from imagined events – is an example of an executive theory. Normal people are able to differentiate real from spurious information at high success rates. This seems to be a learned, or at least a developed ability. Real memories, according to Johnson, can often be distinguished from mere imaginings by the amount of perceptual detail they contain, as well as by the presence of supporting memories, including information about where and when the remembered event occurred – source memory.

The search for the neural locus and functional nature of these executive processes begins with what is already known about executive processes affected when the lateral portions of the prefrontal cortex are damaged. One line of inquiry involves determining how often confabulation tends to occur in the presence or absence of other mental functions known to require frontal activity. In the Wisconsin Card Sorting Test, the patient must first sort cards by the color shown on them, then the rule is changed so that the cards are to be sorted by the shape of the figures on them, etc. Patients with lateral frontal damage get 'stuck' responding the first way, and are unable to change to the new sorting rule, a phenomenon known as 'perseveration.' However, several confabulating patients have been found who perform normally on the standard tests of frontal function, such as the Wisconsin Card Sorting Test, causing speculation that confabulation and failure on such tests are the result of damage to different frontal areas.

It may just be that retrieval theories and executive theories are merely directed at different parts of the confabulation process. The first phase involves the production of a false memory. The second phase involves failure to notice and correct the falsity. Retrieval theories focus on the failure to access the correct memory, while executive theories focus on the correction failure. The executive theorists typically attribute confabulation to a failure in what they call self-monitoring, or self-awareness.

Reality Monitoring

Confabulation may be due to a broader failure to test representations, whether they are from memory or not. The term 'reality monitoring' seems

misleading, since what is being monitored (or not) is a representation, often one to which no reality corresponds – 'representation monitoring' would be more accurate. According to Johnson, episodic memories of an event bind together elements of several different types, some of which represent properties of the event, while others represent features of us, for example, our thoughts or emotions in reaction to witnessing the event. These different properties include the following: colors, sounds, tastes, emotions, objects, and locations, and also information contained in semantic memory. Recall of any one of these features is often enough to draw the entire autobiographical memory back into awareness. When thoughts that have this rich detail present themselves as memories, this can be sufficient to make us regard them as genuine. Because of this, in the mind of a person with a vivid and detailed imagination, memories of imaginings can be mistaken for memories of actual events. According to the proponents of this approach, vivid imaginings can be mistaken for memories, for example, when one believes they said or did something that they only thought or imagined saying or doing.

Further checks can be made: We can check the consistency of the candidate memory with our set of beliefs. Monitoring can involve the noting of inconsistencies among representations currently in consciousness or between those and long-term knowledge. Confabulation patients tend not to notice or worry when they contradict themselves. One patient, for example, contradicted himself in the same sentence, saying that he had just visited a store he formerly owned, and then acknowledging that the store no longer exists. As early as 1915, Pick noted that Korsakoff's patients also have a deficiency in the need to correct contradictions.

People can intentionally tighten their monitoring standards when motivated to do so. Researchers often report that simply admonishing memory patients to be more careful can work to increase the accuracy level of their reported memories. It is interesting to note that we tend not to voluntarily loosen our standards; rather, it tends to happen unconsciously and spontaneously. Johnson and her colleagues distinguish between 'heuristic' checking of candidate memories, which is usually operating automatically when we are remembering, and 'systematic' checking, which is intentional. Heuristic processing consists of fewer component processes and uses readily available information, such as familiarity, perceptual detail, and schemas (e.g., world knowledge, stereotypes), typically activated by a cue. Systematic processing is made up of more component processes and may also involve the retrieval of other memories and knowledge not initially activated.

Systematic processing requires selective attention: the person must explicitly attend to the candidate memory. It also includes self-provided memory cues. We often cue our own memories: When I want to remember someone's name, I imagine her face. I produce a cue for my memory system to use in retrieving the name. I then monitor any representations that the cue gives rise to. I may need to use other information to reject candidate names that come up. Often this cuing process must be used several times in order to reconstruct the memory correctly. As to the neural locus of these monitoring processes, researchers point to bifrontal areas.

The Suppression of Irrelevant Memories

Armin Schnider's research group hypothesizes similarly that the problem in memory confabulation is that the orbitofrontal cortex and its limbic connections are not performing their function of suppressing or inhibiting recalled memories that are not relevant to the current task. They argue that the posterior medial orbitofrontal cortex sorts out the mental associations that pertain to ongoing reality by suppressing memory traces that have no current relevance. As to the question of what the crucial lesion is for producing confabulation, Schnider points to an orbitofrontal-mediodorsal-amygdala circuit, claiming that spontaneous confabulation appears to emanate from interruption of the loop connecting the posterior orbitofrontal cortex directly (via ventral amygdalafugal pathways) and indirectly (via the mediodorsal thalamus) with the amygdala. Schnider's way of connecting confabulation in Korsakoff's with that found in ACoA patients is to point out that the basal forebrain lesions one sees in ACoA

aneurysms often include the posterior medial orbitofrontal cortex. Schnider's localization is supported by two of his findings. First, patients with lesions involving the posterior medial orbitofrontal cortex and basal forebrain confabulate for much longer periods (several months) than patients with anterior medial orbitofrontal lesions. Schnider's group also found posterior medial orbitofrontal cortex activation in normal subjects who performed a memory task requiring that they carefully separate relevant from similar but irrelevant memories.

Some frontal theories of confabulation can be expanded to fit other confabulation syndromes not involving memory problems, but others are so strongly memory based that they cannot be expanded. Theories that attribute confabulation specifically to a memory retrieval problem only apply to memory patients: Korsakoff, for instance, argued that confabulations are simply veridical memories placed out of context. Schnider et al. conclude similarly that spontaneous confabulations arise due to a failure to know the proper temporal place of the recalled memory. Alternatively, many of the executive theories have the potential to be expanded beyond mnemonic confabulation, such as Johnson's view.

Domain Specificity

The question of the domain specificity of confabulation becomes important in assessing and improving the different theories of confabulation. Several cases have now been reported of patients whose confabulation was restricted to autobiographical memory, or who at least confabulated mainly about autobiographical memory. But there are exceptions, for instance, one patient showed spontaneous confabulation across episodic, personal, and general semantic memory. Another patient with frontal damage who only showed confabulation in 'personal' semantic memory – memory of the important facts of one's life.

While retrieval theories are supported by confabulation restricted to certain memory domains, executive or reality monitoring theories have trouble with domain-specific confabulation, because the general understanding of executive processes is that they can be applied in more than one domain. Executive processes may have their own domain specificities, however. If there are several different checking or monitoring processes, this predicts that there are several different types of confabulating patients, depending on which processes are damaged. This also implies that several different neurological tests need to be used to assess each of the types. Dalla Barba and his colleagues developed a confabulation battery that contains questions touching on personal semantic memory (such as one's address, age, or occupation), questions relating to autobiographical memory, questions about orientation in time and place, and questions touching on general semantic memory, such as historical facts. There are also several extremely difficult questions in the battery (e.g., "Who was Marilyn Monroe's father?"), to which normal subjects would be likely to answer "I don't know," on the assumption that confabulators will fail to know that they do not know.

Separating Amnesia and Confabulation

A large body of evidence has accumulated showing dissociations between problems with temporal memory and confabulation. The two phenomena also show different time courses: The confabulation in Korsakoff's often clears while the amnesia remains. Similarly, confabulation clears in most ACoA patients after weeks or months, but the memory problems do not improve. Sometimes it works in the other direction, memory improves, while confabulation continues. Two phenomena militate against the idea that amnesia alone in Korsakoff's patients can explain the confabulation: First, confabulation tends to disappear as patients progress from the confusional stage to the chronic phase, while the amnesia remains. Second, some chronic amnesics patients confabulate a great deal, others less or not at all; and there is no correlation between their propensity to confabulate and the gravity of their amnesia. If memory confabulation results from two independent lesions, this indicates that there are two types of patients:

1. Patients who sustained the memory system lesion first. This is a patient who should admit

his memory problem until the executive problem develops, at which point he should deny it and commence confabulating.

2. Patients who sustained the executive lesion first. The course of this disease might be rather subtle. We also need to leave open the possibility that there are some people who simply do not develop the executive processes needed to check memory reports, and make do with their memories alone and tolerate a high rate of errors. This opens up the interesting possibility that the problem with some Korsakoff 's patients is that they are confabulatory before losing their memory. They have lost the ability to check thoughts or candidate memories. It may pass unnoticed because the patient is substantially correct in what he says. But once the amnesia sets in, the problem becomes painfully obvious.

Episodic Memories of Others

A person's episodic memories are memories of that person from her point of view. The representations that are stored in episodic memory are egocentric in the sense that they represent events as we experienced them, hence they are also called autobiographical memories. Autobiographical episodic memories combine several different subrepresentations including representations of our bodies moving through different spaces and environments and representations of the people and objects we have significant interactions with. We carefully represent each aspect of a particularly significant interaction, exactly what was said, and in which tone of voice. In addition to representations of emotions, episodic memories may also contain representations of other conscious states, such as our thoughts, motives, and intentions at the time of the event. The typical autobiographical memory representation is of a person with a conscious mind, moving through space, and interacting with people and objects.

The episodic memory system is also able to aggregate its information into our existing concepts of important people and things. Once this information enters the system of concepts it becomes part of the semantic memory system also, and is then accessible to the process of thinking itself. Thus the episodic memory system can feed the semantic memory system. For instance, if I travel to Paris, the episodic memories I amass as I see important sites in the city also add information to my semantic representation of Paris, the Eiffel Tower, the Arch of Triumph, etc. All of this information tends to be either conceptual or allocentric in form, but I suggest that the egocentric realm has its own ability to aggregate its representations into a full-blown simulation of persons' minds. As I accumulate information about someone over the course of many interactions with her, I also accumulate information about her thoughts, moods, and emotions, using the egocentric representation system in other mode, to represent her. We might call such memories, 'biographical memories.' These accumulations of simulated mental states constitute representations of the minds of the significant people in our lives. If this is right, our representation of a significant person would contain an allocentric component with representation of how he looks and sounds, and an egocentric representation of his mind, body, and environments from his viewpoint. Our awareness of the minds of others when the egocentric system is representing them is nonexplicit and faint in our minds. Perhaps this is a reason why distinguishing whether an emotion or a simulation of a mind is missing is difficult for the patients. Perhaps another reason why the activity level of the egocentric system needs to be low when it is operating in the other mode is because full-blown conscious representation leads to actual external actions, not merely to represented actions, in much the same way that the dreaming mind malfunctions during REM sleep disorder, when dreamed perceptions cause real actions. Simulations cannot fully employ this representation system without danger of causing real actions, with potentially disastrous consequences.

The experimental and clinical data show that this biographical memory system may have a special fragility in the way it sorts incoming memories according to persons' identities. The system needs to be able to create new concepts when new people are encountered, and to properly segregate this concept from its concepts of other people. For instance, if I start a new job and spend several hours with a new coworker, my brain will begin

creating a representation of this person. But when this person leaves and another new coworker enters, my brain needs to close the first file it opened and start another one. There needs to be a firm partition between these two representations, so that I do not confuse the identities of my coworkers.

Conclusion

The many types of different memory systems that the brain's evolutionary development has created testifies to the value these forces of design place on learning from the past. Several questions remain: Do memory confabulations belong to the larger set of completion phenomena, such as the filling-in of the visual blind spot? The executive processes located in the prefrontal lobes require clear, unambiguous information in order to achieve their primary task, that is, the creation of effective actions. We typically do not have the time to spend examining gaps in our perceptions and memories. Very often in real life, when memories occur, we make a quick plausibility check, sort out any obvious contradictions or impossibilities in the memory, and move forward with the belief that the memory is correct. Another interesting question is whether memory confabulations might involve a type of self-deception. One piece of evidence for this is the way that so many of the recalled memories present the patient in a rosy light. In addition, the idea that merely imagining or thinking about an event can cause us to later confabulate that the event happened may be revealing one of the mechanisms behind self-deception to us: thinking that something is so, usually something positive for us, can work to convince us that it really is so. On the one hand this removes some of the blame for self-deception from us, since it shows that unconscious processes not normally under our control can produce a plausibility effect. On the other, it shows the power that conscious rehearsal of information can have on our rather fragile and complex memory systems.

See also: The Control of Mnemonic Awareness; Autobiographical Memory and Consciousness; Consciousness and Memory in Amnesia.

Suggested Readings

DeLuca J (2001) A cognitive neuroscience perspective on confabulation. Neuro-Psychoanalysis 2: 119–132.

DeLuca J and Diamond BJ (1995) Aneurysm of the anterior communicating artery: A review of the neuroanatomical and neuropsychological sequelae. Journal of Clinical and Experimental Neuropsychology 17: 100–121.

Fischer RS, Alexander MP, D'Esposito MD, and Otto R (1995) Neuropsychological and neuroanatomical correlates of confabulation. Journal of Clinical and Experimental Neuropsychology 17: 20–28.

Fuster JM (1995) Memory in the Cerebral Cortex. Cambridge, MA: MIT Press.

Garry M, Manning CG, Loftus EF, and Sherman SJ (1996) Imagination inflation: Imagining a childhood event inflates confidence that it occurred. Psychonomic Bulletin and Review 12: 359–366.

Hirstein W (2005) Brain Fiction: Self-Deception and the Riddle of Confabulation. Cambridge, MA: MIT Press.

Johnson MK, Hayes SM, D'Esposito MD, and Raye CL (2000) Confabulation. In: Grafman J and Boller F (eds.) Handbook of Neuropsychology. New York: Elsevier.

Kopelman MD (1987) Two types of confabulation. Journal of Neurology, Neurosurgery, and Psychiatry 50: 1482–1487.

Korsakoff SS (1889) Psychic disturbance in conjunction with peripheral neuritis, Victor M and Yakovlev PI (trans.). Neurology 5(1955): 394–406.

Loftus EF (1991) Made in memory: Distortions in recollection after misleading information. In: Bower GH (ed.) The Psychology of Learning and Motivation: Advances in Research and Theory. San Diego: Academic Press.

Moscovitch M (1995) Confabulation. In: Schacter DL (ed.) Memory Distortions. Cambridge, MA: Harvard University Press.

Parkin AJ and Leng NRC (1993) Neuropsychology of the Amnesic Syndrome. Hove: Lawrence Erlbaum Associates.

Schacter DL, Reiman E, Curran T, Yun LS, Bandy D, McDermott KB, and Roediger HL (1996) Neuroanatomical correlates of veridical and illusory recognition memory: Evidence from positron emission tomography. Neuron 17: 267–274.

Schnider A (2001) Spontaneous confabulation, reality monitoring, and the limbic system: A review. Brain Research Reviews 36: 150–160.

Scoville WB and Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery, and Psychiatry 20: 11–21.

Stuss DT, Alexander MP, Lieberman A, and Levine H (1978) An extraordinary form of confabulation. Neurology 28: 1166–1172.

Talland GA (1965) Deranged Memory. New York: Academic Press.

Biographical Sketch

William Hirstein is professor and chair of the Philosophy Department at Elmhurst College, in Elmhurst, Illinois, USA. He received his PhD from the University of California, Davis, in 1994. His graduate and postdoctoral studies were conducted under the supervision of John Searle, V. S. Ramachandran, and Patricia Churchland. He is the author of several books, including *On the Churchlands* (Wadsworth, 2004), and *Brain Fiction: Self-Deception and the Riddle of Confabulation* (MIT, 2005).

Memory: Procedural Memory, Skill, Perceptual-Motor Learning, and Awareness

R Custers and H Veling, Utrecht University, Utrecht, The Netherlands

Glossary

Closed-loop process – Process that uses information about its outcomes as input.
Declarative memory – Memory for knowledge about facts.
Episodic memory – Autobiographical memory, part of declarative memory.
Open-loop process – Process that does not use information about its outcomes as input.
Perceptual-motor learning – Learning of motor skills that rely on perceptual input.
Procedural memory – Memory for knowledge about how actions are executed.
Schema – General knowledge structure in memory about a particular type of action.
Semantic memory – Memory for meaning, part of declarative memory.
Skill – Overlearned behavioral routine resulting from practice.

Introduction

Anybody who has ever learned to ride a bicycle, or play the piano, will admit that mastering such skills requires lots of practice. And when asked how we acquire those skills, that is about all we have to say about it. We may be able to describe the principles by which a bike functions, or the structure of chords, but unlike such declarative knowledge, the knowledge about how our actions are executed – that is, procedural knowledge – that we acquire during skill development is not open to introspection. This seems even more remarkable if one considers the challenges that are faced in such learning processes. We never make the exact same bike ride twice, or play the same arpeggio on the piano, yet somehow we are able to extract general knowledge about the motor actions involved, store it in memory, learn to plan and tune those actions based on perceived feedback, and learn how they can be combined into complex sequences of actions without much awareness of what is going on.

Perceptual-Motor Learning

The problems with storing an action in memory already become apparent if we look at the simple act of, say, grasping a coffee mug. In any instance, we have never executed the exact required action before (the location of the mug and its handle are different) and even more problematic, we can execute the action in an infinite number of ways. One can flex and extend the different joints that are orchestrated by many different muscles in an infinite number of ways with the same result. We are, for example, still able to grasp the mug when holding a phone between our ear and the shoulder of the grasping arm and get the job done in an awkward, but effective, way. This degrees-of-freedom problem suggests that knowledge about actions cannot be stored in terms of exact knowledge of muscle tension and joint positions, but has to be represented in a more general way.

It turns out that what is stored in memory is the general pattern or schema that captures the essential structure of the action, with specific parameters to be filled in during the planning and execution of the action. This requires integration of knowledge from memory and information resulting from perception. The crucial role of visual feedback in the execution of actions has, for example, been demonstrated in studies in which perception was blocked by turning off the lights once participants had planned and started to execute the movement. Except for extremely quick, reflex-like actions – which are more ballistic in nature – performance significantly deteriorated without visual feedback. This demonstrates that even performance of a simple action is

not an open-loop process, in which the action is only planned and then blindly executed, but a closed-loop process that requires integration of perceptual as well as motor information.

When it comes to executing skilled behaviors, it is especially this tuning of action, based on closed-loop processes, that we are mostly unaware of. We may consciously initiate a turn with our bike, but the processes by which we maintain our balance and stabilize our course in the new direction operate largely outside of awareness. This lack of consciously accessible knowledge about tuning becomes apparent in the following phenomenon. When you ask participants in the laboratory to demonstrate the steering movements that they usually make to change lanes with their car, most people turn the wheel in the required direction, and then turn it back to the starting position. In reality, it takes an equal turn in the opposite direction after changing the lane to keep the car on the road. Apparently, although people are consciously aware of the movement they must make to initiate action, they lack conscious knowledge of the subsequent compensating movement that relies on tuning in response to visual feedback.

Sequence Learning

In addition to this lack of conscious knowledge of tuning, people are often oblivious to how sequences of actions are learned. Researchers have since long been intrigued by the question of how rapid sequences of actions in playing piano, or typing, can be learned and executed, especially because there is no time to receive the feedback of one action before performing the next. Although it was first believed that such sequences were stored in memory as action chains, in which each specific action that was retrieved from memory would trigger retrieval and execution of the next associated action, by now it has become clear that people do not so much store the chain of the exact responses, but rather a more abstract, higher-level pattern. This was revealed by studies showing that execution of a specific action-pattern benefited from an earlier learning phase, even when the exact actions during that learning phase (pressing a series of buttons with one's hand) were executed by a different means in the test phase (e.g., pressing the same series of buttons with one's elbow).

Learning of such action sequences is one of the pillars of skill development and has been shown to occur even without conscious awareness. In classic studies on implicit sequence learning, participants responded with different actions to different stimuli that occurred in a fixed, but rather complex pattern. Even though participants did not consciously detect any pattern, they speeded up over time and performance significantly deteriorated when the stimuli were no longer presented according to the fixed pattern.

Organization of Skills in Memory

But even when such sequences are learned there is room for improvement in execution. One of the characteristics of skills is that they keep improving, although ever less dramatically, over time. An important mechanism that contributes to this effect may be more efficient storage of action patterns in memory. Typing lessons capitalize on this knowledge by having people repeat often occurring patterns of keys (e.g., 'T-H-E') over and over again. Although it is clear to the student what has to be learned, the pattern of movements has to be 'stamped in' through practice. This strategy promotes the grouping of specific actions in memory into chunks, which also happens over time without specific practice. These chunks can then be retrieved from memory and executed as one unit.

Brain Processes, Skill-Learning, and Awareness

Studying skill acquisition in people with specific brain lesions, as well as modern neuroimaging methods such as positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have made it possible to gain knowledge about the areas of the brain that are involved during several stages of skill development. One striking finding is that brain structures that are crucial in memory for events have little to do with skill acquisition. Studies with patients suffering from anterograde amnesia have demonstrated that although

those patients are unable to form new memories about events, they develop many skills just as well as control participants. It has been demonstrated that this skill-learning goes even further than perceptual-motor skills. In one study, amnesic and control participants practiced reading of words that were presented in mirror image in a series of learning blocks that extended over a 2-week period. Some specific words were repeated from block to block, whereas the rest of the words were always novel in each block. Although the reading time per word for the repeated words dropped faster for control participants than amnesiacs, mirror-reading skills for novel words were shown to improve equally fast for both groups. Hence, although control participants clearly benefited from the memory of having seen the specific repeated words (which they could consciously recall) in previous blocks, amnesiacs developed the same mirror-reading skill without any recollection of those words.

These findings demonstrate that procedural and declarative memories are supported by distinct brain areas. In declarative memory, the hippocampus has been found to play a crucial role. Bilateral damage to the hippocampus and parahippocampal regions indeed causes retrograde amnesia, with loss of episodic memory and semantic memory. These structures also play a crucial role in certain types of learning, such as trace conditioning, where there is a relatively long time lag between a neutral conditioned stimulus (CS) (e.g., a tone) that predicts an unconditioned stimulus (US) (e.g., a puff of air in the eye). Although after conditioning in such a task animals with an intact hippocampus react with an eye blink to the CS, such learning does not occur when the hippocampus is damaged. The same effect has been demonstrated in humans, where this type of learning is accompanied by activation of the hippocampus and usually with conscious awareness of the relation between the CS and the US. In sum, the hippocampus plays a crucial role in declarative memory, but also in particular forms of learning, where integration of separate events in memory is required.

The brain areas involved in procedural learning, and in particular the learning of motor skills, change as learning of the skill progresses. In the early stage of motor-skill learning, the prefrontal cortex (PFC), anterior cingulated cortex (ACC), and posterior parietal cortex (PPC) direct attentional and control processes that are necessary for grasping the basics of the task and the planning of deliberate, intentional responses. As skills start to form, the cerebellum plays a crucial role in creating 'mental shortcuts' between perception and actions. It integrates afferent signals coming in from the sensory systems with efferent signals that produce motor actions. As such, it links together new and already acquired motor-programs, forming more complex response patterns and assigning them to particular perceptual patterns. Activation of the cerebellum is typically found in the early stages of skill-learning and shifts from the cerebellar cortex to the nuclei as learning advances, until it almost ceases when a basic skill is acquired.

In this stage, the striatum has been found to play an important role in the detection of errors that are produced. When erroneous responses occur, the inappropriate programs are inhibited, a process in which the PFC is again thought to play an important role. As a result, connections between perceptual input and the appropriate motor output are further strengthened. Thus, the skill is further polished until it is executed almost flawlessly.

If this skill is then used frequently over a longer period, the skill becomes overlearned, which will finally lead to changes in the motor areas that play a role in the execution of the specific motor actions. Such cortical plasticity has been demonstrated in, among others, musicians and professional sportsmen, in whom the mapping of crucial motor programs has been reorganized in service of their skills. As such, overlearned skills become more and more hardwired in people's brains, and conscious attention or awareness are no longer required to conduct them.

Acquired Skills, Goals, and Conscious Control

Although the execution of a skill may eventually occur without conscious awareness, skills are often executed in the service of a conscious goal. Grasping a mug and bringing it to one's mouth, for instance, usually serves to take a sip of the fluid contained by the mug (e.g., coffee). Because acquired

skills are stored as abstract high-level patterns, skill execution such as grasping the mug can be elicited by merely perceiving a skill-relevant cue (e.g., a coffee mug) under many different circumstances (e.g., under variations in distance toward the mug, or size of the mug) once a goal is set. As a result of overlearning, little or no conscious control is needed to execute the individual action sequences that capture the essential structure of the action (e.g., when and how much a hand should be opened when reaching for the mug). In fact, it has been shown that focusing conscious attention on the execution of specific components of a complex motor skill can impair performance.

For instance, experienced golfers are better in putting a ball when they are distracted by a secondary task (e.g., monitoring whether a specific tone sounds through a headphone) than when they are instructed to focus attention on their swing. In a similar vein, experienced soccer players, but not novices, handle the ball better with their dominant foot when they are distracted from executing a skill (e.g., dribbling) than when they consciously focus on specific components of the actions that they are executing. An explanation for this effect is that by attending to separate components one overrules the more efficient organizational structure of the skill, causing the building blocks of the skill to function as separate components, in pretty much the same way as before the skill was acquired. Once the organizational structure breaks down, each component is executed separately, which costs more time, and leaves more room for error, than when the abstract high-level pattern is used to execute the separate components. It has been proposed that conscious attention to the step-by-step components of motor skills can lead to choking under pressure, for instance when a soccer player misses a penalty kick. Conscious control can, however, be beneficial if skills are not yet overlearned. It can, for example, improve performance in experienced soccer players when they handle the ball with their nondominant foot, or in novices in general, indicating that conscious attention toward action components can be helpful in practicing, or tuning, an underdeveloped skill.

Although conscious control can be detrimental to performance of complex goal-directed skills, conscious control may be experienced over the overall action if one reflects on one's behaviors, as long as the outcome of the action (e.g., the golf ball rolls in the hole) is congruent with the goal that started the execution of the skill (e.g., putting the ball in the hole). Thus, even though awareness – and conscious control – of specific components of overlearned skills can be very restricted, conscious control can nevertheless be experienced when an intended outcome is reached. It appears that what is needed for an experience of control is that the outcome of an action is congruent with the mental representation of an intended outcome, irrespective of whether one actually consciously controls each and every step of the action sequence.

The lack of conscious control over our skills, however, becomes apparent if our skills produce other outcomes than we consciously intended. As noted above, skill execution can be easily elicited by perceiving a skill-relevant cue when one has a goal in mind. If you want to use your computer, for instance, perceiving the log-in screen may trigger the responses required to log-in (e.g., typing in the correct password) without much conscious deliberation, if this procedure has been performed a great number of times before. Although this efficiency has a great advantage, as it relieves consciousness from tedious tasks such as retrieving a password from declarative memory, it also comes with a cost: It can be hard to prevent skill execution when a goal requires a new response that is different from an overlearned behavioral routine. For instance, when the password of your computer is changed, it is possible that the old password is entered erroneously the first couple of times that you start the computer. This example illustrates that an overlearned behavior (typing in the old password), which is no longer instrumental in attaining the goal (starting your computer), is still easily triggered by the environment. When there is a discrepancy between the outcomes produced by our skills and our intended outcome (starting the computer), we become aware of the situation, and conscious control can be recruited to choose a new path of action in order to reach the intended outcome (e.g., consciously retrieving the new password from declarative memory). Of course, with repeated practice, conscious control will again become less necessary and perceiving the log-in screen will be sufficient to trigger the adjusted skill.

Conclusion

From perceptual-motor to sequence learning, procedural knowledge is acquired as skills develop, although most of the time this knowledge is not accessible to consciousness. Acquiring a skill takes a lot of practice, during which different brain processes are involved in creating stable mental maps of the skill. Although it may seem to be a disadvantage in that there is no other way of acquiring this knowledge than through practice, the major benefit of procedural memory over declarative memory (which relies on different brain areas) is that it is highly stable. Once one has learned to ride a bike one never forgets. Although sometimes these skills may produce different outcomes than consciously intended, on the whole they serve us well, by freeing up precious capacity for other conscious processes. This process is so effective that conscious attention to skills may even deteriorate performance. Thus, skills make it possible to behave very effectively without much conscious interventions. In this light, the lack of introspection in how we acquire them may be a small price to pay.

Acknowledgment

The preparation and writing of this article was financially supported by the Netherlands Organization for Scientific Research (VENI grant 451-06-014, VICI grant 453-06-002).

See also: Habit, Action, and Consciousness; Perception, Action, and Consciousness.

Suggested Readings

Anderson JR, Bothell D, Byrne MD, Douglass S, Lebiere C, and Qin Y (2004) An integrated theory of the mind. Psychological Review 111: 1036.

Cooper R and Shallice T (2000) Contention scheduling and the control of routine activities. Cognitive Neuropsychology 17: 297–338.

Jeannerod M (1997) The Cognitive Neuroscience of Action. Oxford, England: Blackwell.

Lewicki P, Hill T, and Czyzewska M (1992) Nonconscious acquisition of information. American Psychologist 47: 796.

Powers WT (1973) Feedback: Beyond behaviorism. Science 179: 351–356.

Squire LR, Knowlton B, and Musen G (1993) The structure and organization of memory. Annual Review of Psychology 44: 453–495.

Biographical Sketch

Ruud Custers received an education in human–technology interaction at Eindhoven University of Technology, where he graduated cum laude on work investigating the role of memory in the formation of judgments about environments. Subsequently, he moved to Utrecht University to pursue his PhD in experimental social psychology. He received his PhD cum laude in 2006 for his dissertation on the underlying mechanisms on nonconscious goal pursuit, which mainly focused on the role of affective signals in this process. He published several papers in fundamental journals, of which one was regarded as the best paper of the year on social cognition by the International Social Cognition Network in 2006. As an assistant professor at Utrecht University, he continues to study the processes that allow people to pursue goals without conscious awareness.

Harm Veling is trained as an experimental social psychologist at Radboud University Nijmegen where he worked on intention memory and inhibitory control. He received his PhD in 2007 for his dissertation on the inhibitory processes that facilitate execution of previously formed intentions. Next, he continued to work as a postdoc at Utrecht University. His work deals with several topics related to the role of goals and actions in automatic processes of social cognition, with an emphasis on inhibitory processes, and is published in several fundamental journals. One recent discovery in his research concerns the notion that stopping an action that is initially triggered by a rewarding stimulus results in devaluation of the rewarding stimulus. This chain of events suggests a functional behavior regulatory dynamic. He continues to study this intriguing and important topic.

Mental Representation and Consciousness

D D Hutto, University of Hertfordshire, Hatfield, UK

Glossary

Enactivism – The view that experiences are not inner events but are constituted by the activity of organisms engaging or interacting with some environment or other.

Mode of presentation – The manner or aspect under which an individual, object or state of affairs is apprehended by a thinker or subject. Modes of presentation are thought to be the basis of cognitive mediation.

Narrow content – A kind of content that is nonrelational and dependent only on the local, portable properties of psychological individuals.

Phenomenal character – The distinctive quality associated with token experiences (perceptions, sensations, feelings, moods); what-it-is-like to undergo such experiences.

Phenomenal consciousness – Subjective states of mind that involve having experiences with specific phenomenal characters.

Strong representationalism – The claim that the phenomenal character of experience is exhausted by, identical with, or entirely determined by representational content or properties.

Subjectivity – The property of being an experience for something or of someone, often equated with the having of an idiosyncratic first-personal point of view, or perspective.

Supervenience – A nonreductive relation of covariance holding between specified relata (concepts, properties, entities, etc.). Supervenience relations are generally believed to involve asymmetric dependence or determination. Accordingly, if 'x' and 'y' properties exist, such that 'y' depends on or is determined by 'x' then any relevant change in 'x' also incurs a change in 'y.' For example, aesthetic properties are thought by some to depend on physical properties, such that modifying an object's material aspects in relevant ways would automatically modify its aesthetic aspects, that is, enhancing or marring its beauty. Supervenience relations can vary in their modal force (as reflected by operators such as 'necessity' and 'possibility') and scope.

Weak representationalism – The claim that experiential facts are, or supervene on, representational facts.

Wide (or broad) content – Content that is necessarily individuated by environmental factors, or kinds that exist outside the bounds of subjects or organisms.

Introduction

Intentionality and consciousness are the fundamental kinds of mental phenomena. Although they are widely regarded as being entirely distinct some philosophers conjecture that they are intimately related. Prominently it has been claimed that consciousness can be best understood in terms of representational facts or properties. Representationalist theories vary in strength. At their core they seek to establish that subjective, phenomenal consciousness (of the kind that involves the having of first-personal points of view or perspectives on the world – perspectives that incorporate experiences with specific phenomenal characters) is either exhausted by, or supervenes on, capacities for mental representation. These proposals face several serious objections.

Intentionality and Mental Representation

Intentionality and consciousness are the most fundamental and philosophically interesting kinds of

mental phenomena. Explaining how they fit into the world order poses powerful (and some think insuperable) challenges for naturalists who rely exclusively on the resources of hard sciences for prosecuting that task.

Intentionality is the capacity to have thoughts, feelings, and other states of mind that are about, or directed at, particular worldly offerings and states of affairs. Examples include having the thought that "It is going to rain"; harboring a desire for strong coffee, or seeing that the leaves have changed color. Building on the foundations laid down by early modern thinkers, some contemporary philosophers hold that the capacity to be in such states of mind is best explained by the having of inner mental representations with specific truth-apt contents. At the very least, it is accepted that mental representations form a subclass of intentional phenomena.

Mental representations are distinct from ordinary representations. The latter are familiar public items of our everyday acquaintance. They include such things as drawings, maps, and natural language statements. Mental representations are modeled in weaker or stronger ways on these pedestrian entities and are thought to have similar properties and functions. They aim to say how things stand with the world. If they speak truly they can inform their users of the state of the world, successfully guiding their behavior and actions (or at least they can do this when they combine appropriately with conative states).

Minimally, something is a mental representation if it presents some portion of the world as being a certain way; for example as 'being hot' or 'being colored.' The world may of course vary from the way it is presented as being: things are not always as they seem. Importantly, mental representations can be entirely nonconceptual and still be regarded as having content. A mental state will have content and be representational in character if it has correctness conditions that are specifiable, in principle. Conceptual and nonconceptual content are thus thought to be equally representational in nature; they are distinguished only in the way that they represent.

Representing aspects of the world necessarily invokes correctness conditions. Although some philosophers have argued that certain types of representations are immune to error – such as the necessary truths of logic or mathematics – the existence and status of such truths, and whether they take the form of genuine propositions, is a matter of controversy. It is generally accepted that naturally occurring representations that say how things stand with a changing world – for example, those of the perceptual sort – always entail a risk of error or misrepresentation. Explaining how this could be so is taken to be a primary condition of adequacy for naturalistic theories of mental content.

Acknowledging this, it is almost universally agreed that a stronger condition on what it is to be a representation, is warranted. Something is a representation if and only if it presents some portion of the world as being in a certain truth-evaluable way. As such, specific representations target aspects of the world, saying how things might be.

Propositions expressed by well-formed sentences of natural language remain the clearest and least controversial models of contentful representations. Propositions must in some sense correspond to, or otherwise incorporate, the states of affairs that they represent. What this involves and what propositions are comes out differently on different analyses. For example, Russellian propositions are purely extensional; they include objects or individuals that are thought of as constituents of propositions themselves. Such propositions have the property of being true or false, but they do not do so in virtue of corresponding to anything beyond themselves.

In contrast Fregeans hold that propositions have both intensional and extensional aspects. This idea is embodied in Frege's famous distinction between sense and reference. Intensional content, or Fregean sense, can be understood as the way or manner in which the referent in question is presented to, or apprehended by, the subject. Famous examples include the different senses of the proper names Hesperus and Phosphorus. Although these names denote the same planet, Venus, they have different connotations; one labels it 'the morning star' and the other 'the evening star.' As such they have the potential to evoke different thoughts in their respondents. For Frege all genuinely referring names and whole propositions were thought

to have intensional contents, which had to be understood in aspectual terms. Following Frege such contents and their analogues in contemporary theories of representation are commonly known as modes of presentation.

Whereas Frege rejected standard correspondence theories of truth, most contemporary accounts do not. Many subscribe to some version or other of possible world semantics. They assume that the intentional content of any given representation – what it is about in extension – is some possible state of affairs. Such representations can also be understood in terms of specific representational targets – what is meant to be represented. Individual representations are thus made true by the obtaining of specific facts.

Another important distinction is that of vehicles and contents. In case of language-based representations, linguistic signs serve as vehicles of meaning. They are meaningless, locatable spatiotemporal existents, even if the contents they sponsor are not. Nevertheless they are what enable discrete acts of mental representation; they make representation possible. Drawing directly on the analogy of sentences and their meanings, neural states or processes – or something identifiable with such – are imagined to be the basis of cognitive processing involving mental representation.

There is much dispute about the exact form and properties of the vehicles of mental representation. Some, following empiricists, think of them as sensory and imagistic in character. Others endorse a more linguistically inspired conception, holding that the basic units of cognition are amodal symbols; elements of a 'language of thought.'

Psychosemantic Theories of Content

The semantic properties of mental representations, from which – some argue – all other more public representations derive their meaningful content, require special explanation. Psychosemantics is the attempt to produce a workable theory of content.

In the prehistory of cognitive science only two sorts of naturalistic theories were advanced to explain representation: resemblance and causal theories. Both proved inadequate. Contemporary thinkers agree that the vehicles of content need not resemble what they are meant to denote in order to represent. The contentful properties of mental representations do not depend or rely upon there being any kind of resemblance between what is represented and how it is represented. There need be no inner pictures in the head that match or copy outer scenes; nor would the existence of such exotica be sufficient to explain representational relations. Two coins from the same national mint will resemble one another (almost exactly), but they do not thereby represent one another. Being similar in such respects is not enough for representation.

Causal theories are no better placed to provide a naturalistic account of representation. The crudest versions appeal to causal laws. These can be understood as trading on conditional statements of the following kind: if, in the right conditions, subject S perceives an object (or feature) of type X, then S's mental machinery will produce a token of type 'X' (*ceteris paribus*). Theories of this kind suffer from a multitude of problems, but the most devastating is the misrepresentation problem. It is demonstrable that any causal account, which makes an appeal to strict causal laws, rules out the possibility of misrepresentation, and thereby the possibility of representation. The problem is that the tokening of symbols would be caused by many other things than those exclusively in the class of things to be represented. Yet if, *ex hypothesi*, it is only causal relations (as defined by strict laws) that determine how specific symbol types are meaningful (if it is only causal connections that fix their representational content, reference, and truth-conditions) – then such symbols must represent or stand for the entire class of things that would cause them. This rules out any possibility of misrepresentation. But since the possibility of misrepresentation is taken as an essential requirement for representing 'tout court,' it turns out that the imagined symbols represent nothing at all. The fatal consequence of crude causal theories of content is that they leave no room for error.

All contemporary naturalistic theories of representational content acknowledge the need to account for misrepresentation. They recognize the need to go beyond purely causal-informational accounts if they are to explain mental representation.

Some seek to account for the norms that ground content in counterfactual terms, by appealing to the asymmetric dependence of false tokenings on true ones. This is one way of rising to the challenge. But it leaves open questions as to why one class of things is propriety in this regard, and what establishes and maintains this. Any interesting psychosemantic theory must answer such questions without making an appeal to phenomena that are already infused with, or rely on, the existence of mental representations.

Teleosemantic accounts have proved popular for this reason. Such theories look best placed to circumvent the problem of misrepresentation, while providing a substantive naturalistic theory of representational content. To achieve this they make appeal to the notion of organismic proper (or teleo) functions – those that are defined in terms of the ends they serve for organisms. These functions should not be confused with purely systemic functions of the sort that are defined by the role an item or device plays within a more complex system. Rather teleofunctions concern what a device or item is supposed to do as opposed to what it is disposed to do.

In explicating teleofunctions the standard naturalist strategy has been to appeal to a principled notion of biological function. The aim is to account for the basis and source of the norms in question in a scientifically respectable manner. There are a number of ways of doing this, but by far the most popular has been to explain the basis of the norms etiologically – that is, by appeal to the historical conditions and evolutionary pressures under which the devices were formed and forged. Construed thus, proper functions are explained in terms of normal conditions of operation that tell how a function was historically performed on those (perhaps rare) occasions, when it was properly performed. The historical conditions in question are those in which a given biological response originally conferred the sort of benefits that brought about, or contributed to, the selection of its underlying mechanism. Appeal to historically normal conditions, therefore, enables the explanation of why a device, entity, or response proliferated, and this, in turn, enables us to say what it is supposed to be doing, even in those cases in which it fails to achieve its ends.

Unlike other representational theories of content, which tend to focus solely on the input, teleofunctional accounts seek to clarify the notion of representation both in terms of indicative relations between the representation and the represented, as well as, by crucially emphasizing the 'use' that is made of purported inner representations to achieve specific organismic ends or purposes. Such approaches are consumer-based.

Phenomenal Consciousness

Consciousness is an umbrella term. Philosophers recognize many varieties associated with diverse criteria. Which forms are genuine and which are basic is a matter of intensive and ongoing debate. Philosophical taxonomies are meant either to descriptively capture core features of our pretheoretical understanding of consciousness, or to identify its true characteristics, as revealed through analysis. Still, there is no clear consensus on what the concept picks out. The situation reflects, and is exacerbated by the fact that we speak of consciousness in many different ways in ordinary parlance. Nevertheless it is widely agreed that consciousness has some prominent and interesting features that must be either explained or explained away.

We say that a creature or organism is conscious if it is awake and sentient. This minimally implies that it has some degree of occurrent awareness, it does not entail that we can describe the character of such awareness. Such consciousness may be awareness of its surroundings or aspects thereof or it may take a more intransitive form. Either way this is generally thought to involve being in a state of mind with a characteristic feel – one in which there is something-that-it-is-like to be in it. So understood, consciousness is an all or nothing property: one either has it or one does not. Human beings, cats, octopi (apparently), and spiders (perhaps) are kinds of things commonly thought capable of possessing it while inanimate objects, such as chairs, are not.

Consciousness takes specific forms. What-it-is-like to be a human being varies considerably from what-it-is-like to be a dolphin, or more famously still, what-it-is-like to be a bat. Different types of conscious beings enjoy experiences with

phenomenally or qualitatively different characters. Moreover, particular types of experiences have distinctive characteristics. Experiencing itchiness is quite different from experiencing anger. Seeing the peculiar greenness of an *aloe vera* plant differs from seeing the greenness of a Granny Smith apple.

Apparently experiencing makes a difference. Encountering the unusual taste and smell of durian, for example, may evoke reveries or prompt actions. Yet experiences are only sometimes implicated in the guidance of behavior and action and its autonomous or rational control. It is easy to think of examples of complex activity involving sophisticated, but apparently habitual, automatic, or unreflective responses, which are not governed by agents, precisely because those agents are not conscious of what they are doing in ways that would make it possible for them to modify their behavior. For this reason some reserve the accolade of being conscious only for those beings that exhibit a certain degree of control over their actions or those that are capable of reporting or expressing how things appear to them. Here it seems unavoidable that to have such control a creature must be aware of specific features of its environment in more than a general and intransitive sense. In addition, some hold that awareness of this kind implicates at least some degree of self-awareness.

In thinking about the phenomenal character of experience, analytic philosophers often claim that token mental states of specific kinds have distinctive qualitative properties of the sort just described. It is important not to confuse the claim that experiences have phenomenal characters with the claim that these are best understood as causally efficacious mental particulars or objects (often identified with subjectively accessible feels or *qualia*). Equally, acknowledging the existence of phenomenal characters does not entail that these are ineffable or logically private. Moreover, although it is common for phenomenal properties to be identified using abstract categories (such as 'greenness') it is likely that their character is too fine-grained (or analog) to be adequately identified in this way. While sentient beings can selectively attend to and experience certain features of their environment, it may be that they are, nonetheless, only capable of experiencing the world in ways that are highly context-specific and circumstantial. Minimally, to say that a creature or being has an experience of a certain phenomenal character is to acknowledge that there is something-that-it-is-like for it to engage in certain activities. It is to observe that some creatures experience the world in characteristic ways.

It is subjective, phenomenal consciousness, of the kind that equates to having a first-personal point of view, or perspective on the world, involving experiences with specific phenomenal characters, that is the primary concern of those who propose that consciousness can be understood in terms of representational facts or properties.

Representational Theories of Consciousness

The idea that representation and consciousness are intimately connected made its most prominent debut with the advent of Cartesian philosophy in the seventeenth century. Descartes promoted an understanding of minds as being a special sort of mental substance. To have a mind is to have a coherent and unified individual perspective on reality. These unique points of view are imagined to be internally complex. It is possible to notice and attend to specific worldly features, such as the greenness of a particular apple, but this involves being able to see an apple as something more than just the sum of its presented features. To see an apple as something in which greenness, and other properties, might inhere is to see it as having a continued existence over time. To experience a world of objects and their features always occurs against a larger and more complex background in which such items are systematically related to other things. To have experience of the world, as opposed to merely having sentient capacities, is to experience it as structured.

Although Descartes recognized the existence of perceptions, emotions, and sensations, to the extent that these were not under the control of the will – a decisive mark of the involvement of contentful thoughts – he relegated them to purely bodily phenomena, admitting of physiological explanation. For him these were not to be confused

with interesting, bona fide cases of mental states. Only humans enjoy the latter, in his view. This is because we (humans, at least) do not simply get by in the world by responsive and reactive means using a battery of sensory modalities and mechanisms – being moved simply by colors, sounds, tastes or the like – rather we encounter and are conscious of the world and its aspects as being a certain way. It contains a range of properties and things, arranged thus and so. Or, at least, that is how it seems to us. How the world is, in actuality, may be quite different.

The driving intuition behind this Cartesian insight is that all genuine conscious experiences are contentful; they necessarily involve having ideas – the ultimate basis for conceptual judgments. Plausibly, this is the central characteristic of the full blown, perceptual consciousness enjoyed by humans (though it remains an open question whether any other species have similar capacities). In promoting this idea Descartes is credited with having initiated the first cognitive revolution. Following in his footsteps, many of today's philosophers and cognitive scientists also hold that the true phenomenal consciousness must have contentful features.

Contemporary representational theories of consciousness endorse the basic Cartesian picture. The most ambitious versions hold that conscious experience simply equates to taking the world to be a certain way. Accordingly, what-it-is-like to be conscious boils down to having a coherent model of how things might be; one that logically and determinately excludes others. Thus, if some aspect of the world appears to be red then that same region of reality cannot also appear to be green, at the same time. This understanding of conscious experience is representational, because the way one takes things to be, can of course be false. The most ambitious representationalists not only claim that phenomenal consciousness is essentially representational in this respect; they also claim that the phenomenal character of experiences is nothing other than, or can be fully explained in terms of, the contents of tokened representations.

In recent times this alleged link between phenomenal consciousness and mental representation has been less evident. Since the birth of cognitive science, until quite recently, there has been a tendency to adopt an isolationist policy regarding these two aspects of mind. They have generally been treated as entirely separate topics of study. The tendency has been to focus on theories concerning unconscious mental states and subpersonal representations, and to ignore or avoid the question of consciousness altogether. In such a climate, demonstrating that representation and consciousness are essentially linked is a substantive project.

Existing representational theories of consciousness come in a variety of forms and strengths. Some versions claim that representational facts exhaust all the experiential facts; that all experiential facts are representational facts. This is a relatively weak claim. Its proponents only commit to the view that phenomenal consciousness supervenes upon representational facts. The former need not reduce to the latter. Being in a state of mind with a representational content might be essential to being conscious, without it being the case that being conscious is nothing but representing. Taking the world to be a certain way might be one necessary factor amongst others (such as a mental state's having particular functional properties). The phenomenal character of a given experience might be determined in complex ways. Pitched at this level of grain, representationalism can tolerate impurities in its account of what makes a creature (or its mental states) phenomenally conscious.

This sort of proposal is typically cast in terms of covariance relations holding across possible worlds. The basic idea, which admits of considerable modal refinement, is that there is no possible world such that all the representational facts remain the same, yet exhibit a difference with respect to the phenomenal consciousness. For such a law to hold, it need not be the case that conscious properties *per se* reduce to representational properties, all that need be true is that facts about representation and consciousness travel together, always and everywhere. It may be true that there cannot be phenomenal character without representation, without it being the case that the two are more intimately related.

A stronger thesis is that there can be no difference in phenomenal character without a corresponding difference in representational content; either because consciousness is just a kind of representational

content or because it lawfully covaries with changes in such content. Those who believe that the phenomenal character of conscious mental states is exhausted, or entirely determined, by the representational properties of such states support such views.

Whether representationalists advance fact-based or property-based variants, insofar as they only advance supervenience claims, they leave open a range of possibilities about what might best explain the metaphysical basis of the proposed systematic relationships. This might be explained in terms of causal dependencies, mereological part–whole relations, or token identities.

The scope of such proposals is restricted in that even strong representationalists only propose that all experiential properties are representational and not vice versa. The claim is that all conscious experiences are, or are logically related to, representations, not that all representations are, or logically related to, conscious experiences. This is an important qualification because it seems clearly false that every representation does its work by invoking phenomenal properties. For example, sincerely asserting that ''Snow is white'' says that the world is such that snow is white (relative to a time, place, and language). The English sentence itself may have familiar phenomenal properties – indeed ones that we can attend to – but it does not represent the putative fact that ''Snow is white'' by presenting things to the speaker or hearer in a way that relies on it appearing that snow is white. The word 'white' does not denote whiteness by looking white or exhibiting the quality of whiteness.

The English sentence, cited above, is a truth-evaluable representation, but it is not of the mental sort. Still the point carries. For the same verdict ought to apply to items in the mental lexicon and the sentences composed from them (should any exist). For example, if we imagine that modes of presentation of some mental representations are entirely syntactic and linguaform – if they take the form of well-formed strings of amodal symbols – then such representations will clearly share the characteristic of natural language sentences of representing what they do by nonexperiential means.

This does not show that no mental representations could have the relevant properties. But it does show that at best representationalism will be true only of mental representations of a select kind. Representing the world as being a certain way is not sufficient for being an experience with a particular phenomenal character. Not every representational fact entails an experiential fact. For example, some believe that even the low-level activity in early stages of human visual processing involves the manipulation of contentful representations. But, if so, the content of such representations will be quite distinct from the contents of ordinary perceptual experience. The cells that fire in low-level visual processing are orientation-invariant, whereas, our experiential way of perceiving objects is not. This implies that, at best, only a subclass of representations exhibit or involve phenomenality.

It also follows that genuine phenomenal consciousness might be the province of only a small class of organisms; those whose online perceptual processing requires rich, complex, and multilevel forms of representation. It has been argued, for example, that only the abstract representations allegedly involved in the intermediate stages of human visual processing would be of the right kind to possess experiential content. If so, perhaps as Descartes thought, only humans are truly conscious. Importantly, those who defend such views also hold that contentful experiences are logically distinct from (and typically ontologically prior to) perceptual beliefs and judgments. Seeing that the wall is of a certain illumination, with various hues and a certain distance away involves enjoying a nonlinguistic, nonconceptual, representational content. It does not involve making a belief-based judgment.

As a result only a special class of representations will be suitable candidates for rendering true the supervenience claims of representationalists. Yet some proponents of this view draw further distinctions even within this category, restricting their claims to the specific types of perceptual experience found in particular modalities. Accordingly differences in the phenomenal character track differences in content within the visual, auditory, or tactical modalities.

There is an obvious motivation for wanting some suitably strong version of representationalism to be true. If it could be convincingly demonstrated that the phenomenal features of consciousness are really nothing but the representational features of mental states, then this promises an attractive

metaphysical economy. Two seemingly intractable problems in the philosophy of mind would in fact turn out to be a single problem: that of accounting for representational content in naturalistic terms. While naturalizing content is generally acknowledged to be difficult to achieve, it is also thought to be less challenging than providing a straight solution to the hard problem of consciousness. If strong representationalism were true then this would allow for a considerable reduction of effort: the naturalization of representational content, in a way consistent with physicalism or materialism, would be the only hard problem.

Representationalism has some intuitive appeal. It trades on the insight that conscious experience is transparent or diaphanous. To the extent that experience has any qualitative aspects, we only become aware of these by focusing on features of what a given experience is about (or is apparently about). We typically see right through our experiences to the objects of our concern or interest. In most cases it is the character of what our experiences are about that is the focus of our attention and not the experiences themselves. We are conscious of the trees blowing in the wind or of the mountain in the distance, first and foremost. Only on rare occasions, if at all, is it possible to focus our attention on the properties of the way we experience things. And, arguably, even in those cases the content of what we introspect is borrowed from the basic content of the experience itself. Those who defend a strong transparency thesis maintain that even in cases in which we concentrate on features of our experiences, we have no way of specifying their phenomenal character, other than by describing the features of what the experiences are about. Despite its initial plausibility, the truth of the strong transparency thesis has been challenged of late and softer variants promoted. Notably, although the strong transparency thesis helps to motivate representationalism, the program could in principle survive its loss.

General Objections to Representationalism

A standard objection to representationalism is that it fails as a general account of conscious experience, because it fails to encompass experience of pain, feelings, or bodily sensations (such as tickles, itches, bouts of nausea) or undirected, diffuse moods such as depression and elation (those which have no precise focus or target). Its defenders have tried to show that such experiences can be accommodated if we think of them as telling us something about the state of the organism in question – that is, as indicating something about damaged body parts, internal or external, or about the state of the creature more generally. In effect, representationalists hope to demonstrate that all experience is ultimately perceptual in nature. For if it should turn out that there are nonperceptual experiences with distinctive phenomenal characters, it would undermine the general truth of their thesis.

Others have directly attacked the supervenience claims of strong representationalists. A number of thought experiments have been fashioned in order to put these under intuitive pressure. The hope has been to demonstrate that it is logically possible that representational content and phenomenal character can come apart. One way of achieving this asks us to imagine that there could be a case of difference in phenomenal character without any difference in representational content. Thus we are asked to imagine two different subjects that are functionally equivalent in the performance of a given recognition-based task, say, the sorting of red socks from blue socks. It is assumed that in order to perform this task they must represent aspects of the world as being a certain way. But it is then further stipulated that one group of sock-sorters see red as blue, while the others see red as red. But this makes no difference to either group's ability to perform the tasks. If such cases are possible, and not just imaginable, then the qualitative aspects of experiences might vary arbitrarily, relative to their representational contents.

Another tactic is to construct cases in which there is a difference in representational content while the phenomenal character remains the same. Ned Block's Inverted Earth attempts this. The colors of his imagined world are systematically reversed with respect to ours, such that there green things are red, blue things are yellow, and so on. The language spoken by Inverted Earthlings

is also inverted; as a result, it perfectly mirrors that of Earthlings. Just like us they are inclined to voice that 'Grass is green' and 'Roses are red.' Bearing this in mind, we are then asked to imagine that after having undergone color-inversion operations to their visual systems some Earthlings are transported to Inverted Earth – all without their knowledge. On their arrival to the character of those they experiences would be imperceivably different to those they would have on Earth in similar circumstances, even though they would be everywhere misrepresenting the colors of this new environment. This would constitute a difference in representational content without a relevant difference in phenomenal character.

If such imagined scenarios are to serve as genuine counterexamples, it must be assumed that our capacity to imagine them is metaphysically revealing and relevant. A standard response to the use of such thought experiments is simply to deny this. Naturalists in particular generally challenge the idea that such imagined cases have any philosophical bite. A stronger offensive tactic is therefore to identify internal tensions that render specific representationalist proposals incoherent or unworkable.

In proposing a general metaphysical thesis about the nature of phenomenal consciousness, strong representationalism need not commit to any specific theory about the nature of content, such as externalism or internalism, nor need it commit to any particular psychosemantic theory of content, such as biosemantics or causal-informational theories. But evaluating the truth of any interesting version of representationalism ultimately requires assessing specific proposals in the light of such commitments.

Objections to Externalist Variants

Some strong representationalists hold that when it comes to understanding phenomenal character, we should concern ourselves only with what it is that an organism is having an experience of. What is of interest is what the experience is and how what it is about is represented. Representationalists of this stripe are interested in intentional content – what it is that one is conscious of – and not the way in which one is conscious of it.

Such theorists endorse externalism about phenomenal character, if they also accept that intentional content cannot be individuated by internal factors alone; a view that has some support today. Thus if phenomenal character is nothing but (or strongly covaries with) intentional content and intentional content is wide or broad (in that it cannot be individuated solely by appeal to the internal properties of an organism) then phenomenal character will also share this feature.

The standard externalist versions of representationalism hold that to represent is essentially a matter of tracking, or having the function to indicate specific worldly features. But which worldly features do experiences represent? These are identified with external, mind-independent, if subject-relative, 'affordance-like' properties – properties of the world that reliably (or reliably enough) elicit certain dispositions in specific organisms. Surface qualities or properties are what are represented.

Advocates of this sort of view confront the problem of making good on objectivism about the so-called secondary properties. Consider the case of colors. Most philosophers are wary of the idea that there are any worldly features with which the colors we experience can be uniquely identified. It is generally acknowledged that our experience of color depends both on the character of our internal processes as well as our responses to specific external stimuli. To begin with, the colors and shades we see are in part determined by the ratio of activation across three light-sensitive cones in our retinas. These ratios of activation do not correspond in a one-to-one fashion to particular light stimulus configurations, but rather with a disjunction of the latter. Moreover, the distributed response of these cones initially effects changes in the retinal ganglion, and from there further processing occurs in two separate chromatic channels, known as the red-green and the blue-yellow. The interaction between these two channels underpins our chromatic experience in a way that makes the classification of hue perception a fourfold business. The problem for advocates of color externalism is that, if what constitutes our color experience constitutively depends on internal factors of this kind, then it looks as if there is not much prospect of identifying colors with any exclusively,

mind-independent, external features of the world (such as particular wavelengths of light or spectral reflectance signatures or the like).

If there are no interesting, mind-independent features of the world which can be cleanly identified with the various colors we experience then we must ask: Which properties of the environment are we tracking or representing in having color experiences? Unless there is a principled answer to this question externalism about experience is under threat. The problem generalizes, since a similar verdict looks apt to apply to other response-eliciting surface properties as well.

A more serious problem facing those strong representationalists who endorse externalism concerns misrepresentation. Let us suppose that 'phenomenal green' is a mind-independent, response-eliciting property. We know that it can be represented truly in a variety of ways. I can represent its presence correctly even though my way of doing so does not depend on my representing its surface properties as being green *per se* – as when I state that the fourth book on the shelf behind me has a green spine (doing so from memory and without in any way imagining that color). But equally I might also be inclined to deny the presence of 'phenomenal greenness' even when I am, as a matter of fact, confronted with it. I might see something that is actually 'phenomenal green,' but which looks red to me: my mode of presentation may have a 'red' character. Perhaps this will be because the lighting conditions have been carefully adjusted so as to generate precisely this type of illusion. This would be a case of misrepresentation since I would be inclined to act as if (and would judge that) the portion of the world in question was in fact 'red' when it was really 'green.' This sort of situation is clearly possible (and not just imaginable like the Inverted Earth scenario). If so, it looks as if the relevant supervenience claim is false if one assumes that phenomenal character supervenes on wide intentional content. It seems possible to have experiences with phenomenal characters different from the intentional content of what is represented.

In thinking about such cases it is important to realize that to represent successfully essentially requires that a correspondence holds between what is represented and what does the representing. Following the linguistically inspired model of representation, specifying the content of truth-conditional representation involves disquotation. Thus it is no accident that representational contents are defined by making direct reference to the worldly features or objects that representations aim to represent. But in some contexts this can be misleading, since to specify representational content in this way requires focusing squarely and exclusively on cases of successful representation. This raises important questions about how to specify the contents of representations in cases of misrepresentation; those in which there is a mismatch or, more precisely, a lack of correspondence between the state of the organism and the world.

It is accepted that the way organisms respond to the range of nonveridical stimuli in instances of misrepresentation must be understood as strongly dependent upon, or as having been in some way fixed by the way they respond to a select propriety class of referents. It is because they respond appropriately when encountering tokens of those types that their representations are veridical. The precise analysis of this dependency relationship, and the prospects for explaining it naturalistically, vary according to different psychosemantic theories, as noted above.

Nonetheless, whichever theory one favors, problems arise for those who want to defend the idea that phenomenal character equates to wide or broad content. For the phenomenal character of experience, and what is represented, appear not to always and everywhere keep in step. If anything, it looks as if phenomenal character is best identified with the nonintentional properties of certain modes of presentations. This might be understood in terms of representational character, as opposed to representational content. Phenomenal character should be understood as the way or manner in which things make their appearances rather than being identified with what is represented in particular cases (even if we assume that the latter places unavoidable and interesting constraints on the former). To recognize this one must acknowledge that acts of conscious experiencing with phenomenal characters have a dual structure, involving features of which one is aware and features in virtue of which such awareness is made possible.

Objections to Internalist Variants

Variants of strong representationalism that seek to reduce phenomenal character to (or identify it with) external, intentional content face serious problems. But externalism is not the only option for strong representationalists. Some have proposed that there may be an identity between phenomenal character and narrow content.

Narrow content is a special sort of content recognized by internalists. The notion was first introduced in order to show how everyday psychology (which apparently only trades in wide content attributions) could nevertheless be at peace with methodological solipsism. Methodological solipsists hold that it is what agents have in mind (not external features of the world) that matter to what they do and for explanations of what they do. Recognition of the existence of narrow content was meant to show that it is possible, at some level of abstraction, to class behaviors as being psychologically identical, even when the propositional attitudes of the agents initiating those behaviors had different wide contents. This was meant to capture what is psychologically similar about physically identical duplicates when they are each operating in environments in which the wide contents of their thoughts vary, such as when one is on Earth and the other on Twin Earth. Twin Earth is a philosophical construct, a world exactly like this one in every detail, but in which the seas, lakes, and baths are filled with the substance twater. Twater has all the outward properties of water but it has the chemical make-up of XYZ (not H_2O). Consequently, despite the fact that the thoughts of the two imagined agents are about different things, their behavior will be exactly the same. Internalists hope to explain this by proposing that what they have in common is the narrow content of their thoughts. Narrow content is a property of internal states; it is defined in terms of partial functions that map thoughts onto environments. This is the sort of content that the imagined twins can share even though the truth values of their thoughts may differ.

It is not possible to give a positive, free-standing specification of narrow content. It is normally defined indirectly as the type of content that is independent of the way in which subjects are related to their environment. It is emphatically not a kind of truth conditional content. On the other hand, it is not reducible to a mere vehicle of content, understood in terms of its syntactical or functional role. Narrow contents are at best truth-apt in that they become truth-evaluable when and only when they are appropriately anchored to an environment. Hence the best way to characterize narrow content is to think of it as a partial function that needs to be contextualized. Ultimately its specification depends on making an appeal to the wide intentional content. Narrow content is thus inherently and radically inexpressible.

This fact raises the concern that narrow contents (should they exist at all) are misnamed; perhaps they are not really any kind of content. For if narrow contents are best understood as functions (mathematically conceived) from contexts to truth conditions then they are not strictly speaking about anything. At best, they have a special kind of potential to be so. To equate phenomenal character with narrow content would entail that phenomenal character is at best a kind of nonveridical content. This avoids the problems associated with misrepresentation, but only because there is no question of narrow contents being true or false. Such a move would seem to suit certain widely held Cartesian intuitions. For unreformed Cartesians would want to identify phenomenal character with how things seem to us. And how things seem to us is not something about which we can be mistaken. I can be mistaken that the tomato before me is green, but not that it looks green to me. My phenomenal awareness (how it seems) can be at odds with how the world is. If the phenomenal character of experience is just a kind of narrow content then it would be a kind of content for which the question of misrepresentation does not arise. If there is any coherent type of representational content that has such a feature then it is narrow content.

But narrow contents are not generally assumed to be consciously accessible and if the notion proves to be unstable then, once again, phenomenal character will be identified with the nonrepresentational properties of representational states (at best). For example if the notion of narrow content does not stand up to critical scrutiny then phenomenal

character might be identified with noncontentful representational vehicles. But clearly any such straightforward identification on its own will not adequately explain why or how we sometimes represent the world in a phenomenologically salient way. To constitute a satisfying explanation of this kind the vehicles in question would have to have some rather special additional properties, but such properties may, at best, be logically related with certain kinds of representational acts without being representational or contentful properties *per se*.

Objections Concerning Subjectivity

Another worry for representationalists is that more is needed to understand phenomenal consciousness than an account of phenomenal character in any case. Even if there were a workable representationalist story about the latter it would not suffice to explain the former. This is obviously so if, for example, we allow that it is possible to be in a state of mind with a specific phenomenal character without in any way being aware (or even being capable of being aware) of being in such a psychological condition. It is imaginable that certain creatures might undergo pain (in the sense that there is something-that-it-is-like for them to have painful experiences) even though they would be congenitally unable to access these qualitative feels for themselves. If so they would feel pain without being aware of the painful experiences. This seems more than merely theoretically possible. There is some phenomenological evidence that we sometimes undergo experiences with characteristic phenomenal feels, such as being in pain or anger, without noticing, attending or being aware of these at the time of their occurrence. If so, we can distinguish experiences of which organisms are unaware from those that are, to invoke some well-worn metaphors, brought before the mind's eye or illuminated by the spotlight of attention.

Some will object that the very idea of unexperienced experiences is incoherent. They will claim that to have an experience logically entails the existence of a subject (of some sort) – one that actively undergoes or is aware of the experience in question. Experiencing necessarily involves bringing the subject in on the act. This is because experiencing is necessarily experiencing-for something or someone. Whether unexperienced experiences are possible or not, it is clear that experiences are ordinarily experienced. If so a proper account of everyday phenomenal consciousness requires something more than a mere account of phenomenal character; in addition it requires an account of the subjective nature of experiencing.

A number of recent attempts have been made to account for the subjectivity of consciousness in representational terms. These come in the form of higher-order theories of representation of either perception-based or thought-based varieties. The core idea behind all such theories is that first-order mental states are in some sense rendered conscious when they become the objects of other mental states. This is thought to involve a kind of second-order self-monitoring that utilizes either meta-representations or re-representations.

A root problem with such higher theories of representation is that they threaten to make the first-order phenomenal character of experiences superfluous in fixing or determining the what-it-is-likeness of phenomenal experiences. This is because if the subjective aspect of experiencing is determined by second-order representational properties then there is always a risk of misrepresentation. Thus it is possible that what-it-is-like to experience, say, the color green will be fixed by higher order representational properties even in those cases in which the phenomenal character of the lower-order mental state is misrepresented. These will include situations in which one is, in fact, having an experience with a different phenomenal character (say, of red or blue) and those in which one's mental states lack phenomenal character altogether. In such circumstances things might still appear to be green because the higher-order monitoring is doing all of the decisive work in determining how things seem to the subject.

The problem emerges in duplex accounts of this kind because there is always the possibility that the low-order and higher-order states might be out of sync. This is because representation is an extrinsic, external relation and not a constitutive one. This difficulty can be overcome or avoided by supposing that our experiences are in some way

intrinsically subjective. An intimate, nonpositional kind of awareness is quite unlike adopting a representational stance toward one's mental states. Experiences might be Janus-facing – if so, they would not only exhibit world-direct intentionality, but their character is given to the experiential subject in some minimal, prereflective way. The idea that experience necessarily has the feature of givenness is a familiar one in the work of the great phenomenological thinkers and it can be found, arguably, even in Descartes and Kant.

The closest representationlists have come to accommodating this has been to endorse the view that phenomenal consciousness involves the having of self-representing representations. But to make sense of this proposal requires positing the existence of representations of a unique kind, those with quite special (if not naturalistically impossible) properties. Accounting for such representations would be difficult, at best, within the compass of existing theories of psychosemantics.

Nonrepresentational Accounts

Strong first-order representationalist proposals (of both the externalist and internalist varieties) look to be too strong. The way things seem or feel to an organism phenomenally does not appear to equate to any kind of representational content, narrow or wide.

Weaker, impure versions of representationalism – those that cast their identity claims at the level of facts and not properties – are more plausible. Proponents of such accounts concede that not only do select representations have phenomenal features, but also that such features may be best understood in terms of nonrepresentational properties. As such weak representationalism allows that the phenomenal features that give experiences their special characters are not explained by their being a kind of representational content, but, at best, by their role in performing representationally related service.

This claim is weak enough to be consistent with accounts of the nature of experiences that regard them as external, in a way quite different to the proposals of strong representationalists. Emphasizing their ecological and interactive nature, enactivists claim that experiences are not inner events of any sort; rather they are constituted by the activity of organisms engaging with some environment or other. Subjective experiences that have specific phenomenal characters are identified with temporally extended worldly interactions. Brain activity may be necessary, but it is not sufficient for experience. The causal substrate that fixes the character of any token experience is temporally and spatially extended to include aspects of the brains, bodies, and environments of organisms.

Ontologically speaking experiences just are extended spatiotemporal happenings – the reactions, actions, and interactions of embodied creatures situated in the world. As such they are a subclass of events; they are roughly dateable and locatable happenings. While such mental events necessarily involve active, living organisms, they do not occur wholly inside them. Experiences, understood as events or token occurrences, can also be characterized as physical. They are instances of lived, embodied activity, which can be designated and characterized using vocabulary that makes no use of phenomenal predicates. For this reason experiences can be the objects of scientific theorizing and experimentation.

All of this is consistent with weak representationalism, but the more radical versions of enactivism go further, claiming that organisms act first and develop the capacity for thought later. Accordingly they hold that it is not necessary for creatures to think about or represent in order to act successfully in the world. If this is true it reduces the scope of even weak representationalism, since phenomenal consciousness would be more basic and fundamental than the capacity to represent the world as being a certain truth-evaluable way. Experiencing aspects of the world might be thoroughly noncontentful; aspects of the world might not be presented as being in a certain contentful way to subjects at all, even though there would be something-it-is-like to engage with worldly offerings in phenomenologically salient ways.

See also: Inner Speech and Consciousness; Meta-Awareness; Visual Imagery and Consciousness.

Suggested Readings

Block N (1990) Inverted earth. Philosophical Perspectives 4: 53–79.
Brook A (ed.) (2007) The Prehistory of Cognitive Science. Basingstoke: Palgrave.
Byrne A (2001) Intentionalism defended. Philosophical Review 110: 199–240.
Crane T (2001) Elements of Mind. Oxford: Oxford University Press.
Dretske F (1995) Naturalizing the Mind. Cambridge, MA: MIT Press.
Fodor JA (1987) Psychosemantics. Cambridge, MA: MIT Press.
Gallagher S and Zahavi D (2007) The Phenomenological Mind. London: Routledge.
Kind A (2003) What's so transparent about transparency? Philosophical Studies 115: 225–244.
Kriegel U (2002) Phenomenal content. Erkenntnis 57: 175–198.
Lycan W (1996) Consciousness and Experience. Cambridge, MA: MIT Press.
Neander K (1998) The division of phenomenal labor: A problem for representationalist theories of consciousness. Philosophical Perspectives 12: 411–434.
Noë A (2004) Action in Perception. Cambridge, MA: MIT Press.
Rowlands M (2001) The Nature of Consciousness. Cambridge, MA: Cambridge University Press.
Seager W (2000) Theories of Consciousness. London: Routledge.
Tye M (1996) Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind. Cambridge, MA: MIT Press.
Zahavi D (2005) Subjectivity and Selfhood: Investigating the First-Person Perspective. Cambridge, MA: MIT Press.

Biographical Sketch

Daniel D. Hutto was born and schooled in New York, but finished his undergraduate degree as a study abroad student at St. Andrews, Scotland. His maternal roots are Scottish and many of his family still live in Inverness. He returned to New York to teach fourth grade in the Bronx for a year in order to fund his MPhil in logic and metaphysics, after which he carried on his doctoral work in York. He now lives in Hertfordshire with his wife and three boys, having joined the local university in 1993. He served as the head of philosophy from 1999 to 2005, and is currently the research leader for philosophy. He has published on wide range of philosophical topics in journals including: *The Monist*; *Proceedings of the Aristotelian Society*; *Philosophy and Phenomenological Research* and *Mind and Language*. He is the author of *The Presence of Mind* (1999), *Beyond Physicalism* (2000), *Wittgenstein and the End of Philosophy* (2006), and *Folk Psychological Narratives* (2008). He is also the editor of *Narrative and Understanding Persons* (2007) and coeditor of *Current Issues in Idealism* (1996) and *Folk Psychology Re-Assessed* (2007). A special yearbook issue of consciousness and emotion, entitled *Radical Enactivism*, which focuses on his philosophy of intentionality, phenomenology, and narrative, was published in 2006.

Meta-Awareness

J M Chin, University of British Columbia, Vancouver, BC, Canada
J W Schooler, University of California Santa Barbara, Santa Barbara, CA, USA

Glossary

Automaticity – The cognitive processing associated with highly practiced activities.
Flow – A mental state marked by full immersion in a task, such as the feeling of being 'in the zone.'
Meta-awareness – The process of directing attention toward the contents of consciousness, thereby gaining an appraisal of the contents of consciousness.
Metacognition – Knowledge about one's knowledge.
Metaconsciousness – Another term for meta-awareness.
Mindfulness – A state that is enhanced by a practice of meditation, in which one engages in watchful attention and is very present in the moment.
Mind wandering – The experience of having thoughts that are unrelated and often counterproductive to the task and hand.
Recovered memories – Memories, corresponding to what would seem to be highly memorable experiences such as trauma or sexual abuse, that are perceived to have been forgotten and subsequently recalled.
Temporal dissociation – A temporary disconnect between consciousness and meta-awareness in which one goes for some period of time without noticing the contents of their thought.
Translation dissociation – A disconnect between consciousness and meta-awareness, in which meta-awareness, in reappraising consciousness has misrepresented the original experience.
Tuning out – Mind wandering that an individual is aware of engaging in.
Verbal overshadowing – The phenomenon of verbal report interfering with a related task. For instance, describing a face and subsequently having poorer recognition for that face.
Zoning out – Mind wandering without awareness.

Introduction

'Meta-awareness,' a term often used interchangeably with metaconsciousness, is the state of deliberatively attending to the contents of conscious experience. Frequently when researchers speak of consciousness, they distinguish between two general states: unconscious, in which if information is processed at all, it is processed without any concomitant experience, and conscious, in which individuals experience what ever is occupying their minds. In this context, meta-awareness can be thought of as third level of consciousness in which consciousness is turned upon itself in order to re-represent the contents of experience. In other words, meta-awareness is one's explicit appraisal of the current contents of consciousness.

The distinction between unconscious, conscious, and metaconscious processes can be illustrated with the example of mind wandering while reading. Consider the all too familiar case of reading along and then suddenly realizing that despite your best intentions, although your eyes have continued to move across the page, your mind was fundamentally elsewhere. In this example, the pattern recognition procedures that allowed you decode the individual words correspond to an unconscious processes. Conscious experience would primarily correspond to specific contents of mind wandering musings, although

presumably some superficially processed aspects of the reading (e.g., the sounds of some of the words) would also enter consciousness. Initially in this example, meta-awareness would be absent until you notice that you are mind wandering. This abrupt realization (almost like waking up) represents the dawning of metaconsciousness, in which you take stock of what you are thinking about and realize that it has nothing to do with what you are reading. Once meta-awareness of mind wandering is achieved, then you are able to redirect your attention to the narrative and can search for the last sentence you remember actually attending to, and try again.

Several important features of meta-awareness emerge from the mind wandering while reading example. First, as conscious and unconscious processes are presumed to carry on relatively continuously throughout waking life, metaconsciousness is intermittent. Only periodically does one take stock of the contents of thought, which is how it is possible for one to mind wander during reading despite knowing that it is impossible to comprehend text while simultaneously thinking about something entirely unrelated to what one is reading. Second, when one notices that they are mind wandering while reading, there is often an abrupt experience of 'coming to.' It is almost as if one has just awoken even though one was not previously asleep. The abruptness of metaconscious understanding illustrates the qualitative difference between a mental state that is associated with metaconsciousness and one for which metaconsciousness is lacking. Finally, it is important to emphasize that when one notices that the mind has wandered, this fact temporarily becomes the dominant element in consciousness, as one thinks to themselves ''shoot I just spaced out again, let's see: where was I?'' Thus, although metaconsciousness represents a distinct category of thought, it may be nothing more than a particular kind of subject matter of consciousness, namely when the subject of consciousness becomes an acknowledgment of whatever it is that consciousness had just been focusing on. In short, metaconsciousness need not represent a unique mental state, nor need it entail specific dedicated brain regions. It may simply be a topic area (albeit a very important one) of consciousness.

History

Meta-awareness can be thought of as a subtype of a larger category of mental phenomenon known as 'metacognition.' Metacognition corresponds to our knowledge about what we know. The term was introduced by John Flavell in the context of studying children's capacities for monitoring their cognitions. Flavell distinguished two general classes of metacognition. The first was metacognitive knowledge, or in other words understanding how thought operates in the world. Belief that rehearsal improves memory falls under this category of metacognitive knowledge. The second class was metacognitive experiences, which was more focused on the feelings associated with metacognition. The feeling that a sought for word is on the 'tip of one's tongue' falls under this heading.

A key moment in the development of the construct of meta-awareness came with Thomas Nelson and Louis Narens' introduction of a basic model of the control and regulation of cognition in a paper published in 1990. This early model proposed two levels of cognition, the object-level and the meta-level. The object-level deals with cognitions about external objects, while the meta-level deals with cognitions about object-level cognitions. Moreover, information was said to flow between these two levels. Monitoring occurs when the object-level is informed by information from the meta-level. Control, on the other hand, occurs when information from the meta-level modifies the object-level. These applications of metacognitive knowledge provided a shift from work focusing on an understanding of conscious thought, toward work on meta-awareness as a specific representation of consciousness. The distinction between monitoring and control remains prevalent in current work on meta-awareness.

Meta-Awareness and Monitoring

Many aspects of daily life require routine monitoring and adjustment. When one speaks, one needs to adjust their volume to the ambient noise in the room. When one reads, one needs to adjust the rate at which they move their eyes in accordance with comprehension. In many cases, this

type of monitoring and adjustment can carry on behind the scenes without any explicit awareness. However, the tacit monitoring system is limited in the types of things it can monitor. It can, for example, help one to keep one's volume appropriate under normal circumstances, but if one is wearing headphones one has to explicitly attend to one's volume when speaking, otherwise one will be shouting. Similarly, tacit monitoring processes can control eye movements sufficient to recognize the word, but recognizing that one is not attending to what the words are saying requires a more sophisticated monitoring process. Meta-awareness provides this more sophisticated form of monitoring. Comparisons can be drawn between the meta-aware monitoring system and the pilot of an airplane. The autopilot system (tacit monitoring) can efficiently pilot a plane under most circumstances, making minor adjustments to keep the plane on track under most conditions. However, if something major occurs, the pilot (meta-awareness) is needed to make major corrections and decisions. The meta-aware monitoring system has a lot more resources available as it draws from several different systems, but is also more resource-taxing and can potentially interfere with carrying out concurrent tasks. Thus, while it is necessary to invoke meta-awareness periodically to make sure that things are on track, the common absence of meta-awareness is adaptive because it frees up resources that can be applied to the task at hand.

Dissociations between Meta-Awareness and Consciousness

Because meta-awareness is re-representation of the contents of consciousness and not consciousness itself, it is possible that meta-awareness can in some cases be an imperfect or poorly timed translation of conscious experience. In other words, people may fail to take stock of their conscious thoughts, or may do so inaccurately. These dissociations between meta-awareness and consciousness can have important and far-reaching consequences. In 2002, Schooler elaborated on the concept of meta-consciousness by proposing two specific types of dissociations between experiential consciousness and metaconsciousness: *temporal dissociations*, which are discovered experiences that once had eluded meta-awareness, and *translation dissociations*, which encompass occasions when meta-awareness does not accurately reappraise conscious experience. Both types of dissociations are explained in greater detail in the following sections.

Temporal Dissociations

Conscious experience can often occur in the absence of meta-awareness. When triggered, meta-awareness can lead to a reappraisal of elements of conscious thought that once eluded meta-awareness. This experience of discovering an experience that one was previously not meta-aware of is known as a temporal dissociation.

Mind Wandering

Mind wandering without noticing it is a quintessential example of a temporal dissociation meta-awareness. The pervasive phenomenon of mind wandering occurs when attention is decoupled from the task an individual intended it be directed toward. In many situations, mind wandering may be quite adaptive or at least, harmless. For example, when one is walking to work, it may be helpful to think about what one needs to do that day, rather then devoting all attention to the nondemanding task of walking down the sidewalk. However, in other situations, for example, when one is driving in difficult traffic or reading an important paper, mind wandering is counterproductive. The fact that individuals mind wander even when engaged in tasks that they recognize as being undermined by mind wandering illustrates how easy it is to temporarily lose track of what is going on in one's mind, that is, to have a temporal dissociation of meta-awareness.

In recent years, a number of laboratory studies have investigated the process and impact of mind wandering without meta-awareness. Under laboratory conditions, two different approaches have been used to sample mind wandering. The first approach, the probe-caught method, samples the experience of the individual at varying time intervals as they perform a cognitive task. The second

approach, the self-caught method, requires the individual to respond with a button push whenever they catch their own mind wandering. Probe- and self-caught measures of mind wandering yield different information on the occurrence and awareness of mind wandering because they systematically sample the different aspects involved in off-task experiences. The probe-caught technique provides evidence of how readily the mind turns inward, and can be used to study the onset of decoupling, or the speed of drift within attention. On the other hand, the self-caught method requires the individual to recognize that his or her mind is wandering, and so illustrates the engagement of meta-awareness of their own mind wandering. Evidence of the value of distinguishing between probe-caught and self-caught mind wandering comes from the findings that the two measures are differentially associated with task performance. Interestingly, it is the probe-caught mind-wandering episodes that tend to be maximally associated with detriments in performance, including both reading comprehension and memory. The more modest consequences of self-caught mind-wandering episodes suggest that when individuals are meta-aware of mind wandering, they are more effective in circumventing its costs either by more effectively dividing attention or by more efficiently recovering information that was missed during the lapse.

A second way in which meta-awareness during mind wandering has been assessed is simply to ask people following a probe whether or not they had previously been aware of the fact that they were mind wandering. Strikingly this simple procedure reveals consistent differences (in keeping with the self-caught probe-caught distinction introduced above). For example, mind wandering episodes that are characterized as having occurred without meta-awareness (termed 'zone-outs') are typically correlated with performance detriments, whereas mind wandering episodes that occur with meta-awareness (termed 'tune-outs') tend to be less problematic. Neurocognitive measures have revealed a similar story. For example, errors on a simple vigilance task were found to be correlated with zone-outs but not with tune-outs. Similarly, in a functional magnetic resonance imaging (fMRI) study of mind wandering, it was found that the difference in brain activation between on-task and off-task performance was markedly greater when individuals reported being off-task and unaware, relative to off-task and aware.

Automaticity

During mind wandering, it can often seem like the task at hand (e.g., reading, in some cases) has been put on autopilot. There are some tasks, however, that can more easily be put on automatic, and these tasks are also understood better under the umbrella of meta-awareness. Automatic behaviors are often considered to be unconscious, a designation that can prove problematic. For instance, driving an automobile can become automatic, especially if one is attempting to do something like hold a conversation while doing so. The person will then often find that he has arrived at his destination with little memory of the actual drive. The driving, however, was not unconscious as the driver was certainly experiencing the road at some level. Meta-awareness allows psychologists to posit that the driver was conscious of the driving, but not meta-aware of these behaviors.

Unwanted Thoughts

As noted earlier, sometimes the mind can wander even when the individual is explicitly told to reign it in. But, could individuals experience this same gap in meta-awareness when the cost is revisiting a terrible thought or memory? Psychologists, such as Daniel Wegner, have often wondered about why it is so difficult to suppress unwanted thoughts. Meta-awareness shines new light on this troubling issue. Some unwanted thought theorists have hypothesized that these unwanted thoughts lie in the unconscious (or preconscious) mind. An unconscious monitoring system is said to patrol these thoughts and purposefully avoid them, but when this system gets tired or overwhelmed, these thoughts can surface. Research supporting this theory finds that unwanted thoughts are more likely when a person is under a high cognitive load, thus occupying the monitoring system.

Meta-awareness provides another level at which to understand the prevalence of unwanted thoughts. This account suggests unwanted thoughts may simply lie in conscious thoughts, ones that

occupy people's minds but may not penetrate meta-awareness all the time. The monitor, in this account, would then be patrolling conscious thoughts looking for evidence of the unwanted. This formulation of unwanted thoughts can potentially tell researchers more about how unwanted thoughts are banished and where they go.

Recovered Memories

In a similar vein to this research on unwanted thoughts, meta-awareness also provides a useful framework for accounting for recovered memories of sexual abuse. These memories that individuals were previously unaware of, but come streaming back as if from nowhere, are difficult to understand through many traditional psychological theories. Further, there is a good deal of controversy over the truthfulness of some of these memories, an important issue given the often traumatic and sexual nature of some of these memories. Although there are good reasons to believe that recovered memories can be fictitious (particularly when they are recovered in therapy), many of these memories (at least those occurring outside of therapy) can be corroborated and thus treated as memories of actual events. Scrutiny of these corroborated recovered memories demonstrates that they are often consistent with the current conceptualization of meta-awareness.

One way in which the notion of meta-awareness can help us understand recovered memories is with respect to people's estimations that prior to the recovery, the memory had been unrecalled. The characterization of a memory as having been previously forgotten is itself a metacognitive judgment. One is making an appraisal of what one thinks one previously knew. However, if individuals often lack meta-awareness of the contents of their minds, then it is in principle possible that individuals who report recovered memories could in fact have known and thought about the experiences before, but simply failed to note this fact. Several lines of evidence are consistent with this interpretation of at least some recovered memories. First, a number of documented cases of recovered memories have involved individuals who are known to have talked about their experience during the period in which they believed themselves to have been amnesic. Second, individuals with memories that are recovered out of therapy have been found to be particularly susceptible to failures in metacognitive judgments regarding previous episodes of recollection. In other words, they tend to be poor at determining what information they have previously recalled. Third, individuals with recovered memories of sexual abuse tend to be poor at noticing when they are having unwanted thoughts.

Together these findings suggest that reports of recovered memories of abuse may at least sometimes be the consequence of a deficit in meta-awareness, in which individuals lived for a period of time occasionally recalling their abuse, but failed to explicitly notice that they had done so.

Meta-Awareness and Affect

Temporal dissociations need not only deal with lapses of meta-awareness of thoughts, but also of feelings. For example, as the old children's song goes, "If you're happy and you know it, clap your hands." This line certainly implies that it is possible to be happy, but to not have realized it yet. And indeed, the current understanding of meta-awareness suggests it is indeed possible to experience an affective state without having realized it yet.

The experience of flow illustrates the dissociation between experience and meta-awareness of pleasure. One of the most effective ways of assessing the occurrence of pleasure in everyday life is the experience sampling technique in which participants are equipped with a pocket computer that intermittently probes them regarding what they are doing and how much they are enjoying it. Using this methodology, research has found that many of most pleasurable moments occur when individuals are in a state of flow. The flow state occurs when one is deeply absorbed in a task that is both highly challenging yet also accomplishable. What is so striking about research on the flow states is the fact that it indicates that individuals' most positive experiences occur when they are not thinking about themselves, but are rather deeply absorbed in the activity itself. Indeed the flow state is so absorbing that individuals do not have the attentional resources to explicitly notice that they are happy at the time.

As it seems that experience and meta-awareness of positive and negative affect can often become dissociated, then it stands that inducing meta-awareness of affect can change the entire experience of an feeling-laden event. Recent research has shown that the induction of meta-awareness does alter the nature of such events. In one such study, subjects were instructed to continually rate their level of happiness while listening to hedonically ambiguous music. Researchers found that subjects who did rate their happiness throughout the study reported less postmusic happiness than subjects who did not continually rate their happiness throughout the study. Results from this study indicate that inducing meta-awareness of emotion can inextricably change the experience itself. Further, research from flow literature also strongly suggests that introducing meta-awareness during a flow state would interrupt the state and any positive feelings that go along with it.

Translation Dissociations

When reappraising the contents of consciousness, there is a chance that the reappraisal might not be perfectly veridical to the actual experience. These experiences are classified as translation dissociations because there has been a break down of the translation of the experience to meta-awareness. The likelihood of a translation dissociation is particularly great under three sets of circumstances. First, if the experience is essentially nonverbal and a person attempts to verbally reflect on it; then there is an increased chance that he or she will get it wrong. Second, a person could be especially motivated to misrepresent an experience. And third, a lay theory about how an experience should be could lead to a reinterpretation that is unfaithful to the actual experience.

Attempting to verbalize a nonverbal experience can lead to a misrepresentation of that experience, an effect that is often labeled as 'verbal overshadowing.' The hallmark in finding the verbal overshadowing literature deals with the verbalization of faces. Faces are known to be represented holistically in the mind, a quality that makes them difficult to verbalize completely. Numerous studies have shown that subjects instructed to verbally describe a face are poorer at recognizing that face later, as opposed to subjects who simply viewed the face without instructions to verbalize it. Verbal overshadowing is likely due to the recoding of the visual image into words, a modality that lacks the intricacies and nuances to properly describe the complex, holistic nature of a face. In this situation, the meta-aware reappraisal of an experience interferes with the task at hand. The verbal overshadowing effect is not specific to faces, and generalizes to other areas of perceptual memory.

Several studies have shown that many visual stimuli that defy words are vulnerable to verbal overshadowing effects. The detrimental effects of verbalization include stimuli such as color and shapes. Beyond visual memories, verbalization can also interfere with other perceptual modalities, such as audition and taste. One such study demonstrated that untrained wine enthusiasts had poorer memory for the wine they tasted if they attempted to describe wine-tasting experience. These wine drinkers were ostensibly well accustomed to tasting wine, thus possessing the ability to detect various nuances present in the wine. They did not, however, have the vocabulary or expertise to properly describe the wine, leading to a reappraisal that did not do a good job of describing their perceptual experience. On the other hand, trained wine writers showed an improvement of memory when verbally describing the wines, likely due to their ability to accurately and fully verbalize their wine-tasting experience. A meta-awareness of an experience can both aid and obstruct a task depending on one's ability to take stock of and reanalyze that experience.

Constructing a meta-awareness of a nonverbal experience applies not only to the domain of perceptual memory, but to other areas of nonverbal cognition. In the domain of judgment and decision making, verbally describing or rating the qualities of the available choices to be made can lead to substandard outcomes. This detrimental effect of verbalization has been reliably demonstrated in contexts where the choices to be made are affect laden, such as taste and visual appeal. For instance, researchers performed a study veiled as an inquiry into consumer judgments of strawberry jams. They asked some participants (verbalizers) to taste the jams and then list their reasons for liking or not liking the jams, as well as analyzing their reasons.

Control participants tasted the jams, but did not list or analyze their thoughts about the jams. These researchers found that participants who did not list and analyze their reasons made judgments that were more similar to that of expert jam raters (from consumer report magazines) as compared with verbalizers. Other research has shown that analyzing reasons can also promote choices that yield decreased postchoice satisfaction.

The verbal overshadowing effect and other evidence of the sometimes problematic nature of meta-awareness have been shown to be reliable phenomena in the psychological literature. Still, it is important to note that verbal reflection is often helpful. This is so when the experience is easily translated into words. Research has shown that logical problem-solving is aided by verbalization, as are situations when the verbalizer has the expertise or training to create accurate verbalizations. Recall that wine drinkers trained to accurately describe wine show a better memory for wines they verbalize. Regardless, the influence of verbal reflection on memory, judgment, and decision making illustrates the importance of a full, rich, and most importantly, accurate meta-awareness of an experience.

While sometimes an experience is difficult to describe, other times individuals may simply be motivated to misrepresent the experience to themselves. It has been shown that homophobic individuals may not want to recognize when they are aroused in response to depictions of homosexual acts. In other words, they may consciously experience arousal, but not be meta-aware of these feelings due to a strong motivation to suppress such information.

Even if the motivation to develop an accurate meta-awareness exists and verbal reflection is not a problem, other barriers exist to cause translation dissociations. One such barrier is a faulty theory about what a meta-awareness should contain. That is, people may have a faulty theory about what they should be feeling or thinking in a particular situation, which in turn colors their appraisal of their actual experience. A compelling example of this comes from people's reports of their experience of catching a ball. Most people believe that as they watch a ball, their eyes first rise and then go down following the trajectory of the ball. Indeed, this is the case when one watches someone else catch a ball. However, when people catch a ball themselves, they actually maintain the ball at precisely the same visual angle. Nevertheless, when people who just caught a ball are asked what they experienced, they rely on their theory of experience rather than on what they actually did.

The Double-Edged Nature of Meta-Awareness

The various dissociations between meta-awareness and experiential consciousness serve to show that meta-awareness can have both beneficial and detrimental effects. Meta-awareness allows humans the ability to monitor and control their thoughts, which in turn make goal-driven behavior possible. To illustrate, imagine a pilot attempting to land a plane: Without the ability to take stock of his thoughts he would be unable to stave off, or even recognize a potentially disastrous bout of mind wandering during this important process. Phenomena such as verbal overshadowing, however, show that meta-awareness can be incongruent with success at certain tasks. These times when it is perhaps better to not be meta-aware are summed up well by the old children's story of the centipede and his beautiful dance. As the story goes, there was a centipede who, using all his 100 legs, did a wonderful dance that all of the other creatures were jealous of, none more so than the tortoise. The tortoise devised a devious plan to derail the centipede, by sending him a letter asking how the dance was performed. Did he move his 28th leg before the 39th? And was that followed by the 12th and 72nd at the same time? The centipede began to think about what exactly he did and was never able to dance again. This classic story illustrates well the danger of paralysis through analysis, or in other words, the danger of becoming meta-aware (i.e., dissecting the dance) of a conscious experience (i.e., the dance itself). Still, the field of mindfulness may shed light on how to achieve analysis without the accompanying paralysis.

Research on mindfulness meditation training demonstrates a process that may focus on the benefits of meta-awareness, while avoiding its pitfalls. Like other forms of meditation, mindfulness

mediation deals with becoming aware of consciousness. Unlike other types of mediation, it does not involve focusing on a stimulus, but instead strives for a broad observation of several facets of experience. Mindfulness-based interventions have been shown to help decrease stress, anxiety, and lead to several positive outcomes. These results indicate that there may be different sorts of meta-awarenesses to be experienced, and one that involves being broadly mindful may avoid some of the pitfalls of other varieties.

Conclusion

Meta-awareness is likely unique to humans. It represents the ability to step outside one's own thoughts and reflect upon them. This function of meta-awareness endows a high degree of flexibility, such as the ability to monitor and control cognition. While meta-awareness confers incredibly flexibility, it does not always work as one might wish. For one, meta-awareness is limited and costly to sustain. Such can be seen in temporal dissociations such as mind wandering. Attention drifts during these lapses in meta-awareness allow individuals to get off track, and perhaps more importantly fail to notice they are off track. A lack of meta-awareness may even be confused with the feeling of having never known, such as with the problem of recovered memories of sexual abuse.

Meta-awareness is also, by nature, imperfect. When consciousness is reanalyzed and reappraised, mistakes can be made – seldom is the time when the copy is as robust as the original. Translation dissociations such as verbal overshadowing illustrate this pitfall of meta-awareness. In such cases a verbal recoding of a perceptual experience, like the image of a face or the bouquet of a wine, are not as good as the original. This mismatch between experience and reappraisal can cause problems with memory as well as judgments and decisions. Translation dissociations can also occur if the motivation to accurately reflect on a memory is not present, or if theories about how an experience should be are faulty.

Although the two-edged nature of meta-awareness may never be completely avoidable, some forms of meta-awareness may be more beneficial than others. For example, explicit analytic reflection may be more disruptive than the more intuitive mindful awareness that develops through contemplative practices such as meditation. As research continues on the nature of these different flavors of meta-awareness, students and researchers can expect a deeper knowledge of how to harness this fundamentally human ability.

See also: The Control of Mnemonic Awareness; Philosophical Accounts of Self-Awareness and Introspection; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Csikszentmihalyi M and LeFevre J (1989) Optimal experience in work and leisure. Journal of Personality and Social Psychology 56(5): 815–822.

Flavell JH (1979) Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist 34: 906–911.

Lambie JA and Marcel AJ (2002) Consciousness and the varieties of emotion experience: A theoretical framework. Psychological Review 109: 219–259.

Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, and Macrae CN (2007) Wandering minds: The default network and stimulus independent thought. Science 315: 393–395.

Nelson TO (1996) Consciousness and metacognition. American Psychologist 51: 102–116.

Norman DA and Shallice T (1986) Attention to action: Willed and automatic control of behaviour. In: Davidson RJ, Schwartz GE, and Shapiro D (eds.) Consciousness and Self-Regulation: Advances in Research and Theory. New York: Plenum Press.

Schooler JW (2002) Re-representing consciousness: Dissociations between consciousness and meta-consciousness. Trends in Cognitive Science 6: 339–344.

Schooler JW, Ariely D, and Loewenstein G (2003) The pursuit and monitoring of happiness can be self-defeating. In: Carrillo J and Brocas I (eds.) Psychology and Economics, pp. 41–70. Oxford, GB: Oxford University Press.

Schooler J and Schreiber CA (2004) Experience, meta-consciousness, and the paradox of introspection. Journal of Consciousness Studies 11(7–8): 17–39.

Smallwood J, Beach E, Schooler JW, and Handy TC (2008) Going AWOL in the brain: Mind wandering reduces cortical analysis of external events. Journal of Cognitive Neuroscience 20: 458–469.

Smallwood J and Schooler JW (2006) The restless mind. Psychological Bulletin 132: 946–958.

Teasdale JD, Moore RG, Hayhurst H, Pope M, Williams S, and Segal Z (2002) Metacognitive awareness and prevention of relapse in depression: Empirical evidence. Journal of Consulting and Clinical Psychology 70: 275–287.

Winkielman PW and Schooler JW (in press) Unconscious, conscious, and metaconscious in social cognition. In: Strack F and Förster J (eds.) Social Cognition: The Basis of Human Interaction. Philadelphia: Psychology Press.

Biographical Sketch

Jonathan W. Schooler is a professor of psychology in the University of California at Santa Barbara. He pursues research on consciousness, memory, the relationship between language and thought, problem-solving, and decision-making. A *cum laude* graduate of New York's Hamilton College, Dr Schooler earned his PhD in psychology at the University of Washington in 1987. He joined the faculty of the University of Pittsburgh as an assistant professor, and in 2004 he accepted and held the position of a full professor and the Canada Research Chair in Social Cognitive Science until 2007 when he accepted his current position. A fellow of the Association for Psychological Science, Dr Schooler has been the recipient of three Akumal Scholar Awards from the Positive Psychology Network, an Osher Fellowship given by the Exploratorium Science Museum, and a Lilly Foundation Teaching Fellowship. His work has been supported, among others, by the National Institute of Mental Health, the Office of Educational Research, and Canada's Social Sciences and Humanities Research Council. He currently is on the editorial boards of *Consciousness and Cognition* and *Social Cognitive and Affective Neuroscience*. Dr Schooler is the author or coauthor of more than 100 papers published in scientific journals and the editor (with J.C. Cohen) of *Scientific Approaches to Consciousness*.

Jason M. Chin is a PhD student at the University of British Columbia studying social psychology. He completed his undergraduate work at the University of Virginia, where he majored in economics and psychology, graduating *cum laude*. At the University of Virginia, Mr Chin researched the psychological underpinnings of aversion to risk, suggesting that it stems from a tendency to overweight future negative emotional reactions. He earned his MA in psychology in 2005 from his present university for his work on verbal overshadowing, and he currently studies motivational and informational models explaining prosocial behavior. Mr. Chin's research has been supported by the University of British Columbia and the Izaak Killam Memorial Trust. A member of the Society for Personality and Social Psychology, he has presented his research at several conferences and is a coauthor of two scholarly publications. Mr. Chin has also served as a reviewer for several psychological publications.

The Mind–Body Problem

M Rowlands, University of Miami, Coral Gables, FL, USA

Glossary

Consciousness (phenomenal) – What it is like to have or undergo an experience; the ways things seem to feel to you when you have an experience.

Dualism – The claim that the mental phenomena are nonphysical. This can be advanced as a claim about token mental phenomena – the result typically being a form of 'substance dualism.' Or it can be advanced as a claim about types of mental phenomena – resulting in what is known as 'property dualism.'

Functionalism – The claim that mental phenomena are defined by their functional role, that is, by their place in a systematic network of causally related mental phenomena.

Identity theory – The claim that mental phenomena are identical with (i.e., one and the same thing as) physical phenomena. This claim can take two forms. According to the 'type-identity theory,' types of mental phenomena are identical with types of physical phenomena. According to the 'token identity theory,' mental phenomena tokens are identical with physical phenomena tokens.

Intentionality – The aboutness of mental mental phenomena; the directedness of mental phenomena toward objects apparently distinct from them.

Supervenience – A one-way relation of dependence or determination that obtains between properties. A property F is supervenient on a property G when, necessarily, any two objects that are identical with regard to G must also be identical with regard to F.

Tokens – Individual, particular instances of kinds.

Types – General kinds or categories.

Introduction: Mind–Body Problems

The mind body problem has been traditionally understood as the problem of understanding the relation between the mind and the body. However, since the nature of the body is usually assumed to be uncontroversial, this problem has found itself ruminating largely on the nature of the mind. The mind body problem is, in effect, the problem of understanding the nature of the mind. However, since the conclusion of at least some of these ruminations is the rejection of the view of the mind as a substance distinct from mental states and processes, the problem can less tendentiously be formulated as follows: what is the nature of mental phenomena? Upon analysis, however, even this question fragments into two distinct questions, and this fragmentation follows from the nature of mental phenomena.

The concept of a mental phenomenon can be given a preliminary specification by way of ostension. Mental phenomena include beliefs, desires, thoughts, feelings, hopes, fears, expectations, pains, emotions, experiences, and so on. It is common to divide up these phenomena into two categories: sensations and propositional attitudes. Sensations are mental items defined by what it is like to have them or (equivalently) by their possessing phenomenal consciousness. Pain is the paradigm example of a sensation: to be in pain is for things to feel a certain way to the subject of the pain. Propositional attitudes, on the other hand, are mental items defined by their possession of propositional content. Beliefs, desires, and thoughts are paradigm examples of such states. Beliefs are attributed to a subject using an embedded that-clause. Jones believes that p, where p stands for some or other proposition. For example, she believes that the cat is on the mat. The sentence the cat is on the mat has a particular content, or expresses a particular proposition. The content of Jones belief is identical with the proposition

expressed by the sentence that we use to attribute the sentence to Jones. The same is true for thoughts, desires, hopes, fears, expectations, anticipations, and so on. This is why such items are referred to as propositional attitudes.

The distinction between sensations and propositional attitudes is, therefore, a distinction between mental items that are defined by their possession of phenomenal consciousness and those items defined by their possession of propositional or semantic content. While this distinction is common, it is not entirely clear that it is either adequate or illuminating. With regard to the issue of adequacy, for example, it seems far from clear that the distinction exhausts the category of mental phenomena. Experiences, for example, seem to occupy a curiously ambiguous position being characterizable both by their propositional content and what it is like to have them. Some suggest that this ambiguity can be accommodated by recognizing that the content possessed by experiences is of a different kind from that possessed by propositional attitudes: the latter content is conceptual, the former merely nonconceptual. However, this suggestion faces significant difficulties. For the distinction to be illuminating we would have to understand in what the difference between possession of phenomenal consciousness and propositional content consists. However, this difference is far from transparent. Often, the distinction is glossed in terms of the idea that propositional attitudes are about other things, whereas sensations are not. Or, more plausibly, propositional attitudes are about other things in a way that sensations are not. However, this too is controversial. An increasingly popular position today, for example, is a form of representationalism or inseparatism that holds that phenomenal consciousness consists in a certain type of (representational) content. While this view is most commonly applied to the content of experiences, attempts have been made to apply it to sensations also.

This article is not the place to address these issues. These sorts of controversies, and these failures of transparency, are endemic in discussions of the mind body problem. Here, we can simply note them and work around them in so far as this is possible. In this vein, we should note that mental phenomena seem to be characterized by two sorts of property. First, many mental states arguably all, but at the very least many are defined by the fact that there is something that it is like to have or undergo them. This, I shall henceforth refer to simply as consciousness. This is most obviously true of sensations and experiences. The propositional attitudes, on the other hand, can exist in nonconscious or perhaps even unconscious form. Nevertheless, it is arguable that for a state to be mental it must at least be the sort of thing that can, in appropriate circumstances, become conscious. Second, many mental states arguably all, but at the very least many are defined by the fact that they are about other things in such a way that they possess content. This feature of mental states their aboutness is traditionally referred to as their intentionality. If we ignore the rather vexed issue of which states are principally characterized by which consciousness or intentionality and the even more vexed issue of the relation between the two then the claim that consciousness and intentionality are defining features of mental items seems a relatively safe one. Therefore, I shall label consciousness and intentionality as the essences of the mental.

The mind body problem is the problem of answering this question: what is the nature of mental phenomena? Now, however, we see why this question fragments into two: it can be understood as a question about mental items, or as one about the essences of mental items. That is, the question must be replaced with two:

1. What is the nature of mental items?
2. What is the nature of the defining essences of the mental consciousness and intentionality?

I shall refer to the first question as the item problem and the second as the essence problem. The item problem can, in turn, be divided into two because the concept of a mental item can be divided into two. Thus, the question can be understood as a question about mental tokens or one concerning mental types. Mental types are kinds; mental tokens are (usually understood as) concrete dated, unrepeatable instances of those kinds. Pain, in general, would be a type of mental item. The event of pain you suffer at 14.05 p.m. on 9 August 2007, and which lasted for precisely 15 s, would be a token of that type. In the literature,

tokens are often referred to as events, whereas types are referred to as properties. The latter terminology is unfortunate, since it invites confusion of types and essences consciousness and intentionality are also often referred to as properties possessed by mental items. To avoid this, I shall eschew talk of properties, and distinguish clearly between mental types and essences. Pain is a mental type. What it is like to be in pain is a mental essence.

The Item Problem

As we have seen, the item problem is really two problems:

(1a) What is the nature of token mental items?
(1b) What is the nature of mental item types?

(1a) Mental Tokens. If we ignore some of the more idiosyncratic proposals thrown up by history, answers to (1a) divide into two primary sorts. The first is that mental tokens are identical with physical tokens. This view is known as the token-identity theory. This theory is one version of materialism: the view that mental phenomena are ultimately physical. The other is that mental tokens are nonphysical events. This view is known as dualism.

According to dualism, token mental events are nonphysical. But events are often thought of as changes of an object with respect to a property or properties at a time or through a period of time. Therefore, the claim that mental events are nonphysical is typically understood as the claim that they consist in changes in a nonphysical object or substance. The result is what is known as substance dualism.

According to substance dualism, token mental events are identical with events occurring in an immaterial substance. This view is still principally associated with Rene Descartes. Descartes dualism Cartesian dualism was shaped by a particular conception of the defining essence of the mental for Descartes this essence is thought and his insistence that this defining essence could not be realized in any physical structure or mechanism. The essence of physical things is extension. Mental items are, therefore, nonextended. One may question whether Descartes was correct in his conception of the essence both of the mental and physical. But any form of substance dualism is going to regard mental and physical tokens as being of essentially different sorts. And so it is possible to eschew talk of substances and simply formulate dualism in terms of the nonidentity of mental and physical tokens. That is, dualism can be understood as the view that token mental events, states, or processes are distinct from any physical events, states, or processes. We might call this a dualism of particulars. Both substance dualism and this dualism of particulars are attempts to answer (1a).

Problems with dualism (of both kinds) typically divide into two related sorts. The first concerns understanding exactly what the mind is. In its typical formulations, dualism is a negative thesis: token mental events are not identical with token physical events; mental substances are not identical with physical substances. This sort of formulation is often supplemented by additional claims dualists make about what the mind does not have: is not extended, has no physical properties, is not subject to physical laws of nature. Negative definitions such as this are, by themselves, obviously inadequate. First, they do not even suffice to fix the content of the dualist thesis: it is not, in general, possible to explain or identify what an item is by providing a list of everything that item is not. Second, purely negative characterizations invite the charge that dualism is not even in the business of providing a theory of the mind but is, rather, a theoretical hole just waiting for a theory to be put in its place. That is, the worry is that dualism makes no genuine attempt to say what the mind is, how it is constructed, and what laws or principles govern its operation.

Of course, dualism is not restricted to providing a purely negative account of the mind. Then, however, the worry is that the sorts of positive things the dualist is able to say about the mind are insufficient to distinguish dualism from competing views. The materialist, for example, might accept with equanimity Descartes assertion that the essence of the mental is thinking. For thinking is something that the mind does it expresses a functional concept and this, it is, so the materialist would argue, as amenable to materialist analysis just as much as it is to dualist explanation.

The second sort of problem afflicting dualism concerns how to reinsert the mind back into the natural order once it has been identified as being so different from it. It is a commonplace, for example, that mind and world can affect each other. This is often understood in causal terms. Things occurring in the world can cause things to occur in the mind, and vice versa. However, what is the basis of this causal interaction? The danger is that dualism makes mental events so different from physical ones that it becomes impossible to understand how they can causally affect one another. This is, in fact, the traditional objection to Descartes dualism. According to Descartes, the mind has no extension: it does not occupy space. But our ordinary concept of causation seems bound up with the concept of space, in particular, with the idea of spatial contiguity. It may well be that our ordinary concept of causation is na¤ve. But, if so, it is incumbent on the dualist to supply an alternative in terms of which mental physical causal interaction is intelligible. This traditional objection to Descartes dualism is, in fact, an instance of a wider worry: if the mind is as different from anything physical as the dualist maintains, then how do we explain its manifest interconnections causal and developmental with the physical world?

These worries might lead one to suppose that materialism in the form of the token-identity theory wins by default. Most arguments for materialism rest, in one way or another, on the manifest connections between mental and physical phenomena. For example, in developing his anomalous monism an influential version of the token identity theory Donald Davidson restricted his argument to mental events that causally interacted with physical events. Davidson was convinced that he could demonstrate the identity of all such token mental events with token physical events, but remained neutral on the status of any mental event that did not so interact.

However, the token identity also faces its problems: one concerning the idea of identity, the other concerning the idea of the physical. Firstly, it is not clear why mental event-tokens would have to be the sorts of things that that can be identical with physical events tokens. To see why, consider another category of events economic events. When I pay for something in a shop, where does the event of paying take place? Suppose, for example, I am paying by credit card. Does the event of paying occur at the moment of hand-to-hand-transfer of the card? Or does it incorporate the movement of my wallet from my pocket to my hand? Or, does it incorporate the movement of the teller to the machine? Or does it occur during subsequent bank-to-bank transfers of funds? It is not clear that there is any principled way of answering these questions: economic events may simply be spatially ill-defined. If this is true, then it is difficult to understand the sense in which economic events could be identifiable with physical events, if we assume these are spatially well-defined.

The problem for the token-identity theory then, is this: the theory implicitly presupposes a fairly restrictive conception of mental events that they are not at all like economic events, for example. This presupposition has not been subjected to the appropriate amount of critical scrutiny and, to this extent, is unjustified. Faced with this sort of problem, the materialist might be tempted to abandon the idea of mental physical token-identity in favor of some sort of mental physical composition thesis: while mental events are not identical with physical events, they are composed of physical events (with the added proviso that a composite event is not identical with the sum of its parts). This response, however, does not seem to avoid the problem. If physical events are spatially determinate then no matter how we combine physical events, if the composition is mereological we will still end up with a spatially determinate composite event and. And if mental events are not spatially determinate, then they cannot be composed of a spatially determinate sum. In other words, this compositional version of materialism has not jettisoned the conception of mental events that led to the objection to the token-identity theory.

The other option for the materialist is to abandon the idea that physical events are spatially determinate. This is not the same as assuming that physical events are spatially indeterminate, it is simply to remain noncommittal on the issue. By adopting this strategy, the identity theory is no longer vitiated by an unjustified assumption concerning the nature of mental events: the theory can now remain noncommittal about the spatial character of both mental and physical events.

However, this strategy raises problems of its own. In particular, it is not clear how much substance there remains to the idea of identity (or for that matter composition). In particular, the theory now faces the formidable problem of understanding the sense in which one potentially spatially indeterminate event can be identical with another equally spatially indeterminate event.

The second problem concerns what it is for something to be physical. There is actually a cluster of worries here, but perhaps the most significant one was identified by Bertrand Russell as long ago as 1927. It is reasonable to suppose that physics the basic science of matter will tell us what it is for something to be physical. Indeed, if physics does not tell us this, it is difficult to see what would. However, Russell notes that the knowledge of matter supplied by physics is abstract, purely formal, merely a matter of structure. He writes:

> Physics is mathematical not because we know so much about the physical world but because we know so little; it is only its mathematical properties that we can discover. For the rest, our knowledge is negative ... The physical world is only known as regards certain abstract features of its space-time structure features which, because of their abstractness, do not suffice to show whether the physical world is, or is not, different in intrinsic character from the world of mind. (1948: 240)

Physics proceeds by the identification of an abstract mathematical formulism that will detail the relations between things. However, intuitively, not definitely by any means, but this is the way we tend to think about things there are not just relations. When things enter into relations, they do so because of the intrinsic natures of the things that are related. Relations are ontologically dependent on things, and things have natures that explain the relations those things enter into. However, of such natures the natures that ground the relations between things physics, Russell argues, can tell us nothing. The moral, for present purposes, is this: if you think that what makes something physical is its underlying nature, a nature that provides the basis of the relations that physical things enter into, relations described by physics, then physics gives us no idea what it is to be physical. The dualist s failure to tell us anything significant about the mental is, therefore, arguably mirrored by the materialist s failure to tell us anything significant about the physical. And if this is so, any purported victory of materialism over dualism has a distinctly hollow ring.

(1b) Mental Types. Let us overlook these problems, assume we do understand what it is to be physical, and continue with the development of a materialist conception of mental phenomena. It is fairly clear that a token-identity theory will not, by itself, do all the required work. The token-identity theory, in itself, is a very weak doctrine. To begin with, the identity of mental and physical tokens is compatible with a dualism of mental and physical types: the claim that every mental event-token is identical with a physical event-token is compatible with the claim that mental types are distinct from and irreducible to physical types. This creates a problem for the materialist. It would be possible for two individuals to instantiate exactly the same physical types but mentally have nothing in common at all. For example, two molecule-for-molecule duplicates might share no mental types at all. Each mental token they undergo is identical with a physical token, but the mental types that these tokens instantiate are, in each case, different. The condition of token-identity has been satisfied, but, from a materialist perspective, there clearly seems to be something missing. What is missing is the idea that the way the world is physically determines the way it is in every other respect. If we do not at least secure this claim there is little substance left to the idea of materialism. And to secure this claim we must tighten the relation between mental and physical types: we must make the former in some way dependent on the latter.

One way of doing this is to suppose that mental types straightforwardly reduce to physical types. The type-identity theory is a common way of understanding this reductive thesis. According to this theory, mental types are identical with physical types. For example, pain pain as a type, not a specific instance of pain, but pain in general might be identified with C-fiber firing. Pain, so the idea goes, is identical with C-fiber firing in much the same way that water is identical with H_2O; in much the same way that heat is identical with molecular motion, or lightning is identical with an

electrical discharge to earth from a cloud of ionized water particles.

This type-identity theory is thought by many to be fatally susceptible to what is known as the problem of variable realization. As Kripke has emphasized, identity is a necessary relation in this sense: If X is, in fact, identical with Y, then necessarily X is identical with Y. X could not have been nonidentical with Y. For example, if water is, in fact, H_2O, then water could not have been anything other than H_2O. We might think that it could. Why could not water be identical with some entirely different chemical substance? But to entertain this possibility is simply to describe a situation in which there is a substance that looks, tastes, feels, etc., exactly like water but really is something else.

This thesis of the necessity of identity creates problems for the type-identity theory. For it certainly seems possible that mental types can, in different creatures, or even in the same creatures at different times, be realized by different physical types. Imagine, for example, a Martian who felt pain in exactly the same way that we did, but where this feeling was realized not by c-fiber firing but by some entirely different sort of neural state. What is decisive is determining whether the Martian is in pain is the way things feel to it when it undergoes the event. If things feel the requisite way it is in pain is pain irrespective of the state of its neural machinery or whether it even has something we would recognize as neural machinery. This argument does not depend on the actual existence of Martians or other extraterrestrials. Their genuine possibility is enough. If, as the principle of the necessity of identity claims, identity is a necessary relation, then the mere possibility that pain might, in different creatures, be realized by different neural states or structures, is enough to preclude the type-identification of pain with C-fiber firing.

These sorts of reasons the possibility of variable realization coupled with the necessity of identity have led to widespread, if not universal, rejection of the type-identity theory. And if the type-identity theory is, indeed, untenable we need another way of formulating the idea that mental types are dependent on physical types. The usual response, here, is to appeal to the concept of supervenience.

The idea of supervenience is, in essence, a simple one: it is a one-way relation of dependence. With regard to the relation between mental and physical types, the idea of supervenience can be expressed thus: mental types supervene on physical types if and only if any two creatures that are the same with regard to the physical types they instantiate must also be identical with regard to the mental types they instantiate. Equivalently, if a creature changes with regard to the mental types if instantiates, then it must also change with regard to the physical types it instantiates. The dependence is one way. It does not follow, that if any two creatures identical with regard to the mental types they instantiate must also be identical with regard to the physical types they instantiate, or that a creature that changes with regard to the physical types it instantiates must also change with regard to the mental types it instantiates. It is the fact that the concept of supervenience expresses only a one-way relation of dependence that distinguishes it from the type-identity relation.

The concept of supervenience, while in essence a simple one, is notoriously difficult to pin down, and many journal and book pages have been devoted to a delicate juggling of the scope and strength of modal operators flagged in the above definition by the term must required to get the concept just right. We do not have the space here to pursue these issues.

Instead, let us pursue a more pressing issue. What reason, if any, do we have for supposing that this supervenience thesis is true? What reason, that is, do we have for supposing that mental types supervene on physical ones? It is precisely here that the importance of the view known as functionalism can be properly appreciated.

According to functionalism, mental phenomena are identified according to their causal or functional role. Consider an analogy. A carburetor is a physical object located somewhere in the innards of a car s engine (or older cars anyway fuel injection systems have replaced them in more recent models). What is a carburetor? Or, more precisely, what makes something a carburetor? The answer is that a carburetor is defined by what it does. Roughly, it is something that takes in fuel from the fuel inlet manifold, takes in air from the air inlet manifold, mixes the two in an appropriate ratio, and sends the resulting mixture on to the combustion chamber. It is fulfilling this role that makes something a carburetor, and

anything that fulfils this role in a car thereby counts as a carburetor. Most carburetors tend to look pretty similar. But this is at best a contingent fact, because it does not matter what a carburetor looks like as long as it fills this role. The details of its physical structure and implementation are of secondary importance compared to the role it fills, for it is filling this role that makes something a carburetor, and not the details of its physical structure or implementation. Of course, not every physical thing is capable of playing the role of a carburetor. A lump of Jell-O inserted into your car engine would have a hard time mixing fuel and air or doing anything else for that matter. A lump of Jell-O is simply not the right sort of thing for fulfilling the functional role of a carburetor. So, the details of how the functional role is physically implemented are not irrelevant. But as long as you have a suitable physical structure one that is capable of fulfilling the role of a carburetor, then it does not matter what it is as long as it, in fact, fulfills this role. If it does, then it is a carburetor.

Functionalism takes a similar view of the nature of mental properties. That is, such properties are defined by what they do by their functional role. What is it that mental phenomena do? Fundamentally, they relate to each other, to perception and to behavior in various complex, but in principle analyzable, ways. Take a belief, for example, the belief that it is raining. This is a belief that is typically caused by perception of certain environmental conditions, rain being the most obvious. Of course, perception of other environmental conditions might also produce the belief; for example, someone, unbeknownst to you, using a hosepipe outside your window. But rain is the most typical cause of the belief. The belief can, in turn, go on to produce certain sorts of behavior. Because of your belief you might, for example, carry an umbrella with you when you leave the house. However, the belief has these sorts of ramifications for your behavior not in isolation but only in combination with other mental states. You will carry the umbrella because you believe it is raining, but only if you also want to stay dry. Your desire to stay dry is necessary for your behavior too. And this belief-desire combination will produce your behavior only if you also believe that the umbrella will keep you dry; only if you believe that it is not too windy to use an umbrella; indeed, only if you believe that what you have picked up is an umbrella, and so on. What emerges is a complex network of mental states, perception, and behavior. According to functionalism, each mental state is defined by its place in this network: by the relations in which it stands to perception, to other mental states, and to behavior. To specify the place of a mental state in this network is, according to functionalism, to define that state. Of course, any such definition would be grotesquely long. But, its practicalities aside, the strength of functionalism consists in giving us a general vision of what mental phenomena are. The vision is of mental phenomena forming a vast causal system a system of interrelated causal connections where each mental property is individuated by way of its place in this system.

Functionalism provides a clear explanation for why mental types would supervene on physical ones. A mental type is defined by its functional role, and for any such role there is a physical type that fills it. Given the necessity of identity and the possibility of alternative physical types filling that role, we cannot identify the mental type with the physical type. But the filling, by a physical type, of the role that defines a mental type is sufficient for that mental type to supervene on the physical type.

During the second half of the twentieth century, attempts to answer the item question gradually coalesced around a position that is, by now, sufficiently widely accepted to be regarded as the default position. This position is made up of three claims: (1) identity of mental and physical event-tokens, (2) supervenience of mental types on physical types, where this supervenience is underwritten by (3) a functionalist account of mental types. That this combination provides the default position on the item question is, of course, no guarantee that it is correct.

Consciousness and the Essence Problem

Recent work on the mind body problem has tended to focus on the essence rather than the item problem. And the apparatus designed to solve the item problem token-identity, supervenience, and functionalism allows us to make only

limited headway on the essence problem. The essence problem can be developed for both putative defining essences of the mental: consciousness and intentionality. Due to constraints of space, however, I shall discuss only the former.

Many suspect that the existence of phenomenal consciousness what it is like to undergo an experience provides a serious, perhaps even intractable problem for materialist accounts of the mind. The basis of this suspicion is to be found in an intuition, one eloquently expressed by Colin McGinn:

> How is it possible for conscious states to depend on brain states? How can Technicolor phenomenology arise from soggy grey matter? What makes the bodily organ we call the brain so radically different from other bodily organs, say the kidneys the body parts without a trace of consciousness? How could the aggregation of millions of individually insentient neurones generate subjective awareness? We know that brains are the *de facto* causal basis of consciousness, but we have, it seems, no understanding whatsoever of how this can be so. It strikes us as miraculous, eerie, even faintly comic. Somehow, we feel, the water of the brain is turned into the wine of consciousness, but we draw a total blank on the nature of this conversion. (1991: 1)

I shall refer to this as the intuition. Probably the safest way of characterizing the arguments of those who think that consciousness is materialistically problematic is as a series of developments, explications, refinements, and defences of this intuition. If these developments, explications, refinements, and defences are successful, then collectively they add up to a serious obstruction to the project of accommodating consciousness into a materialist world view. The intuition has been refined, elaborated, and supported by several well-known thought experiments: imaginative scenarios whose purpose is to make the intuition particularly concrete and graphic. Here are two of the most famous:

1. Some bats navigate their environment by means of echo location: they emit high-frequency noises and use the pattern of echoes to build up a cognitive map of their environment. Thomas Nagel invites us to imagine we know everything there is to know about the neural mechanisms that allow the bat to do this. Still, Nagel intuits, we would not know what it is like to be a bat. This intuition he develops into a general argument against materialism, an argument that turns on the idea of subjectivity.
2. Frank Jackson invites us to consider the case of Mary. Mary lives in an entirely monochromatic environment never seeing any colors except black, white, and shades of grey. However, despite this impediment, she has become the world s leading neuroscientist knowing, in particular, everything there is to know about the neurophysiology of color vision. That is, she knows everything there is to know about the neural processes involved in visual information processing, the physics of optical processes, the physical makeup of objects in the environment, etc., What she does not know, however, is what it is like to see color. Thus, Jackson argues, when she leaves her room for the first time and sees a bright red object, she learns something new what it is like to see red. Now she knows something she did not know before, and this new knowledge could not have been constructed or derived from her previous knowledge. Jackson thinks this spells the demise of materialism.

Much work has been done examining and evaluating these thought experiments. However, the thought experiments are predicated on the intuition, in the sense that the former are ways of making the latter graphic. Therefore, I shall here focus on the intuition itself. There are, broadly speaking three different stances one can adopt with regard to the intuition: (1) one can reject it, (2) one can accept it but restrict its consequences, or (3) one can accept it. I shall begin with (2), and then discuss (1) as a way of objecting to (2). Finally, I shall make a few brief remarks about (3).

The Epistemological Interpretation: The Explanatory Gap

One thing that the intuition might be taken to show is the existence of an explanatory gap between mental and physical items. Perhaps the most comprehensive development of this idea is due to McGinn. According to McGinn, the explanatory gap is grounded in a conceptual poverty that is, in turn, grounded in a poverty of

faculties. The explanatory gap arises because we do not have the requisite concepts to apply to the natural order concepts that would allow us to see how it produces or constitutes consciousness. But this, in turn, is derived from our lack of the appropriate faculties concept-forming capacities that would allow us to acquire the requisite concepts. McGinn s position is characterized by: (1) ontological naturalism: consciousness is a natural feature of the world, and (2) epistemic irreducibility: there is no and probably can be no explanation of consciousness available to us.

McGinn s development of his mysterian position can be regarded as comprising three interwoven strands. The first strand is concerned with differences in the ways we know about consciousness and the natural world (including the brain). We know about consciousness through introspection. In this, consciousness is unique; our knowledge of the rest of the world is grounded in perception. The second strand develops the idea that consciousness has a nonspatial character. Because our knowledge of consciousness is grounded in introspection, consciousness presents itself to us as nonspatial. And, in this, consciousness is again unique. Thus an idiosyncratic mode of access to consciousness yields an idiosyncratic feature of consciousness. Because of this idiosyncratic feature of consciousness, we will encounter major problems trying to incorporate consciousness into the natural order. The third strand revolves around inherent limitations to our cognitive capacities. Our failure to incorporate consciousness into the natural order is not the disaster many have supposed. To the extent there is a disaster, it is an epistemic, not ontological one. The problems ultimately stem from natural limitations on our cognitive capacities, limitations which make the problem of consciousness insoluble for us, but not for a creature with the appropriate cognitive faculties.

Objections to the Explanatory Gap

In his commitment to ontological naturalism coupled with epistemic irreducibility, McGinn s position is a version of stance (2) with regard to the intuition: acceptance of the intuition combined with restriction on its limitations. Specifically, the intuition is argued to have only epistemic, rather than ontic, consequences. Stance (1) the rejection of the intuition can best be pursued by way of critique of McGinn s position. This position involves two logically distinct claims:

1. An explanation of consciousness must proceed by way of identification of a mechanism.

If an explanation of consciousness required only correlations between neural and conscious states, there would be no deep problem of consciousness. Further, if this underlying mechanism is to explain consciousness, it must do so by eliciting in us a certain kind of insight:

2. The neural mechanism that explains consciousness must allow us to see how consciousness is produced by the brain.

Accordingly, a genuine explanation of consciousness works only to the extent that it allays the feeling of mystery that attends our contemplation of the brain mind link . Both (1) and (2) are not unassailable.

One possible objection to (2) is that it involves a conflation of the concept of explanatory adequacy with what we might call epistemic satisfaction. Some explanations produce in us a feeling of epistemic satisfaction: a Eureka! feeling Now I understand! or in more Wittgensteinian mode, Now I can go on! The molecular explanation of the macroproperties of graphite provides a good example of an explanation likely to elicit this sort of feeling. Graphite consists in layers of carbon arranged into hexagonal rings. The atoms in each layer are covalently bonded to three neighboring atoms at an angle of 120 to each other. Within each layer, the covalent forces binding each atom to its neighbor are relatively strong. However, the layers themselves are bound together only by the very weak van der Waals forces. As a result, adjacent layers can slide over each other resulting in the soft, flaky, nature of graphite, its ability to mark paper, act as a lubricant, etc.

A focus on explanations of this sort might tempt us into thinking that the adequacy of an explanation is to be judged by whether it elicits in us a feeling of epistemic satisfaction. And this assumption is

questionable. Consider, for example, the molecular explanation of solidity in terms of a rigid lattice structure of atoms held together by ionic binding. How, one might reasonably ask, can a solid object be made up mostly of empty space? How could such an item, for example, retain its volume? An obvious response is to explain away any lack of epistemic satisfaction in terms of our empirical ignorance specifically, of relevant atomic or quantum level facts and laws. For example, we might explain the disposition of solids to retain their volume in terms of the characteristics of the specifically ionic bonding that seems to be responsible for this ability. Ionic bonding involves electron transfer of electrons, rather then merely their sharing, and so ionic bonds are very strong. But this merely pushes the problem back a step. Why should bonds that involve transfer of electrons be any stronger than bonds which merely involve their sharing? What reasons are there for supposing that this explanation will be any more epistemically satisfying than the original?

We can push the explanation back further, and explain the salient characteristics of ionic bonding in terms of wave interaction, superposition, and so on. Perhaps, once we acquaint ourselves with the relevant laws of wave dynamics, then everything else will fall into place? But, once again, the same question will arise. Why must explanations cast at this level be any more epistemically satisfying than the original molecular explanation? Is it obvious, for example, why, waves should obey the laws of wave dynamics? More generally, why should the world, at a fundamental level, be an epistemically satisfying place?

The dialectic, here, is tricky because McGinn will, of course, argue precisely that the world is not an epistemically satisfying place, at least not for us; and this is the basis of his mysterian position. The present point, however, is that a lack of epistemic satisfaction need, in itself, be no impediment to recognizing that something is an explanation of a given phenomenon, and an adequate one at that. We can accept that a wave dynamical account of a phenomenon such as solidity is both true and an explanation even if it does not produce in us in any of us the sort of feeling occasioned by the molecular explanation of the macroproperties of graphite. If this is correct, then explanatory adequacy is not a function of epistemic satisfaction: explanatory adequacy does not consist in a specific inner process.

Some explanations ones that we recognize as adequate possess a sort of inchoate, protoversion, of epistemic satisfaction: protoepistemic satisfaction. At the core of this concept is the notion of analogy. Many of our best theories have their origin in provocative initial analogy; one that may be seriously flawed, but subsequently proved to be a fruitful vehicle of understanding. Consider, again, the molecular explanation of solidity. While this may not occasion the sort of epistemic satisfaction elicited by other explanations, it does produce a certain form of enlightenment carried, to a considerable extent, by the relations between properties of the reduced domain and those of the reducing. Thus, suppose we accept that a given solid is composed of a lattice structure of atoms tightly bound together, each oscillating around a fixed point. We can then, with relative ease, accept that the addition of energy to this structure might increase the frequency of this oscillation. And then, also, that the addition of sufficient energy might increase the oscillatory frequency to such an extent that the bonds break down. And the addition of further energy might increase this breakdown further. So, if we accept that solids are made up of a rigid lattice structure of oscillating atoms, then we can also see that the difference between this sort of structure and one where the bonds are more diffuse is something like, somewhat analogous to, the difference between a solid and a liquid. And, in virtue of this sort of rough analogy the molecular explanation of solidity possesses a certain protoepistemic satisfaction.

While it is plausible to suppose that any explanation we recognize as an explanation must elicit some or other psychological states in us, the precise nature of these states may vary considerably from one explanation to another varying from, at one extreme, the full-blown Eureka! feeling to, at the other, a nebulous, imprecise, and analogy-based form of protoepistemic satisfaction. The latter form of understanding can then be reinforced by the sorts of social pressures characteristic of a scientific education.

Consider, now, claim (1). This is the claim that mere correlation of neural and conscious states is not sufficient for an explanation of consciousness: that

would require identification of a mechanism. The distinction between mechanisms and correlations is, however, a questionable one. Specifically, the sort of enlightenment provided by mechanisms consists in the breaking down a correlation into a structured series of smaller correlations, where each of the smaller correlations is more readily intelligible than the original one. Mechanistic explanation is not something radically different from or opposed to the identification of correlations. On the contrary, mechanistic explanation is a specific form of correlation-based explanation. It may be that a correlation between two items can be rendered intelligible by the uncovering of an underlying mechanism. But this is not to replace the correlation with something fundamentally different; it is to break down, and thus explain, the correlation by means of further correlations.

With these points in mind, the best case that can be made for reductive naturalism, and hence against the intuition, involves three claims:

3. There is no fundamental opposition between mechanistic explanation and the identification of correlations.
4. The explanatory adequacy of correlation-based explanation does not require that it elicit in us epistemic satisfaction in any full-blooded sense.

To these principles, we can add a third:

5. There is not an explanation of consciousness. Rather, there are many such explanations as many as there are features of consciousness that require explanation.

Even in the case of properties such as solidity, there is not necessarily, any such thing as the explanation of solidity. Rather, there seem to be at least two. There is an explanation of the disposition to resist deformation and an explanation of the disposition to retain volume. Since not all rigid structures retain volume then an explanation of the former is not, in itself, and explanation of the latter.

We might expect this general point to be reiterated in the case of consciousness. The concept of consciousness almost certainly fragments, upon analysis, into several distinct concepts, including phenomenality, subjectivity, nonrelationality, and so on. If this is so, then it is likely that separate explanations will be required for each of them.

With (3) (5) in mind, consider the much maligned claim of Crick and Koch to have explained consciousness in terms of 40 Hz oscillations in the sensory cortex. Taken in itself, such a claim is, of course, laughable. However, what 40 Hz oscillations might be able to play a role in explaining is not consciousness as such, but one of its features: its gestalt character or, as we might put it, its all-at-onceness. Conscious experience is not presented serially like, for example, a description of that experience in the form of a sentence. It is presented all at once. Part though presumably not all of explaining this feature of consciousness almost certainly involves explaining the brain s capacity for binding information together into a unified whole. And this is precisely what the identification of a single oscillatory frequency might enable us to understand. It would do this not in the sense of providing us with full-blooded epistemic satisfaction with regard to the production of consciousness. Rather, it may yield a form of protoepistemic satisfaction with regard to one aspect of consciousness. That is, we can see that the gestalt character of experience is something like, somewhat analogous to, disparate information that has been bound together in various ways. Consequently, we can understand, in a somewhat nebulous manner, that changes in the quantity and types of information that are bound together at any given time might systematically vary with changes in the content of the visual gestalt.

In short, the best case that can be made against the intuition, certainly in the form presupposed by McGinn, involves arguing (a) that McGinn is committed to principles (1) and (2), but that (b) these principles should be rejected in favor of principles (3), (4), and (5), and then arguing that (c) principles (3), (4), and (5) are precisely the sort of principles that drive, in an admittedly nonreflective manner, current scientific research on consciousness.

The Ontological Interpretation: Consciousness as Outside the Natural Order

If one does not accept the arguments against intuition, then it is possible to regard this intuition as

indicative of something more than merely an epistemic hiatus. That is, one can take intuition as showing that consciousness is not a natural feature of the world. This can be understood in two very different ways.

Chalmers understands the problem of consciousness as one stemming ultimately from the ontological poverty of the sciences of consciousness (although this is reflected in an associated conceptual poverty also). To rectify this, Chalmers advocates what he calls naturalistic dualism. He allows that there is an explanation of consciousness we can understand. But to get this explanation we have to be willing to expand our catalogue of basic substances. So, while there is an ontological gap between consciousness and the basic substances currently countenanced by science and in this sense consciousness is not part of the natural world as this is currently understood this gap can be bridged by suitable adjustment of what we take the basic substances to be. The ontological gap between consciousness and the natural world is one that can be bridged through appropriate adjustment of the sciences of consciousness.

I have argued for a more radical sort of ontological gap between consciousness and the physical world. Consciousness is not a region of reality to which our access is idiosyncratic but rather it exists only in the accessing itself. There is no region of reality to which subjective phenomena belong; they simply belong to our accessing of regions of reality that are, in themselves, perfectly objective. This stems from the hybrid nature of consciousness: it can be both act and object of experience. Consciousness can be both that upon which awareness is directed (i.e., inner sense is possible) and the directing of awareness (the act of inner sensing is numerically distinct from the states or facts that it reveals to the subject). And what it is like to undergo an experience, I argued, is something that attaches to consciousness as act not object. What it is like to have an experience is not something of which we are aware in the having of that experience but, rather, something in virtue of which we are aware of distinct, and nonphenomenal, objects.

This view of consciousness has Kantian roots consciousness is a condition of possibility of objects being presented to a subject under a mode of presentation, and in this sense is a transcendental feature of the world. This transcendentalist view of consciousness, when pushed, has a striking consequence: consciousness is real but nowhere at all. In this, the position shares McGinn s emphasis on space as the problematic feature that undermines reductive explanations of consciousness. But, unlike McGinn s form of mysterianism, it also entails that there can be no explanation of consciousness at all even if our conceptual repertoire were Godlike.

Summary

The mind body problem is really two problems. The item problem concerns the nature of mental items: are they or are they not physical? The essence problem concerns the nature of the defining essences of mental phenomena consciousness and intentionality: can they or can they not be explained in physical terms. With regard to the item problem consensus gradually seems to be coalescing on a combination of (1) mental physical identity at the level of tokens, (2) mental physical supervenience at the level of types, where this supervenience is underwritten by (3) a functionalist account of the nature of mental properties. There is no such consensus with regard to the essence problem however. The problems of explaining the nature of both consciousness and intentionality are among the hottest topics in contemporary philosophy of mind. While there is much agreement about the nature of the problems, substantive solutions are still very much up for grabs.

See also: History of Philosophical Theories of Consciousness; Intentionality and Consciousness; Mental Representation and Consciousness; Self: Body Awareness and Self-Awareness.

Suggested Readings

Chalmers D (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford: Oxford University Press.
Crick F and Koch C (1994) The Astonishing Hypothesis. New York: Scribner.
Davidson D (1970) Mental events. In: Foster L and Swanson J (eds.) Experience and Theory, pp. 79–101. London: Duckworth.
Jackson F (1982) Epiphenomenal qualia. Philosophical Quarterly 32: 127–132.

Jackson F (1986) What Mary didn't know. Journal of Philosophy 83: 291–295.
Kim J (1984) Concepts of supervenience. Philosophy and Phenomenological Research 65: 153–176.
Kripke S (1980) Naming and Necessity. Cambridge, MA: Harvard University Press.
Leder D (1985) Troubles with token identity. Philosophical Studies 47: 79–94.
McGinn C (1991) The Problem of Consciousness. Oxford: Blackwell.
Nagel T (1974) What is it like to be a bat? Philosophical Review 83: 435–450.
Rowlands M (2001) The Nature of Consciousness. Cambridge: Cambridge University Press.
Russell B (1927) The Analysis of Matter. London: Kegan Paul, Trench, Trubner.
Russell B (1948) Human Knowledge; Its Scope and Limits. London: George Allen & Unwin.
Strawson G (2003) Real materialism. In: Anthony L and Hornstein N (eds.) Chomsky and His Critics, pp. 49–88. Oxford: Blackwell.
Tye M (1995) Ten Problems of Consciousness. Cambridge, MA: MIT Press.

Biographical Sketch

Mark Rowlands is a professor of philosophy at the University of Miami. He is the author of a dozen books, translated into 15 languages. These include *The Body in Mind: Understanding Cognitive Processes* (Cambridge University Press 1999), *The Nature of Consciousness* (Cambridge University Press 2001), and *Body Language: Representation in Action* (MIT Press 2006). His autobiography, *The Philosopher and the Wolf*, was published by Granta in 2008.

Mind Wandering and Other Lapses

J Smallwood, University of California, Santa Barbara, CA, USA

Glossary

Action slips – A class of error in which the individual fails to perform a well-practiced task because of a lapse in attention to the task in hand. Also known as an attentional lapse.

Current concerns – The salient personal concerns that an individual may have at any one moment in time. Generally these are considered issues which extend from medium to long-term time frames (e.g., days, weeks, and months).

Default mode network – A term coined to describe a constellation of cortical and subcortical structures which are unusually active when individuals rest in a scanner. One current hypothesis on the default network is that it provides the internal content of mind-wandering episodes.

Ironic processing theory – A theory developed by Dan Wegner to explain the regular wanderings of consciousness. According to ironic processing theory the mind wanders because of the manner that control is instantiated in the brain. One system – the intentional monitoring system is concerned with effortful control of attention, a second – the ironic monitor – checks whether the intentional monitor is currently on task.

Meta-awareness/consciousness – An intermittent state in which we periodically take stock of the contents of consciousness by reflecting on what we are thinking about and comparing this to what we thought we were doing.

Mindless reading – The temporary experience of reading a narrative without making sense of the narrative.

Mind wandering – An intermittent state when attention shifts from the task of the moment and is often accompanied by thoughts related to unrelated personal concerns. The accompanying thoughts are often referred to as task-unrelated thought or stimulus-independent thoughts.

Task-unrelated/stimulus-independent thoughts – The experience of thoughts which are generally internally generated and are unconstrained by the task in hand.

Introduction

One of the distinguishing features of the human mind is the ability to transcend the here and now using imagination. No phenomenon, however, demonstrates the emphasis that our cognitive system places on imaginative processes better than the experience of mind wandering during reading. The strong intuition that our mind has wandered from the text to some other topic of thought, not only underscores the power of the mind to transcend the here and now, it suggests that our imagination has ability to capture attention and so temporarily derail the local goal of reading. This article provides a basic summary of the current state of research into the experience of mind wandering. It describes the challenges that must be overcome when studying the wandering mind and present a simple information-processing model within which to make sense of research. In order to do this, this article considers the results of cognitive, behavioral, and neuroimaging studies which reflect the current state of knowledge of how and why mind wandering occurs in the waking brain.

Among the topics in this volume mind wandering is one of the more enigmatic – it is, for example, probably one of the few topics covered that will be as easy to explain to a layperson as to a cognitive scientist. Such a contradiction arises because, before the recent surge in interest over

the last 4 or 5 years, empirical work in mind wandering was sporadic – despite the obvious importance of mind wandering to our species, cognitive scientists simply did not study it. To understand why research on mind wandering has not matched the resonance of the phenomenon to the human condition, it is necessary to understand why states of task-unrelated thinking are problematic to study.

The Collins Concise English Dictionary defines the verb 'to wander' as 'to move or travel about in, or through (a place) without any definite purpose or destination.' One component of 'wandering,' therefore, is that it implies movement between different places. In order to apply this notion of movement or travel to the mind, it is first necessary to define the places to which the attentional system could travel. On the one hand, the standard influences in attention to which cognitive psychology is familiar are those which are necessary to complete the task. Usually tasks employed in the laboratory emphasize information contained in the external environment and so psychology has a detailed understanding of how attention is deployed on sensory information. When we engage in mind wandering, the focus of attention is withdrawn from the external environment and instead is often directed to information which is derived internally. When we catch ourselves thinking about something other than the topic of what we are reading, we are engaging in what cognitive psychologists refer to as 'task-unrelated' or 'stimulus-independent thought' – labels which capture the notion that these thoughts are not related to the immediate external environment or what we are doing. One aspect of mind wandering which sets it aside from other domains of psychological inquiry is that the content of the episodes themselves is often the product of imaginative processes. One barrier to studying mind wandering, therefore, is the need to develop tools for capturing covert attentional shifts which only indirectly available to scientific enquiry. To date, the most direct method for studying mind wandering is by asking individuals to self-report the contents of consciousness as they perform a task.

A second aspect of 'wandering' is that it implies movement or 'travel' and so the wandering mind is one that is in transition. When we mind wander during reading, focal attention shifts from the narrative toward our own thoughts. It is this state of transition or flux which motivated William James to use the term 'the stream of consciousness.' Studying the 'wandering' mind, therefore, involves studying how attention shows a tendency to ebb and flow between different mental states. A second difficulty in studying mind wandering is that it requires an experimental design which acknowledges the ebb and flow in attention.

A Simple Information-Processing Model of Mind Wandering

The first step in understanding mind wandering is developing a simple framework in which to understand the current knowledge of mind wandering. A simple way to conceptualize mind wandering and other mental states is by considering their relation to the flow of information through an attentional system, such as that presented in Figure 1. This model is based on one that I have been developing over a number of years in conjunction with Jonathan Schooler and others. Figure 1 has three different panels and each describes how a particular mental state relates to different configurations of informational flow through the attention.

Different Mental States as Different Information Flow

The first aspect to note from this model is that information can flow from two basic sources. One source of information to which we can attend arises from the external environment. Over the last 40 years psychology has developed a thorough account of how brain detects, codes, and responds to such sensory information. The second source of information to which we can attend arises from the internal environment. The brain has the ability to store representations of past events, such as memories. When we use our imagination to consider information from the internal milieu, we use these internal representations to free our mind from the constraints imposed by the outside world. Both sources of information – the internal and the external – form the contents of consciousness when they gain access to working memory.

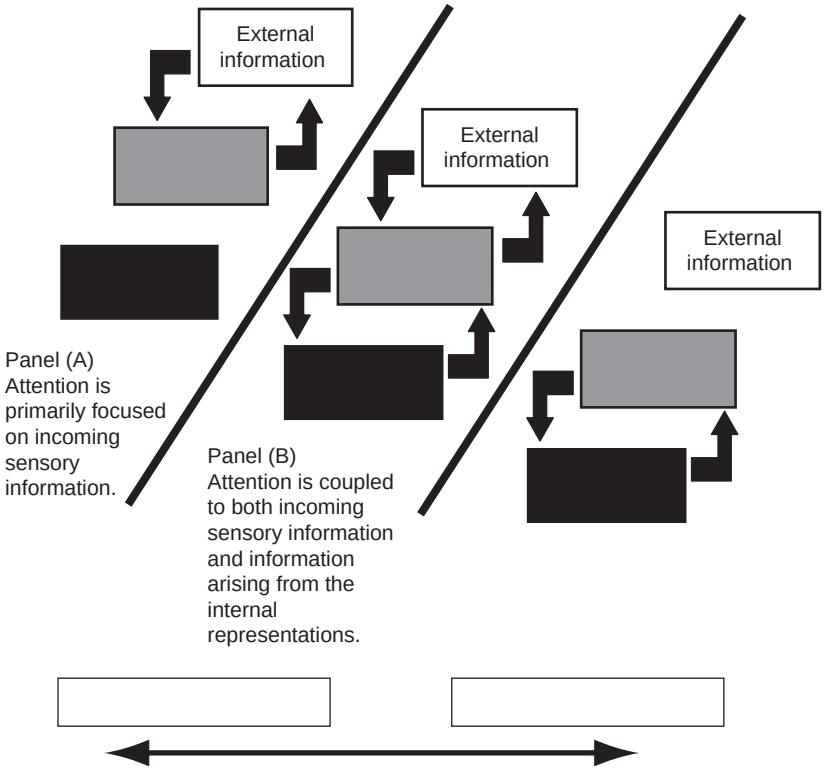


Figure 1 Schematic account of the flow of information through an attentional system and the associated mental states. Panel (A) presents the attentional coupling which accompanies the processing of information which arises from the external environment. Panel (B) the state of coupled attention associated with processing information from both internal and external information in order to perform a task. Panel (C) presents the state of decoupled processing when the mind wanders.

A general assumption in psychology is that working memory resources are limited and so, different sources of information (such as those derived from internal and external environment) compete for access to a general attentional resource usually equitable with concepts such as awareness. Information that gains access to this general-purpose mental resource becomes the focus of attention. So that when the flow of information into working memory is primarily from the internal environment, attention will be focused on this source of information and less attention will be directed externally. The model is a variant on what is called a global workspace model which has been developed by a number of different theorists, including Bernard Baars and Stanislas Dehaene. Global workspace models have provided influential accounts of many different conscious phenomena covered in this volume.

Figure 1 illustrates three states of information flow and the associated mental states. On the right-hand panel attention is focused exclusively on the external environment. The flow of information is represented by the arrows, and in the right-hand panel they indicate that information is cycling between working memory and the external environment. This mental state would involve attention giving priority to incoming sensory information, such as when we are startled by a loud noise. In the central panel, information is flowing from both internal and external environment into working memory. To reflect that both internal and external information is contributing to focal attention, arrows indicate that information is traveling from both internal and external sources into working memory. When we are engaged in a task with an extended narrative such as reading or watching a film, we use both internal and external sources of information to make sense of what is happening. Finally, in the left-hand panel, information flows from internal representations into working memory and so attention is decoupled

from both the task and the environment. In the literature, a focus on internal information processing has been described in a number of different terms – daydreaming, task-unrelated thought, and stimulus-independent thought. All of these labels emphasize a mental state in which attention is giving priority to internal information.

Transitions in Mental States as Changes in the Flow of Information

As noted in the 'Introduction' section, a second key feature of mind wandering is that it involves a state of attentional transition. It is easy to see how the dynamic aspects of mind wandering are captured by the model presented in Figure 1. When we are engaged in normal reading, information flows from both external (e.g., the text) and internal (the representation of the narrative) sources into working memory (Panel B). When the mind wanders from the narrative, the flow of information from the text is suppressed in favor of task-irrelevant information from the internal milieu (Panel C). In terms of Figure 1, therefore, when the mind wanders during reading, the direction of information flow through the attentional system shifts from Panel B to C.

Why the Mind Wanders

While the model presented in Figure 1 explains what is entailed in different mental states it does not explain why the mind should wander in the first instance. Despite being a compelling and frequent aspect of our mental life, explaining why the mind wanders is in fact one of the harder aspects of the phenomenon to accommodate into models of cognitive science.

In order to see why the relation between mind wandering and intentionality is controversial requires that we examine the position that a wandering mind occupies in a standard hierarchical model of attention. In such a model of attention, basic information processes which are more or less directly related to the processing of stimulus information are seen as occupying a relatively low-level position in the hierarchy. An example of a low-level attentional process is the detection of luminance or color information by the visual system. Low-level processes are generally described as influencing the control of attention through a bottom-up process.

Low-level cognitive processes are often contrasted with higher-level cognitive processes which operate on more abstract information. Such abstract information is usually the product of computations performed on low-level sensory information, or alternatively is retrieved from memory systems. The influence of higher-order cognition is described as top-down. Examples of higher-order cognition would involve planning (anticipating future events) or metacognition (reflecting upon particular thoughts). In standard models of attention, higher-level cognitive structures provide the opportunity for the top-down control of behavior. Top-down control allow individuals to strategically modulate the manner in which they respond in the current environment in order to achieve a goal.

When mind wandering is conceived of in terms of the top-down/bottom-up dichotomy it is clear that the phenomenon is somewhat paradoxical. On the one hand, the complex information-processing requirements to introspect in a goal-motivated manner suggests a top-down process, however, the strong sense that we do not intend our minds to wander suggests an absence of top-down control. The puzzling nature of control during mind wandering is satirized by comedian Steven Wright when he jokes, "I was trying to daydream but my mind kept wandering."

Ironic Processes and the Wandering Mind

One explanation for why the mind wanders is suggested by the theory of ironic processes developed by Dan Wegner. Wegner argues that the experience of mind wandering occurs because of the manner with which control is instantiated by the brain. According to Wegner, mental control is instantiated by the cooperation of two systems which combine to produce the control we have over our mental life. The first system – termed the 'intentional monitor' – operates in much the same way that top-down control was described in the previous section. The intentional operator allows us to strategically constrain our attention to a particular topic or activity, and is associated

with the palpable sense of effort when we focus our own thoughts. Wegner's insight was that a system such as the intentional operator could become preoccupied when engaged in a demanding activity. In order to keep the intentional system on track, Wegner suggested that we have a second monitoring system – termed the 'ironic monitor' – which is responsible for ensuring that the intentional monitor does not get side tracked. According to Wegner, it is the interactions of these two systems that lead the mind to wander.

One implication of the theory of ironic processes is that if the intentional monitor was engaged on a demanding task, the ironic monitor would make a greater contribution to the contents of consciousness. To do so, Wegner and others developed what is known as the thought suppression or 'white bear' paradigm. In these studies, participants are asked to suppress thoughts of a particular topic that is not present in the immediate environment – often a 'white bear' – and are either asked to perform a demanding secondary task or not. Participants are asked to indicate (often by ringing a bell) whenever they experience a target thought. Results of this paradigm are quite striking – generally when participants are asked to suppress the target thoughts and are placed under load they report the most intrusions of the unwanted thought. This ironic increase in unwanted thoughts is consistent with Wegners' model because it depends upon (1) the ironic monitor being directed to the unwanted thought by instruction to suppress the thought and (2) the intentional monitor being occupied by the demanding secondary task. The thought suppression paradigm, and related studies provide important support for the notion that the experience of mind wandering is in fact compelled by the manner in which cognitive control in instantiated by the mind.

Awareness and the Recognition of a Wandering Mind

A second possible explanation for why mind wandering occurs has been proposed by Jonathan Schooler. He argues that in addition to the normal division of conscious processes into explicit and implicit, we are also intermittently able to represent the contents of awareness – a concept he has referred to as meta-awareness or metaconsciousness. Mind wandering during reading illustrates a situation when we are aware of the content of awareness (the off topic thoughts) and yet are unaware that we are currently engaging in thoughts which potentially undermine our ability to comprehend what we are reading. According to Schooler, therefore, mind wandering occurs because of the difficulties we have in continually representing the contents of consciousness. On the flip side, catching ourselves mind wandering occurs because we suddenly take stock, or re-represent the contents of awareness, and realize that for some time we were engaging in thoughts which were unrelated to the task of reading. One clear advantage of the notion of meta-awareness is that it captures the compelling sense that in many mind-wandering episodes we fail to recognize for several moments that we were off task. According to Schooler, therefore, the ebb and flow of our minds' wandering occurs in part because we are unable – or unwilling – to continually represent the content of our own experiences.

Three Routes to a Wandering Mind

So far this article has considered a simple information-processing account of information flow which allows mind wandering and other mental states to be clearly conceptualized and has described two of the most compelling explanations for why the mind should wander in the first place. In the next section, this article will consider studies which have examined the influences that lead the mind to wander. Figure 2 presents a schematic diagram of different attentional configurations which could be conceivably be associated with greater likelihood of engaging in mind wandering. Each of these different configurations can be broadly considered as a different route that leads the mind to wander.

The first route to a wandering mind is when the external environment does not exert a sufficient influence on working memory to necessitate that task-irrelevant representations are suppressed. The lack of a demanding external environment is represented in Figure 2 by the relative absence of downward arrows in the left-hand panel. The second

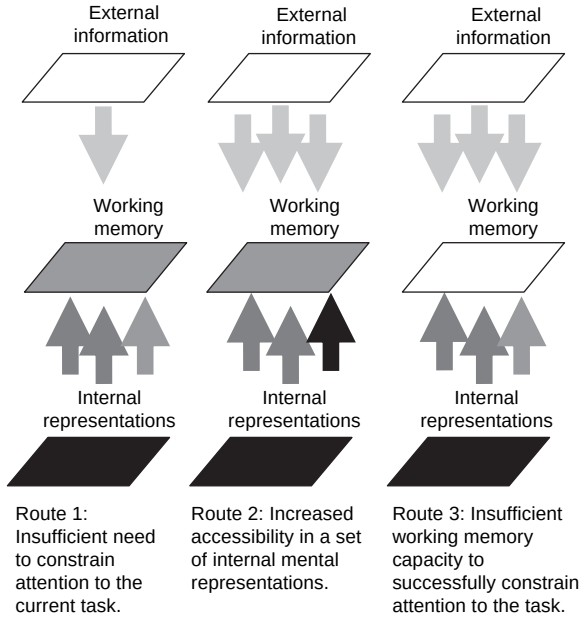


Figure 2 The three routes to a wandering mind. Route one – the external task does not require continuous monitoring and idle working memory resources are available to be co-opted into mind wandering. Route two – an increase in the accessibility of a salient internal mental representation attracts resources away from the outside world. Route three – failures in attentional control allows distractions from the internal or external environment access to working memory.

route to the wandering mind occurs when our attention is drawn to a particularly salient thought. The increase in salience is represented by the single black arrow in the central panel. The third and final route to a wandering mind is when the individual lacks the necessary attentional constraint in order to maintain attention on the task. The inability to constrain attentional resources is represented by the shading in the rectangle in the right-hand panel. In the next three sections, this article will consider experimental evidence that each route can lead to a greater incidence of a wandering mind.

Route One – The Influence of the Task Environment

A tourist, particularly one who is British, can find driving around a foreign city one of the more stressful aspects of their holiday – not only must they attend to the hectic alien traffic system, such as driving on the right-hand side of the road, they must also determine the correct route to take them to their destination. By contrast, the commuter who makes the journey home from work has a different experience. Unlike the tourist, a commuter is familiar with the traffic systems, and because they have a detailed knowledge of the spatial layout of their hometown they need give few thoughts to the task of navigation. As a consequence the commuter has the time and resources to explicitly consider information unrelated to the current circumstance or may instead simply let their minds wander. The tourist does not have such a luxury – their attention is likely to be entirely wrapped up in the task of getting to their destination. This anecdote illustrates one of the well-documented routes to a wandering mind – our minds wander most frequently when the current task does not require that we constrain our attention to task-relevant material.

One way to investigate the effect of task environment is to examine the effect that the rate of events occurring in the task has on the amount of mind wandering. Work conducted in the 1960s by John Antrobus and Jerome Singer demonstrated a clear linear relationship between the frequency of mind wandering and the rate of signal presentation during sustained attention. In these studies, participants were asked to perform a signal-detection task and were periodically asked to report the amount of mind wandering they were engaging in. In these studies, the rate of presentation was varied, and the results showed that slower presentation rates tended to lead to greater mind wandering than did fast rates. The suppression of mind wandering by frequent events is perhaps why when driving in a city we seem to maintain better attention on what we are doing than when driving on an empty highway were incidences of mind wandering seem much more frequent.

In the anecdote described at the start of this section, it is the need to continuously monitor the task that distinguishes the commuter from the tourist. As both commuter and tourist are driving the same route they are engaged in the same task – it is the locals' familiarity with both the traffic system and their knowledge of the route which allows them the opportunity to let their

minds wander. Research by John Teasdale elegantly demonstrated the fact that tasks requiring continuous monitoring tend to suppress the mind's tendency to wander. In a series of studies, participants were shown a series of digits and were asked to either say the number out loud immediately, or to hold it in mind for a short period (c. 2 s). In both task participants were also asked to report the amount of mind wandering they were engaged in. Even though these two tasks are trivially easy, the simple fact that individuals had to hold the digits in working memory for a short period significantly reduced the amount of mind wandering experienced.

The simple involvement of working memory, however, is not the whole story with respect to mind wandering. A quick review of the last time we were absorbed in a good book or a film indicates that often our mind wanders 'least' when the primary task interests or intrigues us. Instead, in the context of tasks such as reading, our experience is more likely to be driven by features of the task which interest us rather than those which are simply demanding to process. Leonard Giambra and others compared the influence of interest and difficulty in reading. In this study participants read a number of texts – preselected for either interest or difficulty during which mind wandering was measured. Their results demonstrated that mind wandering during reading was predicted by interest in text but not difficulty. Presumably, participants find it more absorbing to read an interesting story and as a result they are better able to maintain continuous focus on what they are doing. This finding concurs with our own everyday experience that getting lost in a good novel or film is possibly the best way to temporarily reduce the mind proclivity to wander.

Route Two – The Influence of Current Concerns

The second route to a wandering mind is when our thoughts get attracted to a particularly salient topic. A review of our mind-wandering experiences suggest that we spend a great proportion of time thinking about things which are generally personally relevant to us as an individual. Eric Klinger suggested the term 'current concerns' to reflect that much mind wandering generally has this personally relevant focus.

Several studies have explored how the salience of particular concerns are related to the frequency of mind wandering. In the 1960s during the Vietnam War, Antrobus and others recruited a cohort of individuals to perform a laboratory task in which mind wandering was measured. While the participants were waiting to perform the task, they either heard a fake news broadcast that China had entered the Vietnam War or a control broadcast. During the subsequent attentional task, the erroneous information that China had entered the Vietnam War led to an increase in the frequency of mind wandering and an increase in signal-detection errors. At debriefing, participants reported experiencing reasonably distressing thoughts associated with their concern about the future. This study clearly indicates that the second route to the wandering mind is when a particularly salient concern overshadows performance on the task in hand.

An alternate way to investigate whether mind wandering is associated with particularly salient cognition is to compare populations which differ on the salience of their current concerns. One population that tends to have particularly salient concerns are depressed individuals. Since the 1960s it has been well known that a significant contributing factor to depressed mood are current levels of stress such as daily hassles. Recently, in conjunction with a number of my collaborators I explored whether mind wandering really is associated with particularly pressing concerns. To do so we recruited a cohort of participants and in addition to completing a measure of depression asked them to perform a word-learning task in the laboratory during which we measured mind wandering. While they performed the cognitive tasks we also measured autonomic function (heart rate and the skin conductance level). Results indicated a number of interesting things. First, the ability to remember the words presented was significantly impaired when individuals were off task. Second, the size of the skin conductance response was significantly higher during periods when mind wandering occurred. Finally, participants who were particularly unhappy experienced greater increases in physiological arousal (particularly their heart rate) when their minds wandered.

These data are consistent with the notion that mind-wandering episodes are focused on salient and so arousing information because of the associated physiological responses and so supports the importance of an individuals 'current concerns' in leading the mind to wander.

Route Three – A Lack of Constraint

When we notice that we are mind wandering, as we all frequently do, this act is no mean feat. Among other things catching our minds wandering requires that, first, we can reflect upon the contents of consciousness and, second, assess whether we are still on course to achieve our current goal. Moreover once a mind-wandering episode is identified, the correctly motivated individual will set aside that particular thought and return their attention to what they should have been doing. The ability to catch one's own mind wandering and correct this attentional problem requires sophisticated metacognitive or attentional control. Unlike the previous two routes, the third route to a wandering mind is through difficulties in metacognitive or attentional control which could lead to a general increase in distractibility and so an increase in mind wandering.

Recent studies have confirmed that the ability to control attention impacts upon the ability to control mind wandering. Michael Kane and others explored the relationship between the ability to control attention – often referred to as working memory capacity – and the tendency to engage in mind wandering in every day life. To do so Kane and other recruited a large number of students who performed a number of laboratory tasks measuring their ability to maintain and control information in working memory. After this the participants were given pagers which they carried for a number of days which periodically beeped at which the participants were asked to provide detailed accounts of the content of their thoughts and the context within which it occurred. The results documented an interesting relationship between mind wandering and working memory capacity. Under challenging circumstances individuals who had a higher working memory capacity tended to mind wander less, than at other times. Individuals with a lower working memory capacity, however, tended to mind wander regardless of whether the task was easy or difficult, suggesting that the poorer attentional control prevented these individuals from suppressing mind wandering if and when the situation demanded they do so. Thus, the ability to control the contents of working memory plays an important role in controlling the wandering mind.

The Consequences of the Absent Mind

The previous section of this article considered three different routes that lead the mind to wander. The mind's tendency to wander, however, can lead to disastrous consequences for the integrity of the task in hand. This research covers a wide range of tasks – from simple signal-detection tasks through to more complex tasks such as reading – and so provides an illustration of the far reaching consequences that mind wandering and the associated lapses can have on our daily lives.

Mindless Reading

Throughout this article we have used the example of mind wandering during reading as a simple example of the phenomenon. Reading is a complex task in which the reader must continuously integrate from what they are reading into a larger model of the narrative. Common experience informs us that it is a relatively common experience to engage in mind wandering during reading, and when it occurs we often have to reread whole sections of the text as we have no clear idea of what happened in the narrative. This leads to the clear contention that mind wandering during reading could have implications for what is being read.

To explore exactly what the consequences of mind wandering during reading were, we examined the occurrence of off task thoughts while participants read a detective novel. We chose a detective novel because it was an interesting story but one in which the reader must develop a detailed model of the narrative in order to solve the crime. We reasoned that if mind wandering prevents participants from making sense of what they were

reading it would ultimately prevent them from building a model of the story and so would prevent them from solving the crime. To examine this question, we measured mind wandering at both random intervals in the story and at points when the author delivered a clue which if correctly attended would help the reader solve the crime. At the end of the task participants were asked to answer several questions – one of which required that they had solved a critical aspect of the mystery. Consistent with expectations, analysis indicated that the ability of the individual to solve the crime depended on whether they were on task at the critical parts of the novel. Interestingly, only reports of mind wandering without awareness prevented the reader from solving the crime. By contrast, general mind-wandering propensity (e.g., being off task at random intervals in the text) did not determine whether the individuals solved the crime. The absence of a relation between overall mind-wandering propensity and the participants' ability to solve the crime ruled out alternative explanations such as attentional control as being the main contributor to the study's results. This study clearly indicates that the effects of mind wandering during reading is to interfere with the ability of the individual to create a model of what is being read and that one implication of the failure is that individuals are unable to make the inferences necessary for reading. Given that the majority of education proceeds through reading, one implication of this finding is that mind wandering could well be an under-recognized influence on the educational achievement of both children and adults alike.

Action Slips

The consequences of mind wandering are not limited to tasks that are cognitively demanding it can also lead to error in tasks which at first glance seem trivially easy. Often task in such safety critical situations are largely automated and rely on humans as a watch keeper rather than as the direct controller of the action. While this automation has largely removed direct human error, an indirect consequence of the increased automation is that it provides the ideal environment to encourage the wandering mind. In some contexts, the simple and routine nature of the task makes mind wandering one of the few ways that humans can still contribute to error. Train and automobile driving, provide clear examples of long monotonous tasks in which even momentary inattention to the task in hand can have calamitous consequences.

Errors which occur in seemingly simple tasks are called action slips. Action slips occur whenever a participant makes an error in a routine action sequence that under normal circumstances can be easily performed. Action slips, or, skill-based slips were first studied in the 1980s by the psychologists Jim Reason and Donald Broadbent. From the perspective of an external observer action slips arise from the inattention of the individual rather than because of the difficulty associated with the task. As such action slips arise in the same mysterious manner as does mind wandering – the event is determined by an internal change and not an external event. Initially, this research employed diary studies to assess the frequency that lapses occurred with reasonable frequency in everyday life, although they were hampered by their inability to study the phenomenon experimentally.

In the late 1990s, the study of action slips was moved into the cognitive laboratory by the work of Ian Robertson, Tom Manly, and others. In order to mimic the relatively undemanding nature of many everyday tasks, these researchers examined the performance of individuals on a simple routine task such as withholding a response to a familiar stimulus (often the digit '3') presented in a long sequence of nontargets. This task is deceptive – the vigilance component of the task is minimal as the stimulus is easy to detect. The requirement to withhold a response to an infrequent target, however, requires a surprisingly demanding form of attentional control. Because the target stimulus occurs very infrequently, it is tempting for the individual to perform the task in a rapid stimulus driven manner, or in more familiar terms on 'automatic pilot.' In these circumstances when a target arrives the individual has only a short interval to intervene before the automatic tendency to respond to the target takes over and an error ensues. Even though the detection of a target is trivially easy in the SART, one needs to attend to one's performance in a more or less continuous manner in order to perform the task effectively. This

emphasis on one's own behavior has led to the task being described as the Sustained Attention to Response Task and is often shortened to the acronym the SART. The SART mimics the requirement that occurs in many safety conscious industries because it confronts the individual with a dull task environment and requires that they maintain almost exclusive attention on their performance.

It is clear that the concept of action slips is closely linked to the more subjective aspects of mind wandering – both represent situations in which attention is not fully constrained on what is being done. While there is an obvious similarity between the notion of action slips and mind wandering, just how closely related are these different states? This is the question that Dan Smilek and others set out to answer. They recruited a sample of individuals who completed separate measures of action slips and mind wandering in everyday life and also completed the SART. The results illustrated an interesting dissociation – both measures of action slips and mind wandering in everyday life made unique contributions to performance on the SART. Measures of mind wandering were associated with how fast participants responded to nontargets, while the likelihood of action slips was strongly associated with the amount of errors made on the task. This research clearly documents that mind wandering and action slips are closely related yet experimentally dissociable aspects of conscious experience and leaves open the intriguing question about the aspects of our mental life that lead mind-wandering episodes to become an error.

Imaging the Wandering Mind

As with many areas of psychology, the detailed measurements provided by neuroimaging has helped reveal many fascinating insights into how mind wandering occurs in the waking brain. In fact a key problem associated with mind-wandering research is that the processes which occur are private and so invisible to direct observation and in such circumstances the ability of neuroimaging to reveal the processes which take place in the waking brain is a very useful skill. In this article, we focus on three neuroimaging studies that reveal three key aspects of mind wandering: (1) the extent of cortical processing on the task it is performing while the mind wanders, (2) the neural substrates that produce the introspective content during mind wandering, and (3) the processes which are used to keep the wandering mind in check.

The Absent Task

One advantage of neuroimaging is that it makes it possible to ascertain the depth of processing that an individual is engaged by measuring the response of the brain to events in the task. These evoked responses are referred to as event-related potentials (ERP) and are calculated by measuring the electrical signals generated by the brain using electrodes placed on the scalp. These signals are averaged across different conditions and can be analyzed to estimate the amount of processing that is being deployed to a given stimulus under different task conditions.

In a simple account of mind wandering, we assume that when the mind wanders participants are less aware of the task at hand than when they are focused on what they are doing. In collaboration with Todd Handy and others, we tested whether periods of mind wandering really were associated with less detailed processing of the task than periods of task focus. To do so we recruited a number of volunteers who performed the SART and measured how their brains responded to events in the task when they were mind wandering. We were interested in a component of the event-related potential, known as the P3 which occurs c. 300 ms after stimulus onset and is conceptualized as indicating the amount of effort that is being deployed on the task. We reasoned that if participants really were not giving the task their full attention during mind wandering they should show a smaller P3 in periods when they were off task. The results documented that both reports of mind-wandering and errors on the SART were associated with a reduction in the amplitude of the p3 component of the ERP. Interestingly, further analysis indicated that reports of mind wandering without awareness tended to

be most closely linked to errors, suggesting that failing to recognize that our mind has wandered could help explain why lapses occur.

The Default Network and the Wandering Mind

One of the most controversial findings in neuroimaging over the last several decades is that when participants rest in a scanner, a large number of cortical and subcortical regions are more active than during many demanding experimental tasks. One influential view of this result – described as the default mode hypothesis by Marcus Raichle and others – is that the activity in these structures in periods when participants were ostensibly resting in the scanner reflects a psychological baseline that participants adopt when their task does not require their undivided attention. One crucial feature of the default mode hypothesis is that it predicts that many of the cortical and subcortical structures may play a role in delivering the introspective content during states of mind wandering.

Recent research conducted by Malia Mason has provided important support for the assumption that the default network does play a role in the introspective content when the mind wanders. In a longitudinal study, Mason and others trained participants to perform the same version of a working memory task in three half hour sessions over consecutive days. On the fourth session, participants performed either the same well-practiced version, or a matched novel version. During this session the authors employed thought sampling to estimate how much mind wandering was occurring – participants reported more off task thoughts in practiced than novel blocks. On a fifth session, participants performed both novel and practiced sessions in a functional magnetic resonance imaging (fMRI) scanner. Finally, participants completed a measure of their tendency to daydream. The results indicated that during scanning, a number of structures in the default network were more active in the practiced than nonpracticed blocks. Moreover, the size of activity in several default structures was positively associated with the individual propensity for daydreaming measured via questionnaire. Taken together, the relative activation of default network structures in the practiced blocks of the task, and their association with daydreaming tendency support that assumption that the default network plays a role in mind wandering.

Constraining the Wandering Mind

Given the studies described in this article so far, it seems somewhat miraculous that we are able to engage in any task without continually losing track of what we are doing, and yet despite our minds tendency to wander, as a rule humans are able to successfully navigate their way through many complex tasks environments. It is clear, therefore, that evolution has provided humans with the apparatus to control our minds tendency to wander – but how do we do this? One possibility is that the mind employs the same neural structures to constrain mind wandering as it does to resolve conflicts which arise when we perform other tasks. Models of cognitive control suggest that two neural structures are involved in conflict management – aspects of the lateral prefrontal cortex which are thought to be involved in the strategic regulation of effort to a given task in a sustained manner and the anterior cingulate which is thought to be involved in identifying and responding to transient changes in the conflict when the need arises.

To explore whether the same conflict management processes are involved in controlling mind wandering, Jason Mitchell and colleagues asked participants to perform a thought suppression task in an fMRI scanner. Participants were asked to either suppress or report the experience of thoughts regarding a white bear. The results indicated that relative to the report condition, periods of thought suppression did indeed mobilize lateral prefrontal structures indicating that the brain employed the same control structures to suppress the experience of unwanted thoughts. Similarly, following the experience of unwanted thoughts the anterior cingulate was more active indicating that this structure was implicated in the response to failures in mental control. Taken together, this result clearly indicates that one reason that we are able to refrain from mind wandering in general is because we are able to employ existing executive control strategies to regulate our mental lives.

Future Directions in Mind-Wandering Research

It is an exciting time to be a researcher studying mind wandering. Recent theoretical and experimental developments have shed important light on why the mind wanders and the consequences that can occur when it does. Meanwhile, advance in neuroscience are beginning to reveal in an unprecedented fashion the neural structures that support and control the introspective world that plays such a pivotal role in our mental life. In the future, we are likely to gain important insights into this remarkable feature of the human mind.

One area of future research is likely to be the potential functional basis that mind wandering could play in our daily lives. One important human skill is the ability to solve tasks which require an indirect or counter intuitive solution. Often these sorts of cognitive processes are referred to as insight problems. On the face of it, mind wandering shares a number of basic similarities with insight problem solving. Both insight problem solution and prospective memory share with mind wandering the requirement that the mind is not to tightly constrained to the current task. The relationship between mind wandering and creativity has been relatively overlooked in much of the work on mind wandering, as most of the work has focused on the negative aspects of this concept – as typified by the title of this article. Alternatively, mind wandering could share many features with what researchers refer to as prospective memory – our mind's ability to remember future goals. Studying whether mind wandering is of functional significance to the individual could be important in the future in detailing the function of an experience which makes up such a large component of our waking life.

Another important question facing mind-wandering researchers is the development and validation of an indirect marker a so called mind wandering meter. While this question was almost impossible to imagine several decades ago, in practical terms technological advances in neuroimaging techniques have made it a viable possibility. Such a discovery would revolutionize the study of mind wandering because it would make it possible to identify the experience of mind wandering without interrupting the experience. If successful, this would provide the opportunity for investigators to begin to identify in an online manner the occurence of the wandering mind. Because this measure would be entirely independent of the self-reports of individuals themselves, the extent to which theory predicted fluctuations in this index of mind wandering would provide the strongest source of data to date to approach question such as why the mind wanders.

The lasting message of research on mind wandering, however, is that it reminds us of the essential flux which is such a fundamental property of our mental lives. It is this ebb and flow of attention, so characteristic of processes such as mind wandering, which motivated William James to use the term 'the stream of consciousness' as a metaphor. Rather than assuming that the lack of control over the phenomenon is a limitation, recent success in understanding the theoretical and neural aspects of mind wandering reminds us of the importance that the detailed empirical observation of naturally occurring phenomenon plays in our more general understanding of the mind. It is only by adopting a less constrictive experimental approach than cognitive experimental psychology have favored in the past, that we can shed light on the ebb and flow of attention that leads to states like mind wandering.

See also: Attention: Selective Attention and Consciousness; Automaticity and Consciousness; Neglect and Balint's Syndrome; Philosophical Accounts of Self-Awareness and Introspection.

Suggested Readings

Reviews

Antrobus JS (1999) Toward a neurocognitive processing model of imaginal thought. In: Salovey P and Singer J (eds.) At Play in the Fields of Consciousness: Essay in Honor of Jerome L. Singer, pp. 1–28. Hillsdale, NJ: Erlbaum.

Klinger EC (1999) Thought flow: Properties and mechanisms underlying shifts in content. In: Singer JA and Salovey P (eds.) At Play in the Fields of Consciousness: Essays in the Honour of Jerome L Singer, pp. 29–50. Mahwah, NJ: Lawrence Erlbaum.

Smallwood J and Schooler JW (2006) The restless mind. Psychological Bulletin 132(6): 946–958.

Smallwood J, Fishman DF, and Schooler JW (2007) Counting the cost of the absent mind. Psychonomic Bulletin & Review 14: 230–236.

Robertson IH and Garavan H (2004) Vigilant attention. In: Gazzaniga MS (ed.) The Cognitive Neurosciences, 3rd edn., pp. 563–578. MIT Press.

Wegner DM (1997) Why the mind wanders. In: Cohen JD and Schooler JW (eds.) Scientific Approaches to Consciousness, pp. 295–315. Mahwah, NJ: Erlbaum.

Empirical Papers

Cheyene JA, Carriere JSA, and Smilek D (2006) Absent-mindedness: Lapses of conscious awareness and everyday cognitive failures. Consciousness and Cognition 3: 578–592.

Kane MJ, Brown LH, McVay JC, et al. (2007) For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. Psychological Science 18(7): 614–621.

Mason MF, Norton MI, Van JD, et al. (2007) Wandering minds: The default network and stimulus independent thought. Science 315: 393–395.

Mitchell JP, Heatherton TF, Kelley WM, et al. (2007) Separating the transient and sustained aspects of cognitive control during thought suppression. Psychological Science 18(4): 292–297.

Raichle, et al. (2001) A default mode for brain function. Proceeding of the National Academy of Sciences 98(2): 676–682.

Smallwood J, O'Connor RC, Sudberry MV, et al. (2007) Mind wandering & dysphoria. Cognition & Emotion 21(4): 816–842.

Smallwood J, Beech E, Schooler JW, et al. (2008) Going AWOL in the brain: Mind wandering reduces the cortical analysis of the task environment. Journal of Cognitive Neuroscience 20: 458–469.

Teasdale JD, Proctor L, Lloyd CA, Baddeley AD, et al. (1993) Working memory and stimulus-independent-thought: Effects of memory load and presentation rate. European Journal of Psychology 5: 417–433.

Biographical Sketch

Jonathan Smallwood has had a long-term research interest in the processes which contribute to the experience of thoughts which are not derived from the immediate environment. He began studying this mind wandering in 1996 when he studied for his PhD at the University of Strathclyde in Glasgow, Scotland. After completing his thesis he worked as a research associate in several universities in the Glasgow area, before moving to the University of British Columbia, Vancouver, Canada to complete a postdoctorial with Jonathan Schooler. Currently he lives in the USA and works as a research scientist at the University of Santa Barbara, California. To date he has published more than 30 scholarly articles, 20 of which are on the topic of mind wandering.

Neglect and Balint's Syndrome

P D L Howe, Brigham and Women's Hospital, Harvard Medical School, Cambridge, MA, USA

Glossary

Anosognosia – A neurological disorder, usually caused by brain injury, characterized by the sufferers denying they have a specific impairment, such as hemiplegia.
Balint's syndrome – A neurological disorder characterized by simultanagnosia and optic ataxia, typically caused by bilateral damage to the posterior parietal cortex.
Extinction – A neurological disorder where the ability to attend to stimuli in a particular region of the visual field is impaired by the presence of stimuli outside that region.
Optic ataxia – The inability to guide the hand toward an object using visual information that cannot be explained by motor or sensory deficits.
Simultanagnosia – The inability to perceive more than one object at a time.
Somatoparaphrenia – The belief that part of one's body belongs to another person.

Introduction

Neglect, also known as unilateral neglect, hemispatial neglect, hemineglect, neglect syndrome, or spatial neglect, is a disabling brain disorder, usually caused by a brain lesion. Patients with neglect typically have difficulty paying attention to contralesional space. Thus, if the lesion causing the neglect occurs in the right cerebral cortex, the patient may find it difficult to attend to an object situated to the left of their current point of fixation. Neglect is most often caused by damage to the right cerebral cortex, but can be caused by damage to the left cerebral cortex, though the resultant neglect is usually less severe. Because its symptoms can vary from individual to individual and because it can be caused by a lesion in any of several different locations in the brain, it may not be a single disorder but rather a family of several related disorders. It is especially likely to be caused by lesions near the junction of the temporal lobe and the inferior parietal lobule, but can also be caused by lesions in the frontal cortex, in the white matter, or in subcortical areas such as the thalamus and basal ganglia.

Neglect is necessarily oriented relative to a specific frame of reference. For example, to say that a patient has left-sided neglect one must first specify what frame of reference is used to define the term 'left.' Neglect oriented relative to the patient is referred to as egocentric. Egocentric neglect is most often orientated relative to the patient's point of fixation. For example, a patient with left-sided egocentric neglect will typically ignore objects to the left of their point of fixation. However, egocentric neglect may also be orientated relative to a part of the patient's body. For example, a patient may ignore objects to the left of their left hand. Even when oriented primarily relative to the patient's point of fixation, the degree of egocentric neglect may be influenced by the orientation of parts of the patient's body, such as the head and trunk. For example, an object that is to the left relative to the point of fixation, but to the right relative to the head and to the right relative to the trunk, may be less neglected than an object that is to the left in all three coordinate systems. Although left-sided and right-sided neglect are the most common, other forms of egocentric neglect are possible. For example, a patient may have upper visual hemifield neglect, and so ignore objects that are situated above their point of fixation.

Neglect is not always oriented relative to the patient and may instead be oriented relative to the object that is being viewed, in which case it is referred to as allocentric. Unlike a patient with left-side egocentric neglect, a patient with left-sided allocentric neglect may be able to attend to

objects on their left, but will ignore the left half of each object, regardless of where the object is located. Curiously, if the object has a well-defined left side, a patient with left-sided allocentric neglect may continue to ignore the left side of the object even when the object is rotated by 180 , so that its left side is now situated on the right. For example, when viewing a face with a small blemish on the left side, a patient with left-sided allocentric neglect may continue to ignore the side of the face with the blemish even when the face is rotated by 180 . It seems that when presented with a face, the patient mentally rotates the face until it is orientated in the standard fashion (i.e., the eyes above the mouth) and then ignores the left side.

Although neglect can be so extreme that the patient fails to notice large objects on the neglected side, it need not be total. Instead, it may manifest itself only as a tendency not to respond to stimuli in the neglected region. If neglect is not total, the addition of stimuli to the nonneglected side may further decrease the ability to attend to stimuli on the neglected side, a phenomenon known as extinction. As extinction and neglect can occur independently of each other, they may be distinct disorders. For example, a patient may exhibit left-sided extinction without exhibiting left-sided neglect. Such a patient would have no difficulties attending to an isolated object presented on their left. However, introducing objects on their right, might make it hard (or even impossible) for the patient to continue to attend to the object on the left. Conversely, a patient might exhibit neglect without exhibiting any extinction, so the patient's ability to attend to an object on the neglected side would not be influenced by whether or not objects are presented on the nonneglected side.

Forms of Neglect

Neglect most commonly occurs in the visual domain, in which case it may result in some or all of the following symptoms: Patients may shave or apply makeup to only one side of their face. They may eat from only one side of a plate. When moving in a wheelchair, they may bump into objects on the neglected side. If asked to bisect a line that crosses the visual midline, they may be biased in the nonneglected direction. When copying a picture, they may have a tendency to copy only the nonneglected side. When presented with an image, they may look mainly (or exclusively) at the nonneglected side. Rapid eye movements (REM) to the nonneglected side may occur in REM sleep. If asked to circle all occurrences of a specific letter, they may concentrate on the occurrences that appear on the nonneglected side, circling them repeatedly to the near, or even total, exclusion of those that appear on the neglected side.

Neglect can occur in other sensory domains such as the auditory, olfactory, or somatosensory domains. If it occurs in the somatosensory domain the patient will ignore a region of her body and may even deny ownership of neglected limbs, sometimes believing that they belong to someone else (somatoparaphrenia). Often the neglected area is also paralyzed. In such cases, the patient may be unaware of or deny the paralysis (anosognosia). Neglect can also affect motor responses. Although the patient may have no physical impairments, the patient might have difficulty initiating movement or their movement may be slow.

Although comparatively rare, a patient may exhibit representational neglect and ignore a portion of an imagined scene. For example, when patients with left-sided representational neglect were asked by Bisiach and Luzzatti to imagine viewing the Piazza del Duomo in Milan standing next to the cathedral in the piazza, the patients often failed to mention streets or places on the left side of the piazza. However, when asked to imagine looking directly at the cathedral, so that they imagined viewing the same scene as before, but from the opposite direction, the patients then recalled the objects and places they had previously failed to mention, since these objects were then situated on the right side of the piazza, relative to their new viewpoint.

Processing of Neglected Stimuli

Neglect is not caused by a disruption of the visual system *per se*, but by a disruption of the cortical system that deploys attention. The only reason

why a patient with neglect fails to consciously perceive stimuli located in their neglected region is because the patient cannot attend to them. If patients are prevented from seeing because of an abnormality in their visual system such as in their eyes, optic nerves, lateral geniculate nuclei, or visual cortex, they are said to exhibit blindness, not neglect. To diagnose a patient as having neglect, one must first rule out any other reason why the patient might not be able to see.

Some stimuli can be perceived even when they are not attended, so can be readily seen by a patient with neglect. For example, if a bright spot of light is presented in complete isolation on a uniform black background, most neglect patients will be able to detect it, regardless of where it is located, especially if it is flashed repeatedly.

There is evidence that quite sophisticated processing can occur in the neglected region. For example, if a word is presented on the neglected side, even when it is not consciously seen, it may cause the patient to respond more quickly to similar words presented on the nonneglected side. Similarly, if a patient is simultaneously presented pictures of two different houses and asked which would be better to live in, the patient might reliably choose the house that is not on fire, even though the flames appear only in the neglected hemifield, so are not consciously perceived. When asked to explain their choice, the patient will not be able to do so and will often confabulate. Other studies have asked patients to compare two simultaneously presented pictures and report whether they are the same or different. Patients could do this task even when one or both of the pictures were presented to the neglected hemifield.

Balint's Syndrome

Balint's syndrome is a brain disorder, closely related to neglect, first reported by Reszo Balint in 1909. Whereas neglect is caused by unilateral damage, Balint's syndrome is caused by bilateral damage, typically to the posterior parietal cortex. Unlike a patient with neglect, who is unable to attend to objects in a particular region of the visual field, a patient with Balint's syndrome will be able to attend to an isolated object, regardless of where it is located. However, the patient will find it difficult to point to the object (optic ataxia) or to perceive more than one object at a time (simultanagnosia). While such patients can often perceive the features in a scene, they have difficulty determining which features belong to which object, an issue known as the binding problem. As a result, they have a tendency to conjoin features that belong to different objects and so perceive illusory conjunctions. For example, if a scene contains only red vertical bars and blue horizontal bars, a patient with Balint's syndrome might perceive a blue vertical bar.

Conclusion

Studies on neglect indicate that there is interhemispherical competition between the cortical circuits that control the deployment of attention. For example, damage to the right cerebral cortex may allow the attentional circuits in the left cerebral cortex to dominate. As the left cerebral cortex is responsible for processing the right visual hemifield, attention is directed more often (or even exclusively) to the right visual hemifield, causing the patient to ignore objects that occur to the left of the point of fixation. Similarly, damage to the left cerebral cortex may cause the patient to ignore objects that occur to the right of the point of fixation. If both cerebral hemispheres are damaged, then neither can dominate, so a patient with Balint's syndrome can perceive an isolated object regardless of where it is located. However, due to the damage to their attentional circuits, the patient may be able to perceive only one object at a time.

Several studies have shown that an object located in the neglected region can be represented even when not consciously seen. How this occurs is controversial. A possible explanation is that the attention required for conscious awareness might be different from that required to form object representations, and in neglect, only the first type of attention is inhibited.

See also: Attention: Change Blindness and Inattentional Blindness; Psychopathology and Consciousness.

Suggested Readings

Bartolomeo P (2007) Visual neglect. Current Opinion Neurology 20: 381–386.
Bartolomeo P, Thiebaut de Schotten M, and Doricchi F (2007) Left unilateral neglect as a disconnection syndrome. Cerebral Cortex 17: 2479–2490.
Buxbaum LJ (2006) On the right (and left) track: Twenty years of progress in studying hemispatial neglect. Cognitive Neuropsychology 23: 184–201.
Driver J and Vuilleumier P (2001) Perceptual awareness and its loss in unilateral neglect and extinction. Cognition 79: 39–88.
Husain M and Nachev P (2007) Space and the parietal cortex. Trends in Cognitive Science 11: 30–36.
Milner AD, Karnath H-O, and Vallar G (2002) The Cognitive and Neural Bases of Spatial Neglect. Oxford: Oxford University Press.

Biographical Sketch

Piers Howe graduated as an exhibitioner from Oxford University in 1998, with a masters in physics. Winning a Presidential University graduate fellowship, he obtained his PhD from Boston University in 2003, under the guidance of Stephen Grossberg. His PhD thesis was titled 'Cortical mechanisms of depth and lightness perception: Neural models and psychophysical experiments.' He then worked as a Helen Hay Whitney postdoctoral fellow with Margaret Livingstone at Harvard Medical School, before moving on to Brigham and Women's Hospital to work as a research fellow with Todd Horowitz and Jeremy Wolfe. His research has involved a variety of techniques including computational modeling, macaque neurophysiology, human fMRI, and human behavioral experiments. He has published articles on lightness perception, motion perception, depth perception, and visibility. More recently his focus has shifted to visual attention and to devising computational techniques for determining brain connectivity from fMRI data. He is a member of the Harvard fMRI Center for Neurodegeneration and Repair. He has taught a psychology course at the University of Massachusetts (Boston) and at a cognitive modeling course at Boston University. He lives in Boston, Massachusetts.

The Neural Basis of Perceptual Awareness

J-D Haynes, Charité – Universitätsmedizin Berlin, Berlin, Germany

Glossary

Binocular rivalry – When conflicting stimuli are presented to the two eyes, conscious perception can alternate spontaneously between the input to the left and the right eyes.
Choice probability – The accuracy with which an organism's perceptual choice in a decision task can be predicted from a neural signal.
Global workspace theory – A theory that postulates the neural process underlying conscious awareness is a global distribution of information throughout the brain.
Isomorphism – Literally 'identity in structure' typically refers to the notion that similarities between perceptual experiences are reflected in similarities of the underlying neural signals.
Microconsciousness – The theory that perceptual awareness depends only on suitable representations in sensory brain regions and not on additional activity in prefrontal or parietal cortex.
Perceptual threshold – When the intensity of a barely visible stimulus is gradually increased, there is no abrupt transition between 'unseen' and 'seen' but a gradual transition with an intermediate intensity range where the stimulus is sometimes seen and sometimes not.
Reversible figures – Typically this refers to visual shapes that can be seen in different ways and can give rise to different geometric or semantic interpretations.

Introduction

The rich qualitative properties of perception have made it a key focus of consciousness research for philosophers, neuroscientists, and psychologists alike. The 'qualia' or 'raw feels' of sensory experience such as the redness of red, the timbre of an instrument, or the scent of a specific flower are the most vivid aspects of consciousness. In contrast, our abstract thoughts (such as the feeling of understanding a sentence) appear to have much weaker experiential qualities.

Research on the neural correlates of perceptual consciousness has mainly focused on visual perception, which has been studied like no other field of neuroscience and especially like no other sensory modality. The visual system belongs to the best understood and most researched parts of the brain. A number of key concepts in neuroscience originated from the field of vision such as the concept of receptive field or the role of neural synchronization. The abundance of research in this field has yielded detailed mathematical models that make detailed quantitative predictions about a number of visual phenomena. Because of the rich body of research on visual perception, this article focuses primarily on the neuroscience of perceptual awareness in the visual modality. Many of these findings can be transferred to other modalities where phenomena often have direct counterparts in the visual system.

Crossing the Threshold to Awareness

A starting point for an investigation into the mechanisms of conscious visual experience could be to compare cases where stimuli are clearly visible to cases where stimuli fail to reach awareness. For example, if the intensity of a weak, invisible stimulus is gradually increased, it will at some point be strong enough to reach awareness. The intensity where the transition from 'unseen' to 'seen' occurs is called the perceptual threshold. To compare neural processing with and without awareness, one could conduct a simple experiment

that contrasts neural responses to stimuli that are either above or below the threshold of perception. Any corresponding differences in brain activity could reflect the effect of awareness on neural processing. However, the comparison between seen and unseen stimuli would be confounded because in general the physical intensity of visible stimuli is stronger than the intensity of invisible stimuli. It would thus remain unclear whether any observed neural effects are indeed due to the difference in awareness or due to differences in the physical characteristics of the stimuli.

There are however ways to compare 'seen' and 'unseen' stimuli while at the same time avoiding stimulus confounds. When the intensity of a simple stimulus is gradually increased, there is no abrupt transition between 'unseen' and 'seen' at the perceptual threshold. Instead there is a range of intensities where the stimulus is sometimes seen and sometimes not, thus yielding a certain percentage of 'seen' and 'unseen' responses. Because there is no sharp transition from unseen to seen, the perceptual threshold is usually defined probabilistically as a specific proportion of 'seen' judgments (or alternatively a specific proportion of correct discriminations between stimulus present and stimulus absent). The gradual transition from 'unseen' to 'seen' can be used to separate physical stimulus properties from awareness in two different ways. The first approach is based on the shape of the threshold function that relates increases in physical stimulus intensity to the monotonously increasing proportion of 'seen' responses. The threshold function is s-shaped, thus visibility increases slowly for low and high physical intensities, but strongly for intermediate intensities. Because the s-shape means that visibility undergoes a nonlinear change in a range where physical stimulus properties change linearly, this allows one to separate the physics from perception by identifying brain regions that exhibit an increase in response amplitude that matches the s-shaped threshold function. Using this approach several studies have shown tight links between threshold functions for stimulus intensity and signals in early visual cortex. Threshold functions for the identification of more complex features such as objects are also closely linked to response profiles of cells in brain regions specialized in object recognition.

There is a second, even more powerful, way to study visual awareness using the gradual, probabilistic nature of perceptual thresholds. Stimuli that only reach awareness on a certain proportion of trials are very useful because they allow one to directly compare conscious and unconscious trials for the same physical stimulus parameter. For example, trial-by-trial fluctuations in perception of simple pattern stimuli are reflected by corresponding changes in activity already in primary visual cortex. Thus, already the earliest regions of the cortical visual system can closely reflect conscious visual perception of simple stimulus features. Some studies indicate that already at early stages of processing, the effects of consciousness can be stronger than the effects of physical stimulus characteristics. The differences between processing of stimuli that do or do not reach awareness is also manifest at much higher brain regions, including regions in the prefrontal cortex that are involved in top–down control of processing and in behavioral report.

A broader conceptual framework for understanding what happens in the brain when humans are viewing stimuli around the threshold to awareness is offered by perceptual decision making where subjects are required to perform simple detection and discrimination tasks. In a detection task a subject is asked to judge whether they believe to have seen a stimulus on an individual trial or not. Discrimination tasks come in two basic variants. Either the observer is shown one stimulus and has to judge which of several potential alternative stimuli it was. Or the observer is shown several stimuli and has to judge which one is which. Discrimination tasks do not probe for awareness of a stimulus but for awareness of a difference between stimuli. Performance in perceptual decision making is often accounted for by a sequence of simple information processing steps. In a first step, the presentation of a stimulus evokes a neural process encoding 'sensory evidence' about the stimulus presented. A second step consists of a 'decision variable' that is derived from the sensory evidence. The decision variable collapses all available sensory information in a way that provides for an efficient decision given the current behavioral goals. Finally, the values of the decision variable are mapped to a set of 'judgments.' For categorical

judgments such as 'stimulus A' versus 'stimulus B,' these signals reflecting the outcome of the decision are necessarily dichotomous and are typically directly related to specific motor commands that are used to indicate the judgment. Several studies have used a special type of motion stimuli to unravel the sequence of steps involved in perceptual decision making, so-called 'random dot kinematograms.' These are a blend of a 'signal' and a 'noise' stimulus. The signal stimulus is composed of a field of random dots coherently moving in one direction plus a noise stimulus consisting of randomly moving dots. The task of the subjects is to detect the drift direction of the coherently moving stimuli. The more the stimulus consists of signal dots and the less of noise dots the better the drift direction can be seen, thus yielding a threshold for perceptual motion detection. These stimuli allow one to calculate a so-called 'choice probability' that describes the accuracy with which the perceptual choice for 'unseen' and 'seen' is predicted by the activity of neurons in a specific area. A number of single-cell recordings in animals performing choice tasks with such stimuli has revealed that signals in regions of the brain specialized for motion processing partly predict the outcome of an animal's decision – and hence presumably their perception. However, the prediction from single cells in these regions is far from perfect and choice probabilities are barely above chance. This suggests that the perceptual decision is either encoded in brain areas further downstream, or that it is encoded in pools of neurons.

Perceptual decision making models can partially explain human perceptual choices at the threshold to awareness. But additional assumptions need to be made to account for what a person consciously sees in such tasks. One question is whether the perceptual experience is more closely reflected by the sensory evidence or by the decision variable. If only a single, simple stimulus is presented, the decision variable can be equated with the sensory evidence. But tasks where multiple stimuli have to be compared require a decision variable that computes a comparison, hence a relational property. In these cases the decision variable cannot directly reflect what we see but it reflects differences between things we see. A different question is how the decision making process relates to a person's subjective confidence in the accuracy of their decision. Confidence in a decision is frequently taken to be a good indicator of awareness based on the notion that if we are conscious of something we know that we see it and can be confident about our judgments. But most perceptual decision making models do not treat unseen and seen conditions as qualitatively different cases where visible stimuli undergo a different processing stream than invisible stimuli. Thus, they for example do not capture the finding that visible and invisible stimuli undergo different depths of processing in the brain.

Visual Competition: Masking and Rivalry

A different way to cross the threshold between 'unseen' and 'seen' without changing the intensity of a stimulus is to render its perception difficult by introducing additional, competing stimuli. A large number of experimental approaches follow this logic. For example, in motion-induced blindness a target stimulus can pop in and out of awareness when it is presented in the vicinity of a moving set of dots. In these cases it appears as if the target fails to win the competition for awareness against the highly salient moving stimulus. A related phenomenon is flash suppression where a brief flash in the vicinity of a target (presented either to the same or the opposite eye) can strongly reduce its visibility. Flash suppression has been shown to affect processing already very early in the visual system.

Visual Masking

One of the most prominent experimental procedures for manipulating awareness is visual masking, where the visibility of a target stimulus is decreased by presenting it in close spatial and temporal proximity to a so-called 'mask.' For example, if a target image alone is presented for brief periods it can normally be perceived quite effortlessly. However if the brief target is immediately followed by a second image consisting of an arrangement of random lines and patterns, its visibility is strongly reduced. This phenomenon is

known as backward masking and the mask image is known as a pattern mask. Visual masking has frequently been used to study the neural correlates of consciousness. Depending on conditions, masking can lead to decreases of brain activity at various stages of the visual system. But despite being invisible, masked targets can undergo a considerable degree of processing in the visual system. Object recognition is disrupted by pattern masking only at later stages of the visual system such as the lateral occipital complex (LOC) which is specialized for recognition of objects.

There are a number of different approaches to visual masking. Masks can be effective when presented either before or after the target stimuli and can be spatially either overlapping or nonoverlapping. There are also several related phenomena, including object substitution masking and crowding. A special case is so-called metacontrast masking, where masks do not spatially overlap with the targets but share common contours with them. Metacontrast masks are most effective when there is a small time delay between the target and the mask. With very brief delays or long delays the visibility is high, thus the relationship between the target visibility and the delay follows a characteristic u-shaped function. For this reason, metacontrast masking is particularly interesting because the interaction between the target and the mask has to bridge not only space but also has to seemingly operate backwards in time. The u-shaped function can be used as a signature to identify the neural locations of masking effects in the brain. Although neural activity in V1 can be disrupted by metacontrast masking, this occurs only at later stages of processing, whereas an initial transient response to the target remains unaffected. Information about masked orientation stimuli can still be available at the level of primary visual cortex, suggesting a dissociation between awareness and encoding of features in V1. If targets and masks are presented to different eyes, the masking effects are only seen in higher areas, again suggesting that the contents of consciousness are encoded beyond monocular processing in primary visual cortex. The strongest effects of metacontrast masking occur in higher levels of the visual system and in regions of the prefrontal cortex involved in high-level executive control.

Reversible Figures and Binocular Rivalry

In masking, the competition between the target and the mask is resolved in favor of the mask and the target remains invisible. But when the visual system is confronted with other types of competing or rivaling stimuli, it can sometimes oscillate between different interpretations of the visual input. One type of such multistable stimuli are reversible figures. These are typically 2D shapes that can be interpreted in different ways, such as the famous Necker cube that can be viewed either as a cube seen from the top or from below depending on which square is interpreted as being in front. Many such reversible figures are known, including stimuli that alternate between different directions of motion (e.g., stroboscopic alternating motion and 3D-structure from motion stimuli) or different semantic interpretation (e.g., a rabbit or a duck). Reversible figures are particularly interesting because they open the possibility to study changes in the contents of consciousness without any corresponding change in physical stimulation, thus allowing disentangling conscious perception from sensory factors.

Extensive research on reversible figures has shown that perceptual changes coincide mainly with changes in activity patterns of those brain regions that are specialized for the specific contents of perception. When a face becomes visible during reversible perception, the so-called fusiform face area is activated, a brain region that is specialized for processing of faces. When the perceived direction of motion of similar displays changes, the motion-processing region MT is activated. Choice probabilities for perceptual reports of ambiguous 3D structure from motion stimuli are very high even for individual cells in motion-processing regions of the brain. This suggests that changing the contents of consciousness involves mainly processes in regions specialized for these contents. Similarly, the transition periods between different percepts seem to involve high-level in prefrontal cortex, suggesting that the stability of our conscious perception might be regulated in high-level executive control regions.

A phenomenon similar to reversible figures is binocular rivalry, where a competition arises not between two interpretations of the same figure but

between a conflicting stimulation of the two eyes. When conflicting stimuli are presented to the two eyes, visual perception cannot fuse them. Instead, perception alternates between seeing the stimuli presented to the left and to the right eye. There has been a long controversy regarding the exact locus of such rivalry in the brain. One dominating view is that rivalry is due to a conflict between monocular populations of cells in the early stages of the visual system. Information about the eye-of-origin is largely lost beyond primary visual cortex, and truly monocular cells can only be found in V1 and earlier stages of the visual system. In this monocular account of rivalry, the stimulus presented to one eye is suppressed from awareness because the input to this eye is attenuated at a monocular level of processing. This theory was supported by several behavioral findings. For example, the sensitivity to input from the suppressed eye is reduced, suggesting that processing of the suppressed eye is reduced. Similar support for a monocular selection was that if the input is exchanged between the dominant and suppressed eyes, perception in most cases tends to follow the eye, not the stimulus.

However, an alternative pattern-based view was subsequently suggested. This was based on findings when mixed images are presented to the two eyes that allow for perceptual grouping between the two eyes. To create such stimuli one starts with two images, say a face and a tree. The next step is to mix the two images by exchanging coherent subregions in one image by the corresponding regions in the other image. This results in two complementary images, each showing sections of a face in some regions and sections of a tree in other regions. The question is: If perception during rivalry is dominated by the input from only one eye at a time, one would expect perception to alternate between the patchy image presented to the left eye and the complementary patchy image presented to the right eye. Instead there is a tendency to see coherent percepts that combine input from both eyes into a meaningful figure. This suggests that perception during rivalry depends not only on the eye of input and thus the monocular account of rivalry cannot be the full truth. This pattern-based account of rivalry was supported by single-cell recording studies that suggested that rivalry only affects 20% of cells in regions V1, the latest stage with substantial monocular information. Instead, the main effects of rivalry seemed to be restricted to higher levels of the visual system. In regions of the temporal lobe that are specialized in high-level object recognition, rivalry affects 90% of cells. Similarly, rivalry in humans has been shown to affect high-level regions of the visual system. For example, when rivalrous perception alternates between faces and houses, corresponding increases can be seen in functional magnetic resonance imaging (fMRI) signals measured from regions of cortex specialized for processing faces and houses. However, in the following years studies on humans have gathered evidence that rivalry might affect earlier stages of neural processing than previously believed. It has been repeatedly shown that rivalry affects fMRI signals in primary visual cortex and even has effects on subcortical processing in the lateral geniculate nucleus. The current view on the mechanisms of binocular rivalry combines the monocular and the pattern-based explanations and postulates a selection for awareness at multiple stages of the visual system. Thus, the access to awareness can be regulated at many levels of processing ranging from very early subcortical regions to prefrontal cortex.

Encoding the Contents of Consciousness

Reversible figures and binocular rivalry have long dominated experimental research on visual awareness because they allow dissociating changes in conscious perception from mere stimulation factors. When the contents of consciousness change without corresponding changes in the intensity or level of awareness of conscious perception, this might help isolate where the contents of our consciousness are stored in the brain. The rationale is that a brain region that encodes the contents of our conscious perception will need to change its activity when the contents of perception change. However, at a closer look this approach has a severe flaw. It does not allow one to disentangle brain regions that specifically encode the 'contents' of consciousness from brain regions that are 'unspecifically' involved in switching between different contents.

Besides involvement during perceptual transitions, more needs to be demonstrated to show that a brain region encodes an aspect of conscious perception. It needs to be shown that the neural responses in this brain region change in a 'content-specific' manner. The contents of consciousness can be described along a number of different dimensions. Neuropsychological data from patients with brain lesions show that different dimensions of perceptual space are encoded in a number of different brain regions. For example, there is a double dissociation between regions encoding for color and motion because awareness of both can be disrupted independently following lesions either to color-selective brain regions in the fusiform gyrus versus lesions to the motion-processing region MT. Studies on agnostic patients have revealed that the perception of complex objects can also fail independently of the perception of the simple features they are composed of. This independent drop-out of specific contents of consciousness following specific brain lesions suggests that different aspects of awareness are encoded in separable cortical regions.

Lesions can provide valuable information about brain regions encoding different aspects of awareness. But they still leave several questions open. For example, lesions to primary visual cortex cause an almost complete loss of conscious visual perception. But this does not mean that all contents of our perception are encoded in V1. Cells in primary visual cortex encode flicker at much faster rates than can be perceptually resolved. If V1 really were encoding the contents of conscious perception, then one would expect that the temporal resolution of perception would match the temporal resolution of signals in V1, which is clearly not the case. Similarly, when a stimulus is flashed into just one of the two eyes, primary visual cortex still encodes the eye-of-origin. But perceptually a subject can typically not tell which eye was stimulated, demonstrating again that V1 encodes information that does not reach awareness. Thus, these dissociations raise doubt whether signals in primary visual cortex encode contents of visual consciousness or rather serve a role of relaying information into the cortical visual system without participating in awareness directly.

Such dissociations between encoding of features in V1 and the properties of visual awareness can be taken even a step further by examining precisely how much 'information' is encoded in a neural response and how this compares to the perceptual information available for awareness. Unperceived flicker and unconscious eye-of-origin are cases where a neural process has more information about a stimulus feature than is represented in consciousness. The complementary case would be if neural signals had less information about a stimulus than is encoded in perceptual consciousness. Take as an example color perception. For a neural population to encode color percepts, it must respond with at least one different state to each identifiable color hue. If there were fewer neural states than possible color percepts, then the neural population would not be powerful enough to encode all the possible perceptually distinguishable shades of color. This suggests a useful test: To find out whether a neural population encodes perceived color hue, one can try and decode the perceived hue from signals in that neural population. If there is indeed at least one neural signal that corresponds to each perceptual state, it should be possible to fully decode color perception from this neural signal. This argument directly addresses the content-specificity of neural correlates of consciousness. It can be used to rule out that a brain area that is involved in awareness is merely realizing unspecific enabling factors of awareness, rather than encoding specific perceptual contents.

More generally, the contents of conscious perception can be described along a large number of dimensions. For each point in the visual field, we can define its brightness, its color hue and saturation, its contrast, its speed and direction of motion, depth, and many more. Furthermore, simple features can jointly form geometric shapes and even meaningful objects. The contents of consciousness along a number of different dimensions can be jointly viewed as a complex 'perceptual space.' Using an information-theoretic framework, one can directly search for the population of neurons that encodes each specific subdimension of this perceptual space. Each subdimension of perceptual space can be viewed as a data structure that is encoded in some brain region by some parameter of brain activity.

There are several techniques available for assessing the information encoded in neural populations.

One approach is to record from 'single cells' and to relate their encoding of a stimulus with the perceptual information available to a human or animal. But this neglects the information contained in distributed networks of neurons. A more powerful approach is to take into account the full information contained in 'populations of neurons' within a brain region. For example, to understand how motion percepts are encoded in the motion-processing region MT, one has to take into account not only the activity of single neurons, but of the entire neural population. This is because even a cell that is not tuned exactly to the feature of interest can still carry information about this feature. In fact, information-theoretic analyses have shown that neurons carry only very little information about features where they show the strongest responses. This is because at the peak of the tuning curve the neural responses do not differ very much for different features. Most information is contained in the side bands of the tuning function, where the neuron changes its response rapidly with changes in physical features. Using such decoding approaches it has been shown that single neurons in regions of the medial temporal cortex that are involved in recognition and memory carry information about specific visual contents, such as thoughts about specific individuals. Despite the distributed nature of neural representation, certain cells can exhibit an incredible sparseness, meaning that each visual object is encoded by only a few cells. Such cells respond say to a picture of a specific person, but not to pictures of other people. Information-theoretic approaches to the encoding of contents of consciousness are very powerful. But, depending on the theoretical perspective, they could be considered too powerful, because they can extract information also from 'unusual' brain signals such as for example from the side bands of orientation tuning curves or from deactivations of brain activity. Some theories about awareness in contrast postulate that the contents of consciousness are encoded in the brain in the form of explicit rather than implicit representations. An explicit representation directly signals the presence of a particular feature and does not need further processing to be read out. For example, the encoding of a face in a face-selective cell constitutes an explicit representation whereas the representation of a face in the retina constitutes an implicit representation because it requires additional processing to tell that a face is present in the spatial pattern of retinal signals.

To access the information encoded in entire neural populations, it is necessary to record from multiple cells simultaneously using so-called multielectrode grids. Such recordings can only be done in animals and occasionally in humans with implanted electrodes for diagnosis of epilepsy. A noninvasive alternative approach is offered by decoding techniques for electroencephalography (EEG), magnetoencephalography (MEG), and fMRI signals. These noninvasive techniques cannot resolve individual cells; the resolution is not even sufficient to resolve individual cortical columns (the basic units of information storage in the human brain where cells encoding similar features are clustered together across a span of approximately 0.5 mm of cortex). But these noninvasive techniques can nonetheless provide a handle on information encoded at a fine-grained scale in the visual cortex. This is possible due to small fluctuations in the topography of cortical maps that produce interference patterns that can be picked up with a standard fMRI measurement grid. Using information-theoretic decoding techniques, it is possible to access the information contained in these interference patterns, for example, about a specific type of visual feature, and to compare it to the information available to a human subject who is currently perceiving the same feature. Such decoding-based neuroimaging has been used to reveal which brain regions encode which visual features and how such information relates to the information encoded in visual awareness. This has revealed that orientation-information that does not reach awareness can nonetheless be encoded at the level of primary visual cortex, another proof of a dissociation between V1 signals and visual awareness. Because brain areas further downstream in the visual system do not encode unconscious orientation information, it is plausible that the neural correlates of the contents of visual consciousness start later in the visual system.

Interestingly, the same content-based neuroimaging techniques can be used to go even a step further in mapping conscious experience to brain processes. It is possible to investigate how relationships between elements of conscious experience are reflected in similar relationships between

brain activation patterns encoding them. Such an 'isomorphism' between phenomenal experience and brain activity can for example be found in the way objects are encoded in the human temporal lobe. The perceived similarity between objects is reflected in a similarity between the brain activation patterns in the object-recognition regions of the temporal lobe.

Awareness and Specific Brain Regions

Area V1 or 'striate cortex' is an excellent showcase of the different types of arguments that are put forward when discussing the potential involvement of a brain area in conscious perception. The role of primary visual cortex in visual awareness has been heavily debated. Primary visual cortex is the first cortical stage of information processing and is thus an important entry point for visual information into visual cortex. Despite the fact that other entry points exist where subcortical regions directly project to more high-level visual areas, the majority of information enters the visual system through V1.

The key role of V1 in awareness can be seen from the fact that a lesion in V1 will always cause a fully blind section or 'scotoma' in the visual field. The scotoma exactly matches the corresponding retinotopic location of the visual field. For example, if the upper left section of V1 is fully lesioned, this causes a blind region in the lower right quadrant of the visual field. Similar lesions in higher level visual areas typically do not cause a full dropout of visual sensitivity. They only affect the perception of specific features such as color, motion, or depth. Even lesions in V2 only affect visual acuity and contrast perception. Thus, it appears that V1 plays a special role among the visual areas in that it is necessary for visual awareness. This is supported by a number of demonstrations of close correlation between processing in primary visual cortex and visual awareness. Activity in V1 closely reflects perception of simple visual features such as contrast and brightness. Even simple forms of perceptual integration such as the perception of texture boundaries and contours can be explained from properties of V1 neurons. FMRI signals recorded in human V1 correlate with the conscious percept during binocular rivalry. When experimental subjects train to see subtle differences between simple visual stimuli, a phenomenon called perceptual learning, the improvement in performance correlates with changes in tuning properties in primary visual cortex.

However, when arguing for a role of a specific region in visual awareness from experimental correlations caution is required. The important question needs to be addressed, whether any correlations observed indeed reflect necessary conditions for awareness or whether they are purely incidental or 'epiphenomenal.' The retina can also be considered a necessary condition for visual awareness because a loss of both retinae causes a complete loss in visual perception. This does not mean that the retinae are strictly necessary for awareness. Instead it means that the retina is a necessary step in the normal causal chain of events leading to conscious perception. But conscious visual perception can also be caused by bypassing the retina and directly stimulating visual cortex using implanted electrodes in patients, during surgery or by using transcranial magnetic stimulation (TMS). Also, during visual hallucinations and visual imagery, V1 is not always involved.

An alternative view would be that the disruption of visual perception following lesions to V1 simply reflects the loss of the major 'input' channel to the visual system. It does not directly imply that V1 is always required for every type of conscious visual experience. Furthermore, there are a number of reasons that have cast doubt on a close involvement of V1 in visual awareness. On the one hand there are theoretical reasons. For example, we can directly access, manipulate, and act upon information that is in our consciousness. To explain such access, it seems necessary to assume a direct projection from regions of the brain involved in high-level control of behavior to sensory regions encoding the contents of consciousness. However, there are no direct projections between prefrontal cortex and V1 that could support such access. A further argument can be obtained from the fact that there are striking dissociations between our conscious visual experience and encoding of information in V1. For example, we are not able to consciously tell whether a monocular stimulus is presented to the

left or right eye. But V1 encodes eye-of-origin and also many other features that fail to reach consciousness such as unconscious orientation-information and unperceived high-frequency flicker. In crowding, the visibility of a stimulus that is normally clearly visible is reduced by presenting other stimuli in the surrounding regions of the visual field. Stimuli that fail to reach awareness under crowding conditions are nonetheless processed at least up to primary visual cortex, because they leave traces of orientation-selective adaptation, suggesting that an encoding of information in V1 does not automatically lead to awareness.

Even if V1 does not encode the contents of our consciousness, several findings suggest that it might nonetheless be required for awareness. In some patients large regions of visual cortex beyond V1 can continue to respond to visual stimuli despite lesions to V1 that preclude input arriving through primary visual cortex. This means that activity in such 'extrastriate' visual regions is not sufficient in itself to produce awareness without a contribution from V1. This is further supported by the fact that although residual visual sensitivity can remain in a scotoma, the subject does not subjectively feel to be seeing anything. A subject can be above chance in guessing which stimulus was presented, but they will have the impression to be seeing nothing in that region of the visual field. Such 'blindsight' is presumably mediated by pathways into extrastriate visual cortex that bypass V1, thus lending further support for the notion that activation of extrastriate cortex can be sufficient to support visually-guided behavior but is in itself not sufficient for awareness.

The term 'extrastriate cortex' refers to a group of visually responsive brain regions beyond V1 or 'striate cortex.' An important feature of early extrastriate areas V2, V3, and V4 that they share with V1 is their retinotopic organization. This means that the topography of the visual field is largely preserved in the visual maps of V1–V4, despite undergoing a coordinate transform from Cartesian to polar coordinates. Because detailed spatial information is progressively lost at higher stages of the visual system beyond V4, this means that signals encoding the topographic spatial layout of visual perception can only be found in these early visual regions.

Regions beyond V4 exhibit an increasing specialization for different visual features. Motion is primarily processed in a dedicated area labeled MT located in the back of the temporal lobe. The main cortical color region is located in the fusiform gyrus at the bottom of the temporal lobe, although there is some debate as to whether this is the same region as retinotopic area V4. The main cortical region for object recognition is called the LOC and is located in the lateral occipital lobe and in the posterior fusiform gyrus. Visual contents can be selectively lost from consciousness. For example, color and motion can be independently disrupted following lesions to the corresponding brain regions. After lesions to specific regions of the ventral temporal lobe, patients can lose color perception, thus perceiving the world in shades of grey while other qualities of visual perception such as motion are spared. Similarly, lesions restricted to MT cause loss of motion perception without loss in color perception. The modular encoding of different aspects of visual perception is further supported by direct cortical stimulation to patients undergoing brain surgery. Depending on which visual region is stimulated, patients report seeing complex patterns, colors, or movement.

An interesting property of cells at the higher levels of the visual system is that they respond independent of the detailed physical characteristics and context of a presented object. This phenomenon is termed 'invariance.' Say, an actress might be wearing a different dress and sporting a different hairdo, but still the cell would recognize the invariant person encoded in the image. But this invariance also comes at a price. If a cell responds to a complex visual object independent of the constituent features, it means the cell has lost all information about the fine-grained features such as brightnesses and colors that the object is made up of. This means that the encoding of contents of consciousness necessarily occurs at multiple levels. Brain regions coding complex, invariant aspects of our visual experiences can in principle not be coding the simple features. The fact that the processes underlying consciousness are fragmented, modular and multilevel is well documented, but this stands in direct contradiction to our impression that visual experience is unified rather than divided into a number of different features. The distributed encoding of the contents of consciousness in extrastriate visual areas requires additional

assumptions in order to explain the unity of consciousness. This problem is known as the 'binding' problem. The most prominent but also controversial explanation of binding assumes that features are bound by synchronization of cells in different brain regions.

An important dissociation in the visual system is that between representations for action and representations for awareness. There are two major visual pathways in the extrastriate visual system. The dorsal pathway is involved in spatial transformations and actions, whereas the ventral pathway is involved in object recognition. Patients with lesions of the ventral pathway cannot describe a visual stimulus any more, but can still perform visually-guided actions to the input. Patients with lesions of the dorsal pathway show the opposite pattern of disorders. They cannot perform visually-guided behavior but can consciously describe the stimulus. This has been interpreted to mean that only the ventral visual pathway supports conscious perception, whereas the dorsal pathway supports unconscious visual guidance of behavior, similar to the blindsight that follows lesions to primary visual cortex.

Supramodal regions beyond the visual system also play an important role in perceptual awareness. When a stimulus crosses the threshold to awareness, this also leads to changes in activation of regions of prefrontal cortex. For example, during studies where subjects are required to recognize masked versus unmasked words, conscious perception is correlated with increased activation in several prefrontal brain regions. Similarly, during perception in binocular rivalry, a frontoparietal network is involved in perceptual transitions, suggesting that these regions might be involved in awareness. One possibility is that these prefrontal brain signals reflect the global distribution of information as postulated by the so-called global workspace theory of consciousness. According to this theory, neural representations reach awareness when they are distributed to other brain regions. An alternative view is that content-specific brain activity within individual sensory brain regions is sufficient for awareness of these contents to occur, a theory referred to as microconsciousness. To date it is unclear whether the frontal and parietal processes involved in awareness are indeed content-specific and thus can be claimed to reflect a distribution of information. Instead they could reflect unspecific processes, for example, they might be involved in causing awareness of a stimulus without themselves encoding the contents of consciousness. Alternatively they could be involved in the subject consciously noticing and reporting a change in the contents of their consciousness.

Another important question for the neural mechanisms of awareness is whether they involve large-scale integrative processes that jointly involve a number of brain regions. One line of thought posits that awareness of representations is related to specific dynamic interactions between populations of neurons. For example, one controversial theory suggests that awareness and binding are both closely related to synchronization between the distributed population of cells encoding the different features of an object. A similar hypothesis states that awareness of representations is related to the stage of recurrent cortical processing where feedforward and feedback neural processes overlap in early sensory regions. This is supported by several invasive studies in monkey visual cortex, as well as by studies using TMS to knock out later stages of processing in V1. When MT, the main cortical motion-processing region, is stimulated using TMS, this can create the illusory perception of movement. When a second TMS pulse is administered to V1 just following the first pulse to MT, the perception of movement is abolished. This could mean that awareness of motion depends on intact feedback projections into V1. In contradiction to this finding, patients with the so-called Riddoch syndrome who have lesions to V1 can have selectively spared motion perception despite global blindness. Thus, recurrent processing involving V1 activity cannot be a strictly necessary condition for visual awareness of motion.

Attention and Awareness

Awareness and attention are so closely related that many researchers consider them to reflect the same process. A key feature of both attention and awareness is their selectivity. Only a small amount of sensory information that is available at a given time reaches awareness. This limitation is most clearly apparent when there is a lot of competition for access to our consciousness. For example in a crowded room we might fail to immediately notice

a friend who is present although he is clearly visible. In such situations we have to consciously scan our environment by focusing our attention on one person at a time, and only when attention is focused on a known person we will be able to recognize them. Such cluttered scenes help understand the neural mechanisms that underlie the selectivity of visual awareness. According to the biased competition model of visual attention, stimuli are in competition for access to processing resources. When several stimuli are present in the visual field (or in the receptive field of cell), attention is required to bias or boost the processing of a selected visual feature, and the unselected stimuli are suppressed from further processing. Such effects have been demonstrated for single cells as well as for population responses measured with fMRI. The general finding is that competition increases at advanced stages of visual processing where receptive fields increase in size and competition occurs between an increasing number of elements in the visual field. This model of competitive interactions between processing units provides a model for understanding the selectivity of conscious visual experience.

The selectivity of attention and awareness seems to imply that the two might be two aspects of the same process. In this view attention functions as a gatekeeper to consciousness. There are many demonstrations of the tight relationship between attention and awareness. Lesions to attentional control regions can cause a disorder of awareness known as visual hemineglect, where stimuli can fail to be noticed when presented in the contralesional visual field. Another striking example is a phenomenon called inattentional blindness, where subjects fail to notice highly salient events in their visual field when their attention is directed elsewhere, thus highlighting the importance of attention for visual awareness. A related phenomenon is change-blindness, where subjects fail to notice marked changes occurring in their visual environment. In the typical experiment, a person is monitoring a cluttered visual image that is repeatedly flashing on and off for any changes made between two successive presentations. Most people have the intuitive assumption that they perceive the entire visual scene surrounding them and thus would definitely notice a salient change. In contrast to this intuition, most people fail to perceive quite marked changes unless explicitly paying attention to the particular region of the visual scene where the change occurs. These experiments can be useful for tracking down the neural correlates of consciousness. Major changes to visual scenes that go unnoticed can still continue to be registered and analyzed up to higher stages of the visual system. FMRI signals from high-level object-selective regions signal changes in a visual scene that are not registered. Similar findings have been obtained in single-cell recordings in human medial temporal areas related to object recognition and memory.

Despite the close link between attention and awareness, there are also several findings that demonstrate that the two cannot be fully equated, although they are clearly closely related. One important question is whether stimuli outside the focus of our attention can really not reach our conscious awareness. There are several experiments that suggest that this might be the case. In one line of experiments, the attentional resources of subjects are engaged maximally at one region of the visual field, typically the center of gaze, by requiring them to perform a very difficult visual discrimination task. Despite the full engagement of attention, they can still continue to perceive certain stimuli presented at a different region of the visual field. Simple, salient stimuli such as colors or shapes can be identified without attention. Even more complex aspects of a visual scene can be perceived in parallel without requiring attention. For example, observers can readily and rapidly perceive whether an image contains an animal or not, even when the animal is hidden in a cluttered natural landscape. EEG signals recorded from humans during such tasks show signatures of rapid recognition as early as 120 ms after the onset of a stimulus. In contrast, complex combinations of geometric features cannot be perceived rapidly while attention is directed elsewhere. This suggests that the ambient 'fringe' surrounding the focus of our attention can still enjoy a fair degree of complex, high-level cortical processing and lead to awareness.

Summary

A large number of experimental studies have contributed strongly to our understanding of the neural mechanisms of visual perceptual experience.

The contents of perceptual experiences are encoded in distributed subregions of modality-specific cortex. The access to consciousness appears to be regulated at multiple stages of processing, reaching from early subcortical processing to high-level regions involving prefrontal and parietal cortex. However, several theoretical debates regarding the specific processes involved still await further clarification. It is currently unclear whether representations that reach awareness are globally made available for further processing across the brain. And it is also unclear how the many subdimensions of perceptual space that are distributed across many brain regions give rise to a unified perceptual experience.

See also: Attention: Change Blindness and Inattentional Blindness; Bistable Perception and Consciousness; Blindsight; Neglect and Balint's Syndrome; Perception: The Binding Problem and the Coherence of Perception; Visual Experience and Immediate Memory.

Suggested Readings

Baars BJ (2002) The conscious access hypothesis: Origins and recent evidence. Trends in Cognitive Sciences 6: 47–52.
Crick F and Koch C (1998) Consciousness and neuroscience. Cerebral Cortex 8: 97–107.
Dehaene S and Naccache L (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79: 1–37.
Driver J and Vuilleumier P (2001) Perceptual awareness and its loss in unilateral neglect and extinction. Cognition 79: 39–88.
Engel AK and Singer W (2001) Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences 5: 17–25.
Gold JI and Shadlen MN (2001) Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences 5: 10–16.
Haynes JD and Rees G (2006) Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7: 523–534.
Heeger DJ (1999) Linking visual perception with human brain activity. Current Opinion in Neurobiology 9: 474–479.
Lamme VAF (2003) Why visual attention and awareness are different. Trends in Cognitive Sciences 7: 12–18.
Lamme VAF, Supèr H, Landman R, Roelfsema PR, and Spekreijse H (2000) The role of primary visual cortex (V1) in visual awareness. Vision Research 40: 1507–1521.
Milner AD and Goodale MA (1995) The Visual Brain in Action. Oxford: Oxford University Press.
Rees G, Kreiman G, and Koch C (2002) Neural correlates of consciousness in humans. Nature Reviews Neuroscience 3: 261–270.
Thiele A and Stoner G (2002) Neuronal synchrony does not correlate with motion coherence in cortical area MT. Nature 421: 366–370.
Tong F (2002) Primary visual cortex and visual awareness. Nature Reviews Neuroscience 4: 219–229.
Zeki S (1993) A Vision of the Brain. Oxford: Blackwell.

Biographical Sketch

John-Dylan Haynes is Professor for cognitive and computational neuroimaging at the Bernstein Center for Computational Neuroscience, Berlin, Germany. His main interest lies in consciousness research where he has worked on the neural mechanisms of visual awareness and on unconscious neural determinants of human decisions. His research combines psychophysics, neuroimaging and advanced data analysis methods, such as connectivity measurements and multivariate pattern recognition. The latter can be used to decode mental states from patterns of human brain activity, thus constituting a rudimentary form of "mind-reading." He is currently a board member of the Association for the Scientific Study of Consciousness (ASSC) and will be organizing the 13th meeting of the ASSC in Berlin (with Michael Pauen).

Neurobiological Theories of Consciousness

S Kouider, Laboratoire des Sciences Cognitives et Psycholinguistique, CNRS/EHESS/ENS-DEC, Paris, France

Glossary

Neural correlates of consciousness – They are defined by Christoph Koch as ''The minimal set of neuronal mechanisms or events jointly sufficient for a specific conscious percept or experience.'' They allow to avoid the difficult problem of directly looking for neural bases.

Panpsychism – Reflects the philosophical doctrine that everything (in Greek, 'pan') has a mind ('psyche') and is therefore conscious. Some theories presented in this article endorse a certain form of panpsychism in which anything that transmits information is in a way conscious.

The hard problem – It is the problem of explaining how and why we have the subjective experience of consciousness. It is often contrasted with the easy problem, which consists of describing consciousness as the cognitive ability to discriminate, integrate information, focus attention, etc.

Introduction

Neuroscientists working on the issue of consciousness consider that it is a biological problem. They assume that we will understand how and why we are conscious by studying the cerebral and neuronal features of the brain. These theories have largely benefited from the recent advances in neuropsychology, neurophysiology, and brain imaging in particular. However, neurobiologists have also been influenced, on the one side, by cognitive theories aimed at characterizing the psychological determinants of consciousness, and on the other side, by philosophical issues related to the mind–body problem.

Cognitive Influences on Neurobiological Accounts

Regarding the influence of cognitive theories, the majority of neurobiological accounts can be seen, in fact, as extensions of preexisting cognitive theories (e.g., for instance global workspace theories). Indeed, one of the main tasks exercised by neurobiologists in the last two decades has been to search for cerebral or neuronal equivalents to the functional elements constituting cognitive models (e.g., the dorsolateral prefrontal cortex for voluntary control, or long range axons for connecting brain regions associated with 'unconscious' and 'conscious' processing). Of course, many neurobiologists disagree with this approach. Consciousness, because it is a biological problem, should be reframed the other way around, by focusing primarily on its structural basis rather than relying on cognitive theories, often considered too speculative. Therefore, many neurobiologists consider that an ideal neurobiological science of consciousness should focus on neural structures and mechanisms in order to understand how the organic matter constituting the brain creates consciousness.

The Hard Problem for a Neurobiology of Consciousness

Yet, studying the neural mechanisms 'leading to' consciousness, trying to explain the 'emergence' of consciousness, or focusing on how the brain 'creates' consciousness, as often described in neurobiological literature, sounds as if it involved an immaterial soul that would magically arise from the brain. This is not a new issue for philosophers who have also been wondering about the equivalent mind–body problem since antiquity. More than a century ago, the contemporary 'brain-consciousness' problem was well captured by Thomas Huxley's famous remark: ''How it is that anything so remarkable as a state of consciousness

comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the djinn when Aladdin rubbed his lamp in the story.'' The same issue applies today: understanding consciousness as an 'emergent' property 'arising' from functional elements of the neurocognitive architecture, without falling on a dualistic position where consciousness lies somewhere outside of the brain, poses serious epistemological difficulties and leads to the so-called hard problem of consciousness.

Indeed, many philosophers have concluded that there is not one single problem, but actually two problems that are faced by anyone trying to understand consciousness: they distinguish between the so-called easy problem and hard problem. In a nutshell, the easy problem consists in relying on objective measures of conscious processing in order to explain how one is able to discriminate sensory events, integrate information, report mental states, focus attention, etc. By contrast, the hard problem consists in explaining the first-order, subjective nature of qualias and phenomenal states, the 'what is it like to be conscious' as well as how and why we experience consciousness at all. Addressing the hard rather than the easy problem of consciousness constitutes an important epistemological constraint put forward by philosophers. In particular, contemporary philosophers such as Joseph Levine and later David Chalmers have argued that trying to resolve the hard problem leads to an 'explanatory gap' that science is unable to cross, at least today. Indeed, they stress the fact that it appears impossible to demonstrate that a neural structure leads to a conscious experience, while denying the reverse possibility. In addition, given how different they are, reducing phenomenal states to neural states appears almost impossible.

Looking for Neural Correlates, Not Neural Bases

Should neurobiologists then give up on addressing this issue? Most neurobiologists acknowledge the existence of a hard problem. However, they also endorse the principle that further scientific investigations will ultimately allow us to resolve it. Others explicitly deny the existence of a hard problem in the Chalmerian sense. For some, assigning too much importance to the explanatory gap might actually turn out to be counterproductive and impedes rather than facilitates scientific progress. Accordingly, neurobiologists have mostly focused on the easy problem, considering that this strategy will progressively get us closer to understanding the full issue. They extended the 'contrastive analysis,' originally put forward by Bernard Baars, from the cognitive to the neurobiological domain. While this method initially consisted in contrasting conscious and unconscious processes in order to characterize their cognitive features, the neurobiological approach aims at characterizing the neural features. A typical example consists in comparing the cerebral activity when subjects are presented with subliminal stimuli they cannot report (unconscious processing) with that of visible stimuli they can report (conscious processing).

In other terms, the current first step in trying to understand the link between consciousness and the brain consists in finding out which neural components are specifically involved during conscious processing, but importantly not during unconscious processing. Francis Crick and Christoph Koch have coined the term 'neural correlates of consciousness' (NCC; see Glossary) in order to describe this epistemological approach. According to them, the best strategy for a neurobiological science of consciousness is to search for the NCC. Underneath this approach is the crucial principle that 'correlates' do not imply any relation of causality between the occurrence of conscious mental events and their associated physiological structure. Consequently, this strategy has the advantage of leaving aside, at least for the moment, the hard problem of finding the neural 'bases' of consciousness.

In the following sections, I will provide an overview of the current most influential neurobiological theories of consciousness. These theories will be largely described in an independent manner, such that each of them can be understood individually, that is, without having to frame it in the context of alternative accounts. Only later, in the section labeled 'Neurobiological standpoints on the hard problem' will I evaluate their explanatory power by confronting them in relation to some important conceptual issues (e.g., dissociating access vs. phenomenal consciousness, dissociating attention vs.

consciousness, panpsychism). I will conclude by emphasizing how promising these theories are in getting us closer to resolve the issue of the hard problem.

From Globalist to Localist Accounts of Consciousness

Neurobiological theories of consciousness differ in many respects. One way to portray them in a coherent manner is to follow the large spectrum ranging from globalist to minimally localist accounts. By globalist or localist I refer to the size of the brain states that are assumed to be sufficient for consciousness (extended to large parts of the brain vs. focalized to specific and small brain areas, respectively). I will mainly restrict this review on the cerebral level and exclude alternative theories that are more globalist, in the sense that consciousness encompasses more than neural activity in the brain (e.g., theories proposing that consciousness reflects modification in the electromagnetic field surrounding the brain, as proposed by J. McFadden and by R. John), or, conversely, theories that are more localist, by focusing on single neurons or even lower structural levels (e.g., the quantum-level theory of microtubules by S. Hamerof and R. Penrose). Such theories remain excessively speculative and unspecified to be included in a serious review, at least for the moment.

The Reentrant Dynamic Core Theory

The Reentrant Dynamic Core theory, proposed by Gerald Edelman and Giulio Tononi, is arguably the most globalist account of consciousness. Indeed, in this framework, consciousness is not to be localized in dedicated brain areas or with particular types of neurons. Rather, it is the result of dynamic interactions among widely distributed groups of neurons in the entire thalamocortical network (the 'dynamic core'). This theory offers an interesting tentative of unifying the hard and easy problems, providing a neurobiological explanation for qualias, and explaining both phylogenetic and ontogenetic aspects of the development of consciousness in humans and other species. Yet, although this theory is very appealing, especially given its explanatory power, it is also highly speculative, and based on several assumptions that remain to be demonstrated. I will further discuss the speculative aspects of this theory later (see the section 'Neurobiological standpoints on the hard problem'). For now, I shall provide an overview of the core assumptions underlying this theory.

In order to appreciate the specificity of the Reentrant Dynamic Core theory, it is important to understand that, regardless of its explanation for consciousness, it offers an alternative view on brain structures, considering the wiring of the brain into neuronal assemblies as the result of variation and selection mechanisms that are analogous to those underlying evolutionary theories. This macrolevel account of brain development, formally developed by Edelman in 1978, and called the theory of Neuronal Group Selection (also called the theory of Neural Darwinism) constitutes a key element for understanding the development of consciousness. According to the Neuronal Group Selection theory, the brain is assumed to be a selectionist system in which variant groups of neurons are selected over others in three steps. The first one, termed developmental selection, happens during embryogenesis and early development, and it is largely influenced by epigenetic factors. It consists of several processes such as cell division, cell death, extension, which has the consequence of connecting neurons together into a large number of variant neuronal circuits (labeled primary repertoires), and elimination, which targets unconnected neurons in particular. The second step is called experiential selection and lasts from early infancy through all of the lifespan. It consists of reinforcing, through the influence of behavior and experience with the environment, the synaptic connections of some variants over others, leading to secondary repertoires of neurons. The final step, that of reentry, consists of the formation of massively parallel reciprocal connections among distant maps of neuronal repertoires, allowing them to exchange signals and be spatiotemporally coordinated.

In the Reentrant Dynamic Core theory, consciousness results exactly from this mechanism of reentry among distant groups of neurons within the dynamic core of thalamocortical connections. The spatiotemporal coordination provided by reentry allows the binding of several elements into a single and coherent object or event, providing a solution to the binding problem. It also allows to

explain why conscious experiences appear to be unified.

Particularly important in this theory is the reentry of information between groups of neurons dealing with perceptual categorization (i.e., in posterior areas) and the more frontally located systems responsible for value-category memory. Indeed, the latter will constrain the selectionist process by modulating or altering the synaptic connections within groups of neurons, as a consequence of their influence on behavior (pleasure, pain, etc.). Another key aspect to this whole framework is that groups of neurons are constantly in competition and their survival depends on creating or reinforcing synaptic links with other groups, such as to form large assemblies of reentrant neuronal maps in the dynamic core. The victorious assembly (or 'coalition' in Crick and Koch's terms) will lead to consciousness, at least for a few hundred milliseconds, until a new coalition of neurons bypass it. Indeed, because the variant groups of neurons are assumed to be degenerate, which means that different variants of neuronal assemblies can actually carry out the same function and have the same output, each integrated state in the dynamic core is followed by yet another and differentiated neural state in the core. As such, coalitions of neurons leading to consciousness are temporally transient by nature and widespread along variant regions of the whole dynamic core. Because interactions within the dynamic core are constantly moving, it explains the diversity of consciousness, yet the constant integration through reentry explains the unity of consciousness. Note then that the victory of a coalition of neurons in leading to consciousness is not just a matter of its size; rather it depends on its complexity, that is, its ability to generate at the same time an integrated scene (i.e., appearing as a unitary event) and a differentiated scene (i.e., which can be highly discriminated from other scenes). Measures of complexity in relation to consciousness have been further developed by Tononi in his recent Information Integration theory of consciousness (see the article 'cognitive theories of consciousness' in this volume). According to Edelman and Tononi, these highly discriminatory properties of neural complex systems are the qualias that have been torturing philosophers, nothing more, nothing less! Finally, they distinguish between primary consciousness that allows for a perceptual organization of the environment and whose characteristics have been presented above, and higher-order consciousness that is possessed by humans, and which is related to linguistic and symbolic mental activities. The latter requires further reentry with additional brain regions such as those involved in language production and comprehension.

In sum, this theory assumes that the brain is a dynamic complex system in which consciousness emerges from the interactivity itself. Instead of involving top-down communication between dedicated areas, the Reentrant Dynamic Core theory involves regions in the cortex in active communication with one another and with associated nuclei in the thalamus. Consciousness arises from the differentiated integration of activity in these areas, as information is transmitted recurrently, with local groups of neurons performing their specialized and discriminatory function, while at the same time being unified with other neuronal groups of the dynamic core. Before concluding on the characteristics of this theory, it is important to point out that although the Reentrant Dynamic Core can be considered as a globalist theory, in the sense that it can involve a large set of regions in the brain, it is also a gradual theory of consciousness that can result from minimal neural networks. Indeed, reentry among even a small number of neurons, as far as it reflects differentiated integration, will induce consciousness to a certain degree. We will come back to this theoretical aspect below.

The Global Neuronal Workspace Theory

The Global Neuronal Workspace theory proposed by Stanislas Dehaene with Lionel Naccache and Jean-Pierre Changeux is currently the most explicit and most functional account of the cerebral architecture underlying conscious access. Its functionality has rendered this theory very popular in neuroscience circles. Yet, as we shall see below in the section 'Neurobiological standpoints on the hard problem,' this theory has also been greatly criticized for probing functionality at the price of sacrificing some phenomenological aspects of consciousness. For now, let us focus on the main characteristics of this theory.

This theory is a perfect example of the neurobiological extension of a cognitive theory, that of a global workspace, originally proposed by Bernard Baars in 1988 (see the article 'Cognitive theories of consciousness' in this volume). Dehaene's theory assumes that neurocognitive architecture is composed of two qualitatively distinct types of elements. The first type is represented by a large network of domain-specific processors, in both the cortical and subcortical regions that are each attuned to the processing of a particular type of information. For instance, the occipitotemporal cortex is constituted of many such domain-specific processors, or 'cerebral modules,' where color processing occurs in V4, movement processing in MT/V5, face processing in the fusiform face area (FFA), etc. Although these neural processors can differ widely in complexity and domain specificity, they share several common properties: they are triggered automatically (i.e., mandatorily, by opposition to voluntarily), they are encapsulated (their internal computations are not available to other processors), and importantly, they largely operate unconsciously.

Conscious access involves only the second type of element, namely, the cortical 'workspace' neurons that are particularly dense in the prefrontal, cingulate, and parietal regions. These neurons are characterized by their ability to send and receive projections to many distant areas through long-range excitatory axons, breaking the modularity of the nervous system and allowing the domain-specific processors to exchange information in a global and flexible manner. The global workspace is thus a distributed neural system with long-distance connectivity that can potentially interconnect multiple cerebral modules through workspace neurons. It offers a common communication protocol, by allowing the broadcasting of information to multiple neural targets.

One important aspect of this theory is that encapsulation and automaticity are less rigid than traditional modularist (i.e., Fodorian) accounts of the cognitive system (see the article 'Cognitive theories of consciousness' in this volume). Indeed, once a set of processors have started to communicate through workspace connections, this multi-domain processing stream also becomes more and more 'automatized' through practice, resulting in direct interconnections, without requiring the use of workpace neurons and without involving consciousness. Another important aspect that follows from this theory is that information computed in neural processors that do not possess direct long-range connections with workspace neurons, will always remain unconscious. This idea is rooted in the work of Crick and Koch, postulating that neural activity in V1, because it does not project toward prefrontal neurons, does not participate directly in visual consciousness.

Note that for a mental object to become conscious, it is not sufficient that its activity gives input to the global workspace. Two other conditions have to be met. One is that the content of the mental object must be represented as an explicit firing pattern of neuronal activity, that is, a group of neurons that unambiguously indexes its relevant attributes. A final and important condition is that the top-down amplification mechanisms mobilizing the long-distance workspace connections render the content of consciousness accessible, sharpened, and maintained. A mental object, even if it respects the two first conditions (explicit firing and accessibility to workspace neurons) will still remain buffered in a 'preconscious' (and thus a nonconscious) store until it is attended to and its neural signal amplified. Therefore, top-down attention in this framework is a necessary condition for consciousness. Whether being conscious requires top-down attention constitutes, as we will see below, one of the most debated aspects among neurobiological theories of consciousness.

The Duplex Vision Theory

The Duplex Vision theory proposed by David Milner and Melvyn Goodale postulates that visual perception involves two interconnected, but distinctive pathways in the visual cortex, namely, the dorsal and the ventral stream. After being conveyed along retinal and subcortical (i.e., geniculate) structures, visual information reaches V1 and then involves two streams. The ventral stream projects toward the inferior part of the temporal cortex and, according to this theory, serves to construct a conscious perceptual representation of objects, whereas, the dorsal stream projects toward the posterior parietal cortex and mediates

the control of actions directed at those objects. Apart from these structural considerations, the two streams also differ at the computational and functional levels. On the one side, the ventral stream conveys information about the enduring (i.e., long-lasting) characteristics that will be used to identify the objects correctly, and subsequently to link them to a meaning and classify them in relation to other elements of the visual scene. Computing these enduring characteristics involves relatively long and costly computations. On the other side, the dorsal stream can be regarded as a fast and online visuomotor system dealing with the moment-to-moment information available to the system, which will be used to perform actions in real time.

It should be noted that this dissociation between ventral and dorsal pathways has sometimes been misunderstood as equivalent to conveying the 'what' and 'where' information in the visual cortex, respectively. However, structural and spatial attributes can conjointly be used by both systems. For instance, visual information such as the size, geometrical structure, and location of a target object might be computed both by the dorsal steam, in order to grasp and reach the object, and by the ventral stream in order to segregate the object from a complex visual scene. While the ventral steam involves object perception by comparison with visual perceptual attributes stored in memory, the dorsal steam, because it is fast and updated in real time, involves no storage of the visuomotor attributes extracted from the object, nor the motor program resulting from actions upon that object.

In other terms, while the ventral stream can be seen as conveying vision for perception, the dorsal stream is concerned with vision for action. Both systems work together in the production of adaptative behavior, but they can also function independently as revealed by clinical studies. Indeed, the development of this theory mainly results from neuropsychological investigations, demonstrating a double dissociation between vision for perception and vision for action. On the one side, patients with a lesion in the posterior parietal cortex, the area terminating the visual dorsal stream, suffer from 'optic ataxia,' a deficit in the control of reaching and grasping objects. Despite this deficit, these patients are perfectly able to verbally describe the unreachable objects, either as a whole or in terms of their attributes. In other terms, they can use vision for consciously perceiving objects in their environment, but not for controlling real-time actions directed at those objects. On the other side, patients with damage to a ventral region known as the lateral occipital complex are unable to recognize everyday objects or even simple geometric forms, a deficit labeled 'visual form agnosia.' Yet, such patients are strikingly efficient at grasping objects correctly (for instance by opening their hand as a function of the object size, or by rotating their hand according to the object orientation). Although such patients cannot consciously perceive the object and even its visual attributes (size, shape, or orientation), they can use the same object attributes to control their object-directed actions.

One important consequence of this theory, according to Milner and Goodale, is that because the processes performed by the dorsal stream are very fast and largely automatic, they are largely unconscious, while, by contrast, those conveyed by the ventral stream are assumed to constitute the core of conscious perception. The visual phenomenology generated in ventral regions will in turn be transferred to working memory components in order to use information off-line, when the objects are not stimulating the visual system anymore. According to this theory, although we are typically aware of the actions we perform, no phenomenology is associated with the visual information used by the dorsal stream to control those actions. Hence, the neural computations performed in the dorsal stream remain quite inaccessible to consciousness.

It is of note that when this theory was proposed more than a decade ago, evidence from binocular rivalry (in which conscious perception alternates between two images, say a face and a house presented to each eye) revealed a very strong correlation with conscious perception in the ventral stream (e.g., in the FFA for faces). Therefore, this evidence misled many researchers at that time to deny the possibility of unconscious perception in the ventral stream. However, Geraint Rees and colleagues found that patients with unilateral neglect, a deficit in which they fail to pay attention and then report stimuli on half of their visual field, still exhibit FFA activity for faces in the neglected field. Since neglected stimuli are perfectly reported when cued carefully, neglect is considered as an inability to

report efficiently because of a lack of attention. Given the possibility that consciousness and attention might be distinct (see below), it still remains unclear whether. Rees's finding reflected ventral neural activity without consciousness or just without attention. Yet, more recent evidence with visual masking, obtained in my laboratory has revealed unconscious neural activity in ventral regions, including the FFA, during subliminal face perception, thus clearly demonstrating that the ventral stream is not exclusively related to conscious perceptual processes.

This type of evidence is problematic for the Duplex Vision theory, since this theory predicts that conscious perception should be proportional to neural activity in the ventral stream. Although the possibility of unconscious ventral processing was not taken into account in the original theory, it can be accommodated by assuming a threshold mechanism, as proposed by Zeki for the Minimal Consciousness theory reviewed below. However, including this threshold leads the theory to lose it former appeal, since consciousness is 'only partially' correlated with activity in the ventral stream. Conversely, various recent evidences have shown that the dorsal stream can, under some circumstances, be associated with the consciousness of actions. Thus, although this theory might be effective for distinguishing the neural mechanisms that are responsible for vision for perception and vision for action, this dissociation might turn out to be orthogonal to the dissociation between conscious and unconscious processing.

The Local Recurrence Theory

The Local Recurrence theory put forward by Viktor Lamme is mostly concerned with vision for perception rather than action. It distinguishes between three hierarchical types of neural processes related to consciousness. The first stage involves a 'feedforward sweep' during which the information is fed forward from striate visual regions (i.e., V1) toward extrastriate areas as well as parietal and temporal cortices, without being accompanied by any conscious experience of the visual input. Only during the second stage, involving the 'localized recurrent processing,' is the information fed back to the early visual cortex. It is these recurrent interactions between early and higher visual areas which, according to this theory, lead to visual experience (i.e., phenomenal consciousness, see below). The third and final stage consists of 'widespread recurrent processing,' which involves global interactions (similar to the global workspace model) and extends toward the executive (i.e., the frontoparietal network) and language areas. This final step also involves the attentional, executive, and linguistic processes that are necessary for conscious access and reportability of the stimulus.

An interesting aspect of the Local Recurrence theory is that it offers an explanation for the difference between conscious and unconscious perception in mechanistic rather than in architectural terms. Here, the distinction between subliminal and conscious perception involves the same regions. Subliminal perception reflects the fact that the visual signal is fed forward to higher visual areas without being fed back to early areas (for instance a second stimulus replaces the first one in V1 and then prevents the setting up of recurrence between higher regions and V1). Another interesting aspect of this theory, although provocative and speculative, is that consciousness should not be defined by behavioral indexes such as the subject's introspective reports. Instead, according to Lamme, one should rely on neural indexes of consciousness, one of which is neural recurrence. Indeed, Lamme is concerned with the question of defining phenomenological consciousness when a report is impossible. According to Lamme's theory, involvement of the second step (local recurrence) without involving the third step (global recurrence) represents exactly this situation. We will come back to this aspect when discussing the neurobiological standpoints on the hard problem.

In sum, the theory of Local Recurrence explicitly stipulates that recurrent activity is sufficient for consciousness. Yet, one main difficulty with this theory is that is fails to take into account the recurrent connections that exist between regions that are not associated with consciousness (for instance between V1 and the thalamus). It remains possible that consciousness involves local recurrence 'between some specific cerebral components.' However, local recurrence cannot then be considered as a sufficient condition for consciousness anymore, since it requires the involvement of

additional mechanisms for explaining why it only applies to a restricted set of brain regions.

The Microconsciousness Theory

The microconsciousness theory put forward by Semir Zeki and colleagues is arguably the most localist account of the cerebral architecture underlying consciousness. It is assumed in this theory that instead of a single consciousness, there are multiple consciousnesses that are distributed in time and space. This theory has also been mainly developed in the context of vision research and reflects the large functional specialization of the visual cortex. For instance, evidence from various sources (clinical, brain imaging, etc.) have shown that while the perception of colors is associated with neural activity in area V4, motion perception reflects neural activity in MT/V5. In particular, neuropsychological investigations have demonstrated that the respective cerebral lesions in these two functional sites lead to dissociated forms of conscious perception. Lesions in V4 lead to 'achromatopsia,' the inability to see the world in colors, while motion remains intact, and conversely lesions in MT/V5 result in visual 'akinetopsia,' the inability to perceive visual motion, while color perception remains unaffected. Furthermore, because of the existence of direct connections between subcortical structures and MT/V5 (i.e., without having to be mediated by V1), patients with a lesion in V1 are unable to perceive objects, while they can still experience motion when these objects are moving.

Zeki takes these findings as evidence that consciousness is not a unitary and singular phenomenon, but rather that it involves multiple consciousnesses that are distributed across processing sites (also called essential nodes in this theory), which are independent from each other. Another form of evidence in favor of this theory is that the conscious perception of different attributes is not synchronous and can respect a temporal hierarchy. For instance, psychophysical measures have shown that color is perceived a few tens of milliseconds before motion, reflecting, according to Zeki, the fact that neural activity during perception reaches V4 before reaching MT/V5. This observation is congruent with the microconsciousness theory, where it is postulated that microconsciousnesses are not only distributed in space, but also in time.

A critical characteristic of the microconsciousness theory is that it considers these processing sites to be equivalent to perceptual sites, which means that conscious perception of one specific attribute (e.g., color) is proportional to the strength of activity in its respective processing site (i.e., V4). According to this theory, a microconsciousness associated with a processing site does not necessitate top-down influences from higher (i.e., frontal) areas although these regions might play a role in visual consciousness. Note that although the correlation between neural activity in a processing site and conscious perception is predicted to be highly positive, it cannot be perfect in this theory. Indeed, in order to take into account some clinical and experimental evidences revealing unconscious processing, notably those obtained by Zeki himself, neural activity in a processing site must reach a certain height for a conscious correlate to be generated. Therefore, this theory does not deny the existence of unconscious processing. In fact, a consequence of this postulate is that it obviates the need to separate brain regions that are linked to consciousness from those that are linked to unconscious processing, as we saw above if the Local Recurrence theory. The transition between unconscious and conscious perception reflects the crossing of a threshold 'within' the processing site, though the neural and behavioral characteristics of this transition remain to be specified.

Note that this theory does not deny that the attributes of each processing site are bound together at one point. However, it assumes that binding is a post-conscious process occurring 'after' consciousness of the specific attributes to be bound together has taken place. This second step is termed 'macroconsciousness.' Although microconsciousness involves only one perceptual attribute (e.g., color), macroconsciousness reflects the phenomenal experience associated with the bound object (i.e., including its form, color, motion, etc.). It occurs higher and later in the hierarchy, and depends upon the presence of the previous one. In addition, Zeki proposes that there is also a third and last level coined as 'unified consciousness,' reflecting a more global form of consciousness,

which involves linguistic and communication skills. This third level remains largely unspecified, but it is roughly equivalent to the brain regions leading to consciousness in the global workspace model.

One main difficulty with this theory is that any processing region in the brain should, at first glance, constitute an NCC in the multiple-consciousness theory. As such, it remains unclear why conscious perception is not associated with activity in most brain regions, including the cerebellum and subcortical regions, especially those conveying visual signals (e.g., the Lateral Geniculate Nucleus). Zeki takes this problem into account, at least by admitting it, and tries to solve it by explaining that neural activity in a processing site probably also necessitates the involvement of additional systems, in particular the reticular activating system maintaining arousal. Yet, the theory loses its force, since neural activity, in certain perceptual sites, is not a sufficient condition for microconsciousness anymore (not mentioning the fact that the reticular system must be involved during unconscious processing, since it also influences the subcortical regions). In addition, another difficulty for the Multiple Consciousness theory is that visual regions can lead to the binding of several attributes in the absence of consciousness, as shown both by Dehaene and colleagues using visual masking, and by Wojeiulik and Kanwisher in a patient with a bilateral parietal damage suggesting that the binding mechanisms that are supposed to lead to macroconsciousness can in fact operate in the absence of consciousness. Therefore, empirical evidence contradicts the claim that binding has to be a post-conscious process. Finally, and maybe more crucially, the inclusion of a threshold mechanism between unconscious and conscious processing, and hence the possibility of unconscious processing, is far from being obvious in this framework. This constitutes, as we saw above for the Duplex Vision theory, an empirical constraint (i.e., there has to be a threshold even if its nature remains unspecified) that implies a radical change in the main message conveyed by these two theories (i.e., consciousness is 'only partially' correlated with activity in the ventral/perceptual sites). In particular, it then becomes difficult to understand the frequent claim by Zeki that processing sites in the visual brain are also perceptual sites.

Neurobiological Standpoints on the Hard Problem

Now that we have seen the main characteristics of these competitive theories individually, the next sections will describe how they face (or deny) conceptual issues related to the hard problem of explaining conscious experience.

Relation to Access and Phenomenal Consciousness

The neuroscientific and philosophical issues of consciousness have never been so closely related. A consequence of this close interplay is that conceptual issues raised by philosophers are progressively influencing neurobiological theories. Arguably, the most influential issue in recent years has been the potential distinction between phenomenal and access consciousness proposed by Ned Block. In short (and in the context of visual perception), this dissociative approach assumes that the NCC for phenomenal consciousness reflects the qualitative subjective experience (i.e., the qualia) associated with the percept. According to Block, the phenomenological richness that one can experience when seeing a complex visual scene goes far beyond what the observer can report. Hence, conscious access is assumed to reflect another NCC that is more linked to reportability, stimulus discrimination, introspection, etc. This distinction has been appealing to many neuroscientists in recent years, and several research programs are currently investigating the possibility to dissociate different NCCs for these two forms of consciousness.

An important contrastive feature regarding the five theories reviewed above is whether they accept or reject this dissociative approach. It turns out that all of them, except the Global Neuronal Workspace theory, have recently acknowledged or even incorporated this distinction. In the Reentrant Dynamic Core theory, the distinction between phenomenal and access consciousness is analogous to the one between primary consciousness and higher-order consciousness, respectively. In a recent extension of the Duplex Vision theory, neural activity in the ventral stream has been directly associated with phenomenal consciousness by Goodale. Similarly, Zeki has recently

linked micro- and macroconsciousness with the phenomenal consciousness of specific attributes (colors, contrasts, etc.) and bound objects, respectively, while unified consciousness is analogous to access consciousness. In the Local Recurrence theory, phenomenal consciousness has been a very central concept at the origin of the construction of this framework. It is assumed to be a by-product of local recurrent loops, while access consciousness involves additional widespread loops between posterior and anterior regions of the brain. Importantly, all these theories assume that the NCC for access consciousness involves more or less the network of brain regions constituting the workspace in Dehaene's theory. Therefore, the Global Neuronal Workspace theory is often considered as a restrictive theory of access consciousness, and has been criticized for confounding the subjective experience of consciousness with the 'subsequent' executive processes that are used to access its content.

Dehaene and his colleagues not only reject this dissociation in terms of two NCCs, but they also put doubts on its psychological reality. In particular, they deny the possibility of phenomenal consciousness without access and consider this dissociation to be flawed for at least two important reasons. First, following Larry Weiskrantz, they assume that reportability is the only reliable index of consciousness and thus stimuli that cannot be reported/accessed are by definition not conscious in any way. Second, following Kevin O'Regan and Alva Noë, they consider that apparent cases of a rich phenomenal experience without access actually reflect the so-called illusion of seeing. For instance, in 'change blindness' situations, observers are usually overconfident about their capacity of seeing an entire visual scene, while actually, they fail to notice when important modifications occur at unattended locations. As such, the illusion of phenomenal richness reflects the fact that observers think that they can see more than they actually do. Because of this overconfidence, it remains unclear how one can empirically probe observers to describe the content of unattended or inaccessible perceptual events. As one can see here, this conceptual issue of phenomenal consciousness without access is closely linked to the empirical issue of consciousness without attention. This latter issue is addressed in the next section.

Relation to the Distinction between Attention and Consciousness

Although attention and consciousness have long been considered to be similar, if not identical phenomena, it is largely acknowledged today that they can be dissociated. Yet, a double dissociation remains to be demonstrated. Indeed, since the seminal study by Naccache, Blandin, and Dehaene showed that top-down attention modulates subliminal priming, there is now a general consensus regarding the fact that attention is independent of consciousness and that subjects can attend to objects even if they do not consciously perceive them. However, the reverse dissociation where consciousness is itself independent of attention remains highly debated. More precisely, the possibility that subjects can be conscious of objects without any attention is at the center of a very intense controversy.

Neurobiologists acknowledging Block's dissociative approach actually assume that top-down attention is driven by workspace regions and correspond to a central mechanism, which, along with language and working memory, define access consciousness. For instance, in the case of visual perception, the signal in the posterior visual regions (i.e., in the occipitotemporal cortex) is assumed to be attentionally amplified in a top-down fashion by the same widespread parietofrontal areas as the workspace regions in Dehaene' theory. Thus, similarly to the possibility of phenomenal consciousness without access, all the neurobiological theories seen above, except that of the Global Neuronal Workspace, acknowledge the possibility of consciousness without attention. Some *a priori* support for the dissociative approach can be found in studies showing that visible (i.e., supraliminal) though unattended stimuli do not involve workspace regions, but rather an increase of activity in posterior regions, as I have recently evidenced myself. Yet, as we will see below, it remains extremely difficult to demonstrate that there is any form of consciousness for visible but unattended stimuli.

Here also, Dehaene and colleagues disagree with this dissociation and consider, on the contrary, that attention embodies conscious processing. According to them, only attended stimuli can be reported and are thus consciously

processed: demonstrating that an observer is conscious of an unattended stimulus, without relying on any sort of report by the subject, seems extremely difficult or even impossible. In other words, it appears that in order to assess consciousness of the stimulus, one necessarily needs to direct the observer's attention on the stimulus, leading to the conclusion that consciousness without attention is an illogical possibility! It is of note that a few attempts, notably by Koch and colleagues, have been made to demonstrate consciousness without attention. In particular, they have relied on situations termed 'near-absence of attention' and where a stimulus is presented in the periphery while the subject is performing the task on a central target. Under these conditions, Koch and colleagues have found that subjects can still consciously perceive the peripheral stimulus, at least indistinctively. Yet, unfortunately, it remains impossible to demonstrate that there has not been any residual attentional amplification in this situation (additionally, the mere fact that this situation is called 'near-absence of' and not 'full-absence of attention' appears to be highly symptomatic and suggestive of the difficulty to demonstrate consciousness without attention).

Importantly, the majority of neurobiologists including Dehaene and colleagues, acknowledge that the processing of supraliminal but unattended stimuli has a special status, in-between the first step of subliminal processing, and the last step of conscious access. However, they totally disagree on a crucial question which is, 'whether there is any form of consciousness associated with this intermediate level of processing or rather whether it is just another form of nonconscious perception.' Proponents of the Global Neuronal Workspace theory have been very explicit on this issue and consider that this intermediate level involves just another, more elaborated form of nonconscious processing. They termed it the 'preconscious' stage. According to them, a stimulus reaching this preconscious state becomes potentially accessible (it could quickly gain access to conscious report if it was attended), but it is not consciously accessed at the moment, mainly because attention is maintained away (e.g., by processing a concurrent stimulus as in situations of attentional blink, inattentional blindness, or change blindness).

It is interesting to note that since the recent inclusion of the preconscious stage in the Global Neuronal Workspace theory in 2006, it is becoming harder to distinguish it from the Local Recurrence theory on a purely structural basis. Indeed, a tripartite taxonomy is provided in both accounts. The initial stage reflects local feedforward activity at the neural level and corresponds to unconscious (i.e., subliminal) processing, while the third stage involves a global form of recurrence in the brain and reflects conscious access. Yet, the intermediate stage is also structurally analogous in these two theories, as it involves a local form of recurrence in both cases. Rather, it is on the psychological dimension that these two theories diverge, Dehaene considering this intermediate level to reflect a nonconscious stage, while Lamme takes it as reflecting phenomenal consciousness before access. This shows us how much neurobiological theories are still dependant on psychological (i.e., subjective) indexes of consciousness.

In sum, apart from proponents of the Global Neuronal Workspace theory, there is an increasing tendency in most neurobiological accounts to consider that there is an intermediate level associated with phenomenal consciousness without access. Consequently, new indexes that do not rely on subjective reports and that mainly reflect neural mechanisms are being proposed. For instance, Lamme proposes that consciousness should be indexed by the observation of local recurrent loops. In Edelman and Tononi's theory, consciousness is indexed in terms of neural complexity reflected by differentiated integration. Before addressing these approaches in the next section, it is necessary to remind the reader that the dissociation between consciousness with and without attention, similarly to the distinction between access and phenomenal consciousness, remains highly difficult to tackle experimentally. This difficulty is mainly due to the fact that subjective reports, which necessarily involve access/attention components, remain so far the best index of whether someone is conscious. Unless a more reliable index of consciousness is found, these dissociations might possibly turn out to be immune to scientific investigation. Here also, maybe we should wait and see; or maybe new and less constrained epistemological directions should be explored.

Relation to Neurocognitive Panpsychism

As we have seen in the previous section, there is an increasing tendency to consider that subjective reports cannot be trusted when indexing consciousness. Among the theories we have seen above, proponents of two theories in particular, that of Local Recurrence and that of the Reentrant Dynamic Core (and in particular the resulting Information Integration theory put forward by Tononi), have proposed alternative indexes that are supposedly more reliable than subjective reports. Both theories consider consciousness as an emergent property of any system sharing specific core mechanisms of either recurrence or differentiated integration, respectively. Therefore the most reliable index of consciousness should, according to these accounts, reflect the quantification of these core mechanisms. In particular, these neural indexes are assumed to offer a more reliable estimation of phenomenal experience, well above the poor estimation provided by subjective reports.

This approach can be appealing since it could potentially lead to a quantification of consciousness in nonhuman species, in preverbal infants, and in artificial intelligence systems. Yet, it leads to a form of panpsychism (see Glossary) that one can term as 'neurocognitive panpsychism.' Both the Local Recurrence theory and the Reentrant Dynamic Core theory are leading to a similar, though more restrictive panpsychist views: any system in the world can be conscious as long as it shares information following the core mechanisms that are specific to these theories. They constitute neurocognitive versions of panpsychism, because these theories are primarily aimed at offering neurobiological accounts of consciousness, while they can also be extended to any cognitive system in the world that shares the same mechanical properties for information-processing. As such, both theories follow the principle that consciousness arises when information is transmitted among neurons interacting along the core mechanisms of these theories.

Yet, there are several conceptual difficulties with these theoretical approaches (the term conceptual difficulties and not impossibilities is preferred here because none of these issues is impossible to resolve *per se*; however, they might turn out to be extremely difficult to demonstrate as well). A first conceptual difficulty is that this doctrine leads to the conclusion that any system displaying these specific core mechanisms will experience consciousness, whatever implementation supports it and, more importantly, whatever its size. This fact relates to Chalmers' argument that even a system as small as a thermostat, because it recurrently shares information between a temperature sensor and a controller for switching the heater on or off, will experience consciousness to some degree. Indeed, the theory of Local Recurrence leads to the fact that any two neurons in interaction would be sufficient for consciousness. The Reentrant Dynamic Core theory would require a few more neurons, but not that many to generate consciousness (precisely nine neurons would be sufficient according to Herzog, Esfeld, and Gerstner). Although such possibilities can hardly be totally rejected, it remains as unclear how they could ever be demonstrated.

A second main issue lies in the paradox of trying to prove that neural indexes are more respectable because they supposedly probe phenomenal and not access consciousness. Indeed, although it is clear that neural indexes offer interesting possibilities when report is impossible (for instance in cases of locked-in syndromes or for prelinguistic babies), they still cannot be taken as reflecting more than conscious access. This principle follows from the fact that neural indexes of any sort have to be validated by confronting them at one point with some kind of report, hence with access and not phenomenal consciousness. For instance, demonstrating that recurrent processing is sufficient for consciousness initially requires consciousness to be indexed by probing whether the system is indeed conscious, and the best way to do so is still to rely on subjective reports in humans. Then, a neural index can only be validated when a correlation has been established with subjective reports. Assuming that one can demonstrate that recurrence, even among two neurons, correlates with some degree of consciousness, this can be achieved only through report and that degree of consciousness will only reflect a partial form of access, not phenomenal consciousness. Once again, the interesting aspects of this approach cannot be neglected, since it can allow to subsequently probe whether the core mechanisms represented by this

index can be found in systems or species that cannot report, suggesting that they 'might probably' be conscious. However, neural indexes cannot be taken as reflecting a superior or richer phenomenal form of consciousness since they can only be validated through reports. In other terms, these theories paradoxically emphasize that their core mechanisms do not reflect access consciousness, while actually only this form of consciousness can be used to demonstrate their validity.

A third conceptual issue is that even if it turns out that one of these neural indexes turns out to be perfectly correlated with consciousness, and thus becomes a perfectly reliable measure of consciousness, then one might still ask whether we have made any progress. This issue is less problematic and might turn out to be resolved in the future. However, it is worth mentioning, since it remains unclear whether we are searching in the right direction. Indeed, although neural indexes might turn out to be very useful in many circumstances, they do not escape the limitation of NCCs. They would still primarily inform us on whether someone (or something) is or is not conscious. Yet, they would not directly help us understand why they are conscious. If it turns out that recurrence or differentiated integration are sufficient for consciousness, we would still have to accept that consciousness is 'emerging' from the brain each time these core mechanisms come into play. This issue is more moderate because once we have found a perfect neural index of consciousness, it might get us closer to the explanatory gap. Yet, it remains unclear how it will allow us to cross that gap.

Conclusion

By the end of this article, one might wonder where we stand with regard to the original issue of how the brain leads to consciousness. Over the last two decades, dozens of theories of the NCC have been proposed (leading David Chalmers to ironically speak about an 'NCC zoo') and a kind of trial-and-error strategy has been applied to evaluate them, and in most cases, to reject them one by one. Here we have focused more specifically on five recent neurobiological frameworks that are the most popular accounts in this field of research. We have seen that, actually, consciousness can hardly be restricted to neural activity in the ventral stream, nor is it simply reflecting recurrent activity or minimal activity in processing sites. It remains unclear whether consciousness reflects or does not reflect differentiated integration within a dynamic core, as emphasized by Edelman and Tononi's speculative, but powerful theory, or whether it simply reflects activity in workspace regions, as outlined by Dehaene's well-specified, but apparently restrictive account of conscious experience. One main difficulty in disentangling this issue and contrasting these theories, is that it remains unknown whether neurobiological frameworks should be restricted to conscious access and subjective reports, or whether they should extend to other (i.e., inaccessible, unattended, phenomenal) forms of consciousness. Therefore, future research will unavoidably be bound to decide whether neurobiological accounts should take into account or reject the hard problem of understanding a subjective form of conscious experience that cannot be simply defined as conscious access. Thus, facing the hard problem will probably not be any more a main issue on its own. Instead, it will be the problem of deciding whether there is a hard problem at all, leading to a 'super-hard problem' of consciousness and obviously complicating the whole issue. Whether we should stay focused on variants of the NCC until the solution pops out and reduces the hard problem to an easy one, or whether we actually need new laws (of physics, life, or anything else) to cross the explanatory gap, only time will tell. Yet, although none of the neurobiological theories provided so far have been able to provide a 'conclusive' explanation regarding this matter, the search for a neurobiological explanation of consciousness still constitutes one of the most exciting challenges of contemporary science.

See also: The Neurochemistry of Consciousness.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. New York: Cambridge University Press.
Block N (2005) Two neural correlates of consciousness. Trends in Cognitive Science 9: 46–52.

Chalmers D (1996) The Conscious Mind. New York: Oxford University Press.
Crick F and Koch C (1990) Towards a neurobiological theory of consciousness. Seminars in Neuroscience 2: 263–275.
Dehaene S, Changeux JP, Naccache L, Sackur J, and Sergent C (2006) Conscious, preconscious, and subliminal processing: A testable taxonomy. Trends in Cognitive Sciences 10: 204–211.
Dehaene S and Naccache L (2001) Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. Cognition 79: 1–37.
Edelman GM and Tononi G (2000) A Universe of Consciousness: How Matter Becomes Imagination. New York, NY: Basic Books.
Gallese V (2007) The ''conscious'' dorsal stream: Embodied simulation and its role in space and action conscious awareness. Psyche 13(1): 1–20.
Goodale M (2007) Duplex vision: Separate cortical pathways for conscious perception and the control of action. In: Velmans M and Schneider S (eds.) The Blackwell Companion to Consciousness, pp. 616–627. Oxford: Blackwell.
Koch C (2004) The Quest for Consciousness: A Neurobiological Approach. Denver, CO: Roberts.
Koch C and Tsuchiya N (2007) Attention and consciousness: Two distinct brain processes. Trends in Cognitive Sciences 11: 16–22.
Kouider S and Dehaene S (2007) Levels of processing during non-conscious perception: A critical review. Philosophical Transactions of the Royal Society of London B 362: 857–875.
Lamme VA (2006) Towards a true neural stance on consciousness. Trends in Cognitive Sciences 10: 494–501.
Tononi G (2004) An information integration theory of consciousness. BMC Neuroscience 5: 42.
Zeki S (2007) A theory of micro-consciousness. In: Velmans M and Schneider S (eds.) The Blackwell Companion to Consciousness, pp. 580–588. Oxford: Blackwell.

Biographical Sketch

Sid Kouider is a cognitive neuroscientist working at the Ecole Normale Supérieure (Paris, France) on the neurobiological and psychological foundations of consciousness. His work focuses on contrasting conscious and unconscious processes, both at the psychological and neural level, using various behavioral and brain imaging methods. Recently, he extended this line of research to study the neural correlates of consciousness in prelinguistic babies.

The Neurochemistry of Consciousness

J Smythies, University of California, San Diego, CA, USA

Glossary

AMPA – Alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid
CaMK11 – Alpha Ca^{2+}-calmodulin-dependent kinase 11
EPSP – Excitatory post-synaptic potential
ERK – Extracellular signal-regulated kinase
GABA – Gamma-amino butyric acid
GHB – Gamma-hydroxy butyric acid
GIRK – G-protein mediated inwardly rectifying K^+
JaK – Janus kinase
LDTN – Lateral dorsal tegmental nucleus
LSD – Lysergic acid diethylamide
MAP – Microtubule-associated protein
NMDA – *N*-methyl D-aspartate
Oct-6 – Oct-6 is a transcription factor that belongs to the POU transcription factor family (class-3 subfamily)
PPN – Pedunculopontine nucleus
PSD – Postsynaptic density
S-100 – A brain protein
SAP – Synapse-associated kinase

Introduction

The term consciousness is commonly used in three different ways. One is the clinical sense in which we distinguish a person in coma from a normally alert being. The second sense (as used e.g., by Crick, 1998) is awareness, or alertness that is closely allied to attention. The third is the phenomenal sense, in which we refer to the totality of a person s sensations, images, emotions, and thoughts, available to introspection, and studied by introspectionist psychologists, such as Ramachandran and Gregory. In the first we study neural mechanisms whose malfunction lead to coma. The second covers a broad range of studies. In the third, we study the complex brain mechanisms that construct the structures (including the visual field and the somatic sensory field also known as the body image) that compose a person s phenomenal consciousness.

Clinical Consciousness

Coma can be induced by a wide variety of agents and diseases. Among the most interesting, from the neurochemical point of view, are general anesthetics, and the remarkable effect induced by the combination of desferrioxamine and prochlorperazine.

General Anesthetics

Over a century ago Meyer and Overton proposed that anesthetic agents act by interacting with lipophilic components of cell membranes. Today it is still recognized that lipophilicity is important, but attention has now focused on specific receptors. Most anesthetics act at multiple molecular sites. The most prominent action is enhancement of GABA(A) receptor-mediated neuronal inhibition, in both cortical and subcortical regions, with selectivity for receptors that contain beta2 and especially beta3 subunits. This can be modulated by simultaneous inactivation of structures, such as the median septum, and ventral tegmental area, that normally activate the cortex, or that mediate the hippocampal output, such as the nucleus accumbens and ventral pallidum.

Every general anesthetic acts on at least one type, and in some cases several types, of ligand-gated ion channels. (The family of such channels composes nicotinic, GABA(A), glycine, serotonin, and glutamate receptors). There is evidence that some anesthetic agents act partly at inhibitory glycine channels, and that some depress excitatory acetylcholine nicotinic receptors. They may target specific sites within the channel lumen (ACh), or on the outer side of the alpha helix lining that

channel (GABA(A) and glycine channels). Some also inhibit MGluR5 receptors. Recently individual amino acid residues have been identified that are necessary for anesthetic function. These are located on the extracellular portions of the alpha subunits of the GABA(A) receptor.

Gamma-hydroxybutyric acid (GHB) is a naturally occurring metabolite of GABA that may produce coma by action at its own receptors (GHBR) and at GABA(B) receptors. The situation is complicated by the fact that GHB can also be metabolized to GABA.

The coma induced by traumatic brain injury is usually attributed to glutamate neurotoxicity, there is also evidence, however, that potentiation of GABA(A) receptor function, via activation of CaMK11, may also be involved. GABA(A) receptors may also be involved in the comas induced by inert gases such as nitrogen.

In contrast, two anesthetic agents ketamine and nitrous oxide act primarily by blocking excitatory NMDA receptors (to induce coma), and by enhancing analgesic opioid mu receptors.

The Desferrioxamine/Prochlorperazine Coma

In 1985 Blake and his coworkers reported the serendipitous discovery in human subjects that a combination of normal doses of the iron-chelating agent desferrioxamine with the dopamine-receptor blocker prochlorperazine induced a deep and long-lasting coma (1 3 days with complete recovery). Neither drug by itself induced any such effect. Catecholamines, such as dopamine, are known to take part in conjunction with iron ions *in vitro* in the complex O Brien cycle. In this, iron ions cycle between the ferric and the ferrous states, and the catecholamine molecules cycle with their quinones. This system acts as an effective dismuter of superoxide, converting five molecules of the highly toxic superoxide per cycle into two molecules of oxygen and three of the less toxic hydrogen peroxide. Thus, the combination of dopamine and iron in the O Brien cycle may provide an essential mechanism in the neuron to protect against superoxide toxicity. Mitochondria convert 5% of their oxygen into superoxide. If iron is removed from the neuron by desferrioxamine, and dopamine is prevented from entering the neuron by prochlorperazine, this may result in the failure of the O Brien cycle allowing neurotoxic levels of superoxide to build up, leading to coma.

This hypothesis needs some further explanation of the role of catecholamines like dopamine in the brain. It is usually thought that such neuromodulators exert their effects on the postsynaptic neuron only by activation of their specific receptors and subsequent postsynaptic cascades. The neurotransmitter molecule itself, in this case dopamine, is thought to stay outside the cell, to be dealt with by reuptake into the terminal, or by further metabolism. However, it is now known that, when a receptor (in this case the dopamine D1R) is activated, it is usually immediately endocytosed carrying the molecule of, in this case, dopamine into the endosome apparatus. Here the dopamine molecule and the protein receptor separate, and the latter is recycled back to the surface, or, if damaged, to the proteosome to be broken down into its constituent amino acids for reuse.

The iron transporter molecule transferrin, on binding a molecule of iron, is also endocytosed into the neuron together with its cargo, and is trafficked to the same endosome to which the dopamine receptor is trafficked. Thus the ingredients of the O Brien cycle free iron, dopamine, and superoxide (from a neighboring mitochondrion) can come into the required close contact, in the endosome, for the O Brien cycle to operate. This cycle may play an important role in the normal cell. Thus catecholamines, including dopamine, may act on neurons, in part, by their direct antioxidant chemical properties, rather than by only triggering postsynaptic cascades. Catecholamines are potent direct free-radical scavengers and antioxidants. Dopamine D2 receptors also promote antioxidant activity by inducing the synthesis of the antioxidant enzyme superoxide dismutase, and the neuroprotective enzymes glutathione synthase, gamma-glutamylcysteine synthesase, glutathione peroxidase, glutathione reductase, and glutathione-S-transferase.

The Cholinergic System

The cholinergic system in the brain plays a vital role in the maintenance of the clinical conscious

state. Brain injury studies show that much of the brain can be disabled, or removed, without affecting consciousness. However, damage to the midbrain often results in coma. The key nuclei involved in this reaction are the cholinergic pedunculopontine nucleus (PPN) and lateral dorsal tegmental nucleus (LDTN) in the pons and medulla. These project widely to all thalamic nuclei, including the reticular nucleus, and to the limbic system, as well as to the higher cholinergic nucleus basalis of Meynert in the basal forebrain. They do not project widely to the neocortex, except for the medial prefrontal cortex. The nucleus of Meynert supplies most of the cortex with cholinergic fibers. Lesions of the nucleus basalis do not commonly lead to coma. Another brain area, lesions of which commonly lead to coma, contains the complex intralaminar nuclei of the thalamus. This sends a glutamatergic projection to the cortex. During waking this generates single action potentials at regular intervals that activate the cortex. In contrast during sleep this changes to burst firing that generates the synchronized EEG pattern typical of sleep. A brain area essential for rapid eye movement (REM) sleep is the nucleus reticularis pontis oralis.

Crick, Llinas, and others have suggested that the correlates of consciousness in the brain are provided by reentrant thalamocortical projections, especially linked to the reticular nucleus. This system is powerfully modulated by the brain stem cholinergic nuclei PPN and LDTN. Crick suggested that the control of the searchlight of attention was mediated by the reticular nucleus of the thalamus. However, this searchlight includes olfaction as well as the other four senses. We can switch our attention to an experienced odor as quickly and easily as we can to a sight or a sound. Now, there are no anatomical connections between the olfactory cortex and the thalamic reticular nucleus. Therefore the basic switch that controls conscious attention may be provided by the midbrain cholinergic nuclei acting in conjunction with the reticular nucleus. Thus the global control of consciousness may be affected by this cholinergic system. The further effects of this system on awareness and phenomenal consciousness will be discussed later.

Awareness/Alertness

The degree of alertness of a person, and the direction of selective attention is controlled by a most complex brain mechanism in which a number of anatomical structures and neuromodulators take part. Amongst these the reticular nucleus of the thalamus plays a key role. It sits sequentially astride all the sensory inflow tracts in the brain, except the olfactory input, and receives glutamatergic input from these. Its output consists wholly of GABAergic neurons whose axons are widely distributed sequentially to other thalamic nuclei. This nucleus also receives extensive projections from the cholinergic PP/LDTN, raphe, and locus coeruleus in the brain stem. Thus it is in a position to modulate much of the sensory input to the brain and direct selective attention.

The neuromodulators associated with alertness and selective attention are acetylcholine, norepinephrine and dopamine, adenosine, histamine, and hypocretins. The currently available evidence suggests a role for noradrenaline in alerting/arousal, whereas acetylcholine may control arousal and selective attention. Dopamine is principally concerned with reinforcement. The role of serotonin is very complex and is dealt with here and below.

Acetylcholine. Collaterals of the sensory tracts in the midbrain enter the reticular formation where they activate the cholinergic PP/LDTN. These nuclei in turn project to the sensory relay nuclei and to the reticular nucleus of the thalamus. In the relay nuclei this input activates neuronal transmission via nicotinic receptors. In the reticular nucleus this input from the PP/LDTN inhibits (via muscarinic type 2 receptors) the inhibitory GABAergic cells of the reticular nucleus, which in turn further promotes relay of sensory information to the cortex. During sleep, activity in the sensory tracts diminishes, and this cerebral activation is reversed. Muscarinic blockade by scopolamine leads to an altered state of consciousness (twilight sleep). Nicotine administration leads to improved attentional performance in a number of tests (e.g., the Stroop test). It has been suggested that an important role of acetylcholine related to consciousness is to modulate the intradendritic cytoskeletal structure of cortical pyramidal cells by a mechanism that includes MAP-2.

Norepinephrine (NE) and dopamine (DA). Both of these neuromodulators promote the alert state: dopamine promotes pleasure as well, having a powerful role in reinforcement mechanisms. Reduction in NE activity facilitates disengagement of spatial attention, and a frontal NE system promotes focused attention by attenuating the effects of distractors. DA plays a key role in shifting attention. It has been suggested that a DA pathway mediates a lower tonic arousal system, whereas an upper phasic NE arousal system is involved in response selection. During slow wave sleep (SWS) the NE neurons in the locus coeruleus are inhibited by an input from GABAergic neurons in the lateral preoptic area. During REM sleep inhibition by a second population of GABAergic neurons in the periaqueductal gray and dorsal paragigantocellular nucleus leads to a complete cessation of NE cell activity.

Serotonin. This neuromodulator promotes waking and inhibits sleep activity. Raphe neurons (as well as neurons that release norepinephrine and histamine) fire steadily during waking, more slowly during SWS and hardly at all in REM sleep. This latter effect is brought about by 5HT activation of long-range GABAergic projection exhibiting 5HT 2A/2C receptors that inhibit the cholinergic cells of the PP/LDTN.

Histamine. Histamine increases alertness and inhibits both SWS and REM sleep. Histaminergic cell bodies are located in the tuberomammillary nucleus in the posterior hypothalamus, and project widely to cortex, basal forebrain, thalamus, and tegmentum. Histamine, acting on H1 receptors on relay cells in thalamic relay nuclei, changes their firing from burst to single tonic firing, thus promoting transmission of information. H2 receptors facilitate cortical and hippocampal activity.

Hypocretins/orexins. Two polypeptides, hypocretin 1 and 2 (also known as orexin A and B), are produced by cells in the lateral and posterior hypothalamus. They stimulate wakefulness by binding to their G-protein linked receptor. Defects in their action lead to narcolepsy.

Adenosine. Adenosine is thought to be a sleep-inducing factor. During sleep deprivation extracellular levels of adenosine rise steadily in the basal forebrain and inhibit the basic forebrain cholinergic neurons that promote wakefulness. Caffeine, which promotes wakefulness, acts as an adenosine receptor inhibitor.

Phenomenal Consciousness

This section discusses the neurochemical factors related to the processes of constructing the elements of phenomenal consciousness sensations, images, emotions, mood, and thoughts.

The Serotonin System

Psychedelic drugs, such as mescaline and lysergic acid diethylamide (LSD), produce remarkable changes in conscious perception and cognition well described by Aldous Huxley in *The Doors of Perception.* Their principal mode of action is as partial agonists at serotonin 2A receptors (with some involvement of 2C receptors). In rat brain, LSD greatly increases 5-HT(2A)R-mediated Fos-like immunoreactivity in some limbic areas (medial prefrontal cortex, anterior cingulate, central nucleus of the amygdala, but not nucleus accumbens). Activation of 5-HT(2A)Rs by LSD also enhances non-synchronous, late components of glutamatergic excitatory postsynaptic potentials (EPSPs) at apical dendrites in layer V cortical cells.

The cell bodies of all serotonin neurons in the brain are contained in the raphe nuclei of the midbrain and pons (as well as two small adjacent areas the ventrolateral and dorsal reticular formation). There are three raphe nuclei. The dorsal raphe nucleus projects by thin axons mainly to the cortex, the median nucleus projects by thick fibers bearing large varicosities to limited limbic and frontal cortical areas, and to the hippocampus, and the caudal nucleus which projects to the spinal cord. The raphe nuclei supply every portion of the grey matter (with the exception of the superior olivary complex) with a dense network of fine axons that bear many boutons-en-passage. Many of these are extrasynaptic, so that serotonin uses mainly volume transmission. In the case of many transmitters, transmission is mostly synaptic, so that a precise network of connections is maintained. In volume transmission the transmitter molecule is, as it were, sprayed throughout the target cortex, carrying its informational signal as widely as possible at minimal informational cost.

Many raphe neurons also contain glutamate, which is released concomitantly with serotonin upon activation at the synapse.

Serotonin is an important cellular chemical, even very low down in the phylogenetic chain. In plants it plays an important role in photosynthesis. In the brain it has many cellular functions including modulation of neurogenesis, apoptosis, dendritic and spine refinement, cell migration, and synaptic plasticity. It likewise modulates many neuronal functions including autonomic, circadian rhythms, appetite, aggression, sensorimotor activity, sexual behavior, learning, memory, personality, as well as perception and mood.

There are seven main types of serotonin receptors. The ones of current interest in this review are 5-HT(2A)Rs that are located both presynaptically and postsynaptically in all cortical layers except layer IV. They are located mainly on pyramidal cells, particularly on their apical dendrites, on inhibitory neurons, and on axon terminals. In human cortex they are ranked quantitatively cortex > hippocampus > > thalamus, with none in the cerebellum and brainstem.

The effect of 5HT receptors on neurons is very complex, depending on target cell and species. 5-HT(2A)Rs usually activate the target neuron. For example in rat, neocortical layer V pyramidal cells 5-HT(2A)Rs induce EPSPs by both pre- and postsynaptic mechanisms. In mouse, prefrontal pyramidal cells 5-HT(2A)Rs reduce rapidly inactivating $Na^{þ}$ currents. This does not affect somatic action potentials in the target neuron, but has the important effect of reducing the amplitude of Hebbian back-propagated dendritic potentials.

5-HT(2A)Rs depolarize neurons in human neocortical slices, and guinea-pig motoneurons, by inhibiting a resting $K^{þ}$ currant conductance. However, in the rat, 5-HT(2A)Rs hyperpolarize C6 glioma cells by activating an outward $K^{þ}$ current, via increased phosphoinositol hydrolysis, and increased intracellular calcium levels. Serotonin also produces EPSPs in target neurons by promoting the presynaptic release of glutamate from glutamatergic terminals via activation of phosphokinase C. Serotonin, acting on its 2A receptors, also excites neurons by inhibiting inhibitory GABA interneurons, via phosphorylation of the GABA(A) receptor, consequent on activation of phosphokinase C anchored to the scaffolding protein RACK. 5-HT (2A)Rs also release dopamine at its terminals, and inhibits nitric oxide synthase activity induced by cytokines in C6 glioma cells.

In contrast, 5-HT(1A)Rs usually inhibit the target neuron by opening a channel mediating a G-protein mediated inwardly rectifying $K^{þ}$ (GIRK) inhibitory current, and by opening a channel mediating a slowly activated outwardly rectifying $K^{þ}$ current. It also inhibits AMPA-evoked currents by down regulation of $Ca^{2þ}$/calmodulin-dependant PK11 and by inhibiting cGMP responses. Serotonin (2A) receptors utilize a number of postsynaptic cascades including the phosophoinositol pathway, activation of phospolipases, activation of ERK, JaK, and the transporter processes $Na^{þ}$-proton exchanger and $Na^{þ}/K^{þ}$-ATPase.

40 Hz gamma oscillatory behavior in populations of neurons has been linked to the genesis of conscious phenomena. It is therefore of interest that 5-HT(2A) receptors can inhibit oscillations in rat inferior olive nuclei by depressing both T-type $Ca^{2þ}$ currents and a resting $K^{þ}$ current.

The conclusion that can be drawn from this data supports Aghajanian s hypothesis that psychedelic drugs act by promoting glutamatergic neurotransmission in selected areas of the brain.

However, we need to be more specific than that. In the first place, why should increased glutamatergic activity in the visual brain induced by serotonin acting at its 2A receptors result in the marvelous hallucinations witnessed by the subject, who has taken the drug, when agents that act directly at glutamatergic receptors do no such thing? Furthermore, why should the hallucinations be so beautiful? What does it tell us about the neurophysiological basis of esthetics?

The great British neurologist Macdonald Critchley summed it up as follows:

> The usual emotional content of the hallucinosis is best described as one of amazement, awe, interest, and delight. The character of the visions is such as to impress the most prosaic and unimaginative scientific observer in a manner of which no natural beauty or grandeur is capable, almost all writers have insisted that the most skilful pen or brush could not do justice to the marvel of the hallucinations.

It seems most remarkable that a person, who shows in normal life little or no artistic ability, nevertheless has a brain that can produce all this amazing abundance of high art simply because its 5-HT(2A) receptors are stimulated. Psychedelic drugs simply modulate the firing of cerebral neurons by action on mundane receptors, ion channels, and postsynaptic cascades, just as all the other nonpsychedelic neurotransmitters and neuromodulators do. Aldous Huxley was so struck by this paradox that he invoked Henri Bergson s suggestion that one function of the brain might be to limit our contact with reality, acting as some sort of filter keeping certain aspects of reality out of consciousness. Huxley suggested that our inability to see these phenomena under normal conditions is because this part of the mind is under continual inhibition by certain brain mechanisms, and, when the latter in turn are inhibited by drugs such as mescaline and LSD, then these antipodes of the mind become visible. Jung (personal communication) has suggested that the origin of the visions would lie in the collective unconscious, or *mundus archetypus*. An alternative hypothesis is that psychedelic drugs modulate the nonlinear dynamics of holographic brain mechanisms in some as yet undetermined manner.

Visual beauty depends on those subtle properties of color, shape, balance, proportion, and harmony explored by great artists. The very simple line drawings of Matisse, for example, evoke the utmost pleasure in the viewer. A very simple deviation destroys this effect. These particular forms drawn by Matisse evoke in the visual cortex certain patterns of neuronal excitation, that must lead in turn to a discharge in those efferent cortical neurons that project to the pleasure centers of the brain, such as the septal nuclei and the ventral tegmentum. In contrast, geometrically very similar patterns drawn by Tom, Dick, or Harry fail to stimulate these efferent neurons. How this is done, and what evolutionary purpose this reaction serves, remain problems in search of explanations. Similar considerations apply to how neurocomputational mechanisms in the auditory cortex react to Mozart s music by stimulating comparable efferent neurons to limbic pleasure centers.

The acetylcholine system is also involved in type 3 consciousness as muscarinic agonists can induce delirium, which is a global disorder of consciousness. Opiate receptors are also involved in type 3 consciousness as their stimulation induces profound changes in perception and mood. Salvinorin A stimulates only kappa opiate receptors and is one of the most potent psychedelic drugs known.

Acetylcholine, Virtual Reality, and Television Technology

Recently a number of experimental observations have established the important facts, that the visual brain uses virtual reality in many of its functions, and that the neuromodulator acetylcholine plays a major role in this process. For years it was thought that the visual brain works as follows. Photons bounce off physical objects in the stimulus field to land on the retina, there to form a topographic map of the external world. The retina functions as a digital camera that translates this spatial pattern of stimulation of its rods and cones, into a spatiotemporal pattern of excitation of the optic nerve and optic radiation. These patterns are fed to the primary visual cortex where they are processed by a series of parallel neurocomputational mechanisms. The output of these mechanisms is then fed to three separate and parallel higher visual mechanisms that perform neurocomputations relating to shape, color, and movement. This process involves both topographic and vectorial coding. Finally, by a process not at present understood, the final result of these computations are functionally reunited to produce the unified percept that appears in the visual field in consciousness, in which the color is inside the shape and both move together. The fact that there are three separable processes involved is supported by observations of how vision returns after lesions of the occipital lobe. A patient, whose vision is returning after such an injury, at first sees only pure movement, usually rotary, without any objects or color. Then light appears as a pure Ganzfeld a featureless field of white light. Then colors return, first as space or film colors free floating in visual space. This is followed by the appearance of objects, usually in fragments. Finally these fragments join up to form complete objects into which the film colors enter. But, in all this explanation, it was held that the final picture produced in perception in consciousness represented,

as faithfully as it could, what was actually out there at any one moment. It is now known that this assumption is wrong. We actually see, not what is out there but what the brain computes is most probably out there .

The first evidence for this was provided by Ramachandran, who showed that the brain fills in scotomata. A small lesion in the visual cortex will result in a small equivalent blind spot in vision. A natural form of this is the blind spot that we all have over the site where the optic nerve leaves the retina. But we do not experience any such blind spot as such. If we locate the blind spot over a plaid pattern we will see a complete plaid pattern, because the brain fills in the gap from its memory banks. Since then many experiments have been carried out that demonstrate this phenomenon. In one particularly illuminating experiment subjects were shown two different pictures one to each eye, so that retinal rivalry occurred. The first presentation was of two photographs, one to each eye, the first of a group of monkeys and the other was filled with leaves in a tropical jungle. They saw what was out there, that is, monkeys alternating with leaves in retinal rivalry. The experimenters cut up both photographs into irregular bits and then rearranged these to create two pastiches A and B, so that wherever A showed monkeys, B showed leaves, and vice versa. Now, when the subjects looked at A and B in the same way, they did not see the actual two pastiches A and B in retinal rivalry (which was the actual stimulus), but they saw a complete picture of monkeys alternating with a complete picture of leaves. In other words the brain had rearranged the stimuli, and had suppressed the most unlikely pastiches in favor of what it was more familiar with, and hence was more probable a monkey picture and a leaf picture which it created with assistance from its memory banks.

A similar process occurs during a saccade the REM from one visual fixation point to another. Experiments have shown that, during a saccade, we are technically blind. The input from the retina to the cortex is suppressed and the visual field during the transition is filled in from the brain s memory banks. The purpose of this is to avoid the intense vertigo that would otherwise result. The reader can demonstrate this him or herself. Look into a mirror and direct your attention to your left eye. Then do a saccadic movement to look at your right eye. You will notice that you cannot observe any movement of your eyes. Yet, if you stand next to someone else, and look at their eyes in the mirror whilst they are repeating this maneuver, you will notice that you can easily see the movement of their eyes.

This use of virtual reality by the visual system can be explained by reference to television technology, and, in particular, to information compression. Television engineers have found that transmitting all the information picked up by the TV cameras to form the final picture on the TV screen is needlessly expensive and wasteful. This is because, in most TV scenes, only a few targets change much and much of the background does not. The latter can be filled in much more cheaply by probability estimates-based virtual reality, provided by the memory banks of the system. So the art of the TV engineer is to create the optimum mixture of actual events in the studio (expensive but accurate) with fill-ins from memory banks (cheap but progressively more inaccurate). The brain, astonishingly, does the same.

The cholinergic input to the cortex comes preponderantly from the nucleus basalis of Meynert. This has two target areas in the cortex. The first (A) is to layer IV, where it up-regulates afferent (visual) input by means of excitatory nicotinic receptors on layer IV pyramidal cells. The second (B) is to superficial layers, where it down-regulates corticocortical transmission via muscarinic receptors. In the resting state, system B, that carries information from memory banks, is dominant, so that the visual system is tilted towards virtual reality (computationally parsimonious) presentation. Then, any unexpected, or potentially important, new stimulus causes activation of cholinergic neurons in the nucleus basalis. The acetylcholine thus released in the cortex activates the retinal input A (carrying computationally expensive reality) and inhibits the corticocortical input B (carrying virtual reality), so that the new potentially important stimulus can receive full attention.

This convincing evidence that the brain uses television technology has an important impact on philosophical theories of consciousness and perception. Common sense, as well as many philosophers who do not pay much attention to neuroscience, still believe that the theory of perception known as naïve or direct realism is true. In the case of vision this

theory holds that phenomenal objects that which we experience are literally physical objects, or at least the surfaces of physical objects. Sensations, they say, are just the way that such external objects appear to us. The central nervous system is supposed to mediate this process in some unexplained way. In other words vision works something like a telescope, and not at all like television. The rival scientific theory of physiological realism holds, in contrast, that phenomenal objects (our sensations) are literally constructs of the nervous system. In between the retina and the visual field in consciousness lies bank upon bank of neurocomputational mechanisms, that use, we now know, aspects of television technology. In other words, vision works like television, and not at all like a telescope.

The Adrenergic System in the Medulla and the Neurobiology of Bipolar Disorder

The relevance of studies of abnormal mental states (i.e., those found in certain diseases) to attempts to discover the normal function of the brain is based on the fact that this technique belongs to a tried and tested tradition in clinical neurology and neuropsychiatry. A great deal of our understanding of how the normal brain works has come from such clinical studies. The names of Broca, Wernicke, Head and Holmes, Penfield, Ramachandran, and many others come to mind. In this article I will focus particularly on bipolar disorder and schizophrenia.

One important feature of consciousness is the way that, although composed of many different functional components (perception, ideation, emotions, cognition, action), it nevertheless seems often to act as a unity. This is particularly true with regard to the phenomenon of mood. Emotions are transient episodes that usually are short-lived. Moods, in contrast, last longer and color all aspects of mental life. The basic neurobiological function of mood seems to be linked to the fundamental behavioral instincts of approach behavior and retreat behavior. When we are depressed the world looks a gloomy place. We dwell on its bleak sadness. Trivial worries grow to feelings of despair. Our minds become filled with gloomy thoughts. Any action becomes a burden. We tend to retreat from the world. Everything is repulsive. In contrast, when we are in a good mood, the world appears bright and cheerful. If young, our thoughts lightly turn to thoughts of love. We view the growing mortgage with indifference. Everything appears to be for the best in this, the best of all possible worlds. Everything appears attractive.

Furthermore, abnormalities in this system seems clearly responsible for the clinical condition of manic-depressive psychosis (bipolar disorder), in which all the phenomena described above for mood are taken to extremes. In mania all sensations are intensified. Colors are more intense, and everything appears more bright and beautiful with an intense inner meaning (salience). The sense of beauty is vastly increased. Natural scenes become transcendentally beautiful beyond belief. The patient is also flooded with religious feelings and ideas. Everything becomes attractive. These symptoms strongly recall the effect of psychedelic drugs such as LSD. Other symptoms, such as extreme ecstasy and mental clarity, resemble the effects of the dopamine-releaser cocaine. Thus mania resembles the effect of simultaneous over activity in the brain dopamine system and at serotonin 2A receptors. In major depression all these symptoms are reversed to their opposites. In place of ecstasy there is unutterable misery, and unbearable physical malaise. Patients are consumed by a sense of complete isolation from God, from all fellow humans and from the entire world. Everything is guilt, fog, and darkness. The dominant emotion is fear overpowering, paralyzing fear, and despair. Everything becomes repulsive.

So what could be the neurobiological basis of this close synchronization of mood with sensation, emotion, ideation, cognition, and behavior? Mood and emotion are regulated by extremely complex brain (and visceral) mechanisms in which all the neuromodulators take part. This synchronization typical of mood swings could be engineered by multiple interactions between the many nerve networks involved, or it is possible that one component has a greater regulatory role than the others. In other words, the orchestra may have a conductor. One candidate for this position is the adrenergic system in the brain. Very little is known about the functions of this system, which has achieved a Cinderella status amongst the brain s neuromodulators. It is certainly a very primitive system, since its C1 C3 cell bodies are located in a more primitive location the medulla than the

others, which have their cell bodies higher-up in the pons, midbrain, and basal forebrain. Some of its known activities are related to low-level functions, such as cardiac and respiratory control. This has led to its relative neglect. However, it has extensive rostral projections that reach as far up as the thalamus, and, more particularly from our present point of view, to the nuclei of all the other neurotransmitters extensively to the substantia nigra (DA), the locus coeruleus (NE), and the raphe nucleus (5-HT), plus a small projection to the LDTN (cholinergic); as well as extensive projections to key limbic structures, such as the hypothalamus, the central tegmentum, the periaqueductal grey, and the midline paraventricular nucleus of the thalamus (a key center involved in correlation of emotion with cognition). It is difficult to see why a system, that is supposed to control only low-level functions, such as heart rate, respiration, and blood pressure, would need this extensive projection to most of the key higher areas of the limbic system that are so closely involved in all higher mental functions including consciousness. Progress has been hampered by the fact that the usual lesion and stimulation experiments, used to explore function on other brain areas, are difficult in the medulla owing to the proximity of vital centers.

Therefore this system may have some other as yet undiscovered higher function. Thus, the medullary adrenergic system is well placed and well connected to act as the putative conductor of this orchestra, that controls the basic mechanisms of mood, and that waves its baton too excitedly in mania, and too lethargically in depression. Just as the cholinergic PP/LDTN in the pons and midbrain orchestrate consciousness in the brain, so the even more fundamental C1 C3 adrenergic nuclei in the medulla could orchestrate the basic behaviors of approach and retreat, with all their emotional and cognitive accompaniments, that constitute mood. Further research on this neglected system seems indicated.

The Microanatomy of Cognition and Schizophrenia

There is a quantity of data that supports the conclusion that reactive oxygen species (such as hydrogen peroxide and superoxide) and antioxidants (such as ascorbate and glutathione) play key roles in many brain mechanisms involved in neurocomputation, synaptic plasticity, and the brain activities that construct consciousness. The complexity of this system is such that it cannot be described under any simple rubric. But its relation to consciousness can be demonstrated by showing its disorders found in schizophrenia (as well as the desferrioximine/prochlorperazine coma discussed above). Schizophrenia is characterized by hallucinations, delusions, and a general fragmentation of the conscious mind. Until recently it was thought to be caused by the production of some endogenous neurotoxin, or by some imbalance in the brain s neurotransmitter system, such dopamine or glutamate. There is now, however, strong evidence that schizophrenia is due to a disorder of synaptic plasticity in the brain, in particular damage to the neuropil. The neuropil consists of the fine network of axons, dendrites, and dendritic spines that fill the space in the brain between neuronal and astrocytic cell bodies. In many cases of schizophrenia, particularly in type 2 cases, the volume of the neuropil is reduced by up to 50%. This means that the communications between neurons, on which neurocomputation depends, has been severely jeopardized. Computer modeling has shown that, in a complex computational network, if the quantity of interunit connections is reduced by this amount, the whole computational network malfunctions. In place of its usual appropriate activity, it becomes dominated by fixed, circulating masses of inappropriate activity, comparable to the hallucinations and delusions found in schizophrenia.

This faulty neuropil is thought to be produced partly by genetic or epigenetic factors, so that the synaptic construction and modulation system is not built properly in the first place, rendering it to malfunction at some point in its history, and partly by environmental insults such as viral infections *in utero*. One relevant factor is thought to be neurotoxic oxidative stress. Schizophrenic patients show many signs of synaptic pathology and oxidative stress.

These include the following pathologies:

- Abnormalities in postsynaptic proteins found in the postsynaptic densities of NMDA and AMPA receptors, such as PSD95 and SAP102.
- The expression of growth-associated protein-43 (GAP-43). This is a presynaptic lipoprotein involved in the control of neuronal development and synaptic plasticity.

- Reduced expression of synaptophysin and the vesicular glutamate transporter.
- A robust decrease in reelin expression. This is a key protein in neural development secreted by GABAergic neurons that modulates *inter alia* spine maturation.
- Oxidative damage to membrane lipids has been reported as well as several other abnormalities in indices of oxidative stress.
- Other preliminary findings suggest that there may also be abnormalities in the MAP-kinase cascade that modulates synaptic plasticity, in scaffolding proteins in the postsynaptic density, and in the function of brain-derived neurotrophic factors, oct-6 and S100R.
- LSD induces expression of many genes (such as c-fos, ARC, aria3, and others) related to proteins involved in modulating synaptic plasticity.

These studies show that the brain in schizophrenia presents evidence of various abnormalities in the complex neurochemical processes involved in synapse formation, augmentation, maintenance, and deletion.

See also: Altered and Exceptional States of Consciousness; General Anesthesia; Neurobiological Theories of Consciousness; Psychoactive Drugs and Alterations to Consciousness.

Suggested Readings

Karczmar AG (2007) Exploring the Vertebrate Central Cholinergic Nervous System. New York: Springer.
Monti JM, Pandi-Perumal SR, Jacobs BL, and Nutt DJ (eds.) (2008) Serotonin and Sleep. Basel: Birkhäuser.
Perry E, Ashton H, and Young AH (eds.) (2002) Neurochemistry of Consciousness: Neurotransmitters in the Mind. Amsterdam: John Benjamins.
Rowell PP, Yolk KA, Li J, and Bickford ME (2003) Investigations of the cholinergic modulation of GABA release in rat thalamus slices. Neuroscience 116: 447–453.
Smythies J (1996) The functional neuroanatomy of awareness. Consciousness & Cognition 6: 455–481.
Smythies J (1997) The biochemical basis of synaptic plasticity and neural computation: A new theory. Proceedings of the Royal Society of London. Series B 264: 575–579.
Smythies J (1999) The biochemical basis of coma. Psycholoquy, www.cogsci.soton.ac.uk/psyc.
Smythies J (2002) The Dynamic Neuron. Cambridge, MA: MIT Press.
Smythies J (ed.) (2004) Disorders of Synaptic Plasticity and Schizophrenia. Vol. 59: International Review of Neurobiology. San Diego: Elsevier.
Smythies J (2005) The Neuromodulators. Vol. 64: International Review of Neurobiology. San Diego: Elsevier.
Steriade M and McCarley RW (2006) Brain Control of Wakefulness. New York: Springer.
Yu AJ and Dayan P (2002) Acetylcholine in cortical inference. Neural Networks 15: 719–730.

Biographical Sketch

John Smythies MD is a graduate of Cambridge University and is currently a research scientist at UCSD and a senior research fellow at the Institute of Neurology, University College, London. Previously he has held the positions of C.B. Ireland professor of psychiatric research at the University of Alabama Medical Center, and reader in psychiatry at Edinburgh University. He has served as President of the International Society for Psychoneuroendocrinology (1970 1974), consultant to the World Health Organization (1963 1968), and editor of the *International Review of Neurobiology* (1956 1991). He is a fellow of the Royal College of Physicians (London) and was elected a member of the Athenaeum in 1968. He is the author of over 200 scientific papers and 16 books.

Neuroscience of Volition and Action

M Jeannerod, Institut des Sciences Cognitives, Lyon, France

Introduction

The main feature of voluntarily controlled actions is that they are generated from within, and that their generation requires the preexistence of an internal state (a representation) whereby they can be encoded, stored, and ultimately performed independently from external influences. Thus, they cannot be accounted for by automatic sensorimotor mechanisms, where a causal relationship between an immediate (sensory) cause and a visible behavior can be identified. Instead, voluntary actions are to be considered, not as mere responses to a stimulus, but as the self-generated expression of conscious cognitive states. A central question in this article will be to determine to what extent voluntary actions can be consciously controlled. This issue encompasses a wide range of technical as well as theoretical problems in neuroscience, psychology, and even philosophy.

Historical Landmarks

The existence of actions generated in the absence of sensory input was a central issue in nineteenth century neurology. Those who were influenced by thinkers like Charlton Bastian or William James defended the idea that peripheral influences, immediate or delayed, was a necessary condition for an action to appear. Sherrington, in showing the absence or poverty of spontaneous motor behavior in monkeys with deafferented limbs, provided experimental support to this theory. Alternatively, Hugo Liepmann, starting from the background of clinical neurology, abandoned the opposition between peripheral or central origin of the action. Instead, he proposed a completely new scheme in concentrating on how an action can be assembled from its elementary constituents. According to Liepmann, an action must proceed from an internal 'plan': ''The main representation of the goal can only be reached if a plan (*Entwurf*) is built internally, concerning the direction, contiguity, succession, and rhythm of the elementary acts.'' To account for the implementation of the plan, he proposed that the elementary bits of the action were assembled according to the main representation: The result of this process was what he called a 'movement formula' (*Bewegungsformel*), that is, an anticipatory hierarchical structure where all the aspects of an action were represented. Liepmann's legacy is still quite influential in neuropsychology and cognitive neuroscience of action. Later authors replaced the term of movement formula by those of 'engram,' 'schema,' or 'internal model,' but kept the notion of a representational level with a hierarchical organization.

Another, closely related, debate was about whether actions could be consciously monitored and controlled. This debate became known as the 'Two Williams debate.' On one side, William James, although he clearly considered will as an internal driving force, defended the opinion that the consciousness of our movements is based on *a posteriori* information from sensory organs. On the other side, Wilhelm Wundt held that conscious knowledge about our actions was based on *a priori* efferent information of a central origin. Wundt thought that we can perceive specific feelings, which he called the feelings of innervation (*Innservationsgefühl*), when we produce a movement. He was a proponent of the notion that the mind has a mental content that can be probed by appropriate methodology: To him, voluntary actions were an expression of this mental content.

What Do We Know about Our Actions?

Whoever of James or Wundt was right, the fact is that we remain unaware of many of our own actions. It is common experience that when leaving

home we ask ourselves questions like ''Did I lock the door?'' or ''Did I turn off the light?'' immediately after having done it. We may also drive the car back home and suddenly realize that we are at destination without having the least idea of how we did it. But, to the contrary, there are situations where we remain fully aware throughout the action. When I do something for the first time, for example, start a new laptop computer, I try to carefully follow the instruction manual and to consciously control each step of the action.

Here, we first examine the question of what consciousness of action is about. The first answer to this query that comes to mind is that consciousness of action refers to what the action is about. Actions have goals: To be aware of the goal one reaches for is one way of being conscious of the action undertaken to reach that goal. Being aware of the goal, however, does not imply to be aware of how it is being reached. Indeed, our above definition of voluntary control should imply that an agent can consciously access the internal representation of the action he is performing. To examine this point, we need experimental situations where the subjects have to ask themselves questions about what they really did to achieve a motor task. An example of such situations are those that produce a conflict between the different signals that are generated at the time of execution of an action (e.g., visual signals, proprioceptive signals, signals arising from central motor commands), and that are normally congruent with each other. Subjects' reports about their feelings in situations where these signals become mutually incongruent thus provide a direct insight into their ability to consciously monitor these signals. This type of experiments involving a sensorimotor conflict has a long tradition in psychology. They can typically be created by optical devices: Looking at one's moving hand through an inverting prism, for example, makes the movements to appear inverted, such that, in order to reach for a visual target seen on my right, I have to move my hand to the left. The main objective of these classical experiments was to study the process of adaptation, through which visuomotor coordination progressively rearranges for matching the hand movements to the apparent location of the visual targets. But only a few systematic attempts were made at studying the motor awareness of the subjects, that is, their insight about how they performed their own movements when unknowingly faced with this type of conflict.

The research on motor awareness was initiated by Torsten Nielsen more than 40 years ago. In a more recent version of Nielsen's experiment performed by P. Fourneret and M. Jeannerod, the subjective reports of subjects about what they had just done in trying to reach for a visual target were quantified. Participants were instructed to draw straight lines between a starting position and a target, using a stylus on a digital tablet. The output of the stylus and the target were displayed on a computer screen. The participants saw the computer screen in a mirror placed so as to hide their hand. On some trials, the line seen in the mirror was electronically made to deviate from the line actually drawn by the subject. Thus, to reach the target, the subject had to deviate his or her movement in the direction opposite to that of the line seen in the mirror. At the end of each trial, the subject was asked to indicate verbally in which direction he or she thought his or her hand had actually moved. The results were twofold: first, the subjects were consistently able to trace lines that reached the target, that is, they accurately corrected for the deviation; second, they gave verbal responses indicating that they thought their hand had moved in the direction of the target, hence ignoring the actual movements they had performed. Thus, although they correctly performed the task, they were unable to consciously monitor the discordance between the different signals generated by their own movements, and falsely attributed the direction of their hand to that of the line. In other words, they tended to adhere to the visible aspect of their performance, and to ignore the way it had been achieved. In fact, the access to motor awareness seems to be a matter of threshold. When, in the above experiment, the deviation of the line was progressively increased, the subjects' behavior suddenly changed. Their accuracy in reaching the target deteriorated and they became aware of the deviation: They could report that the movement of their hand erred in a direction different from that seen on the screen and that, in order to fulfill the instruction of reaching the target, they had to deliberately orient their hand

movement in a direction different from that of the target. The breakpoint at which the subjects shifted from an automatic compensation to a conscious adjustment to the deviation was around 14 on average.

Thus, the awareness of a discordance between an action and its sensory consequences emerges when the magnitude of the discordance exceeds a certain amount. In another experiment based on the same principle and made by G. Knoblich and T. Kircher, the participants had to draw circles on a writing pad. As in the above experiments, they did not see their hand, but they saw an image of their movement, represented by a moving dot on a computer screen. The velocity of the moving dot was either the same as that of the subject's movement, or it could be unexpectedly accelerated by a variable factor of up to 80%. To compensate for the change in velocity and to keep the dot move in a circle, as requested by the instruction, subjects had to decrease the velocity of their hand movement by a corresponding amount. They had to indicate any perceived change in velocity of the moving dot. Although they failed to detect the changes in velocity when they were of a small amplitude, their detection rate increased for faster velocity changes. Yet, subjects were found to be able to compensate for all changes in velocity, including for those that they did not consciously detect.

These experimental situations have a counterpart in everyday life, for example, when an action which was unfolding automatically and outside the subject's awareness fails because of some unexpected obstacle: The failure triggers a *prise de conscience* of the ongoing action. Such situations where an action is delayed, incompletely executed or blocked, typically create a mismatch between the desired motor output and its observed result. This mismatch would represent the 'stimulus' that triggers the conscious experience. Another set of experiments using neuroimaging techniques have directly examined the neural mechanism involved in this change in conscious awareness. They found that the neural activity of cortical areas located in the posterior and ventral parts of the right parietal lobe increases during a visuomotor conflict similar as those described in the above behavioral experiments. This increase appears to be a function of the degree of the conflict. The changes in activity of these areas thus reflect the incongruence between normally congruent signals from central or sensory origin generated during a self-generated movement: The central command signals generated for moving the hand in a certain direction are contradicted by the visual signals showing a movement in a different direction. The role of the right parietal lobe in monitoring such signals is supported by clinical observations. Patients with lesion in this region show striking neuropsychological symptoms that testify to an alteration of self-consciousness and consciousness of action: neglect of contralateral space, neglect of the corresponding half of the body, denial of ownership, or even denial of the frequently associated hemiplegia (anosognosia).

Is Consciousness of Action an Afferent or an Efferent Phenomenon?

Several sources of information are potentially available to make a conscious judgment about one's own motor performance. Among the sensory sources are the visual and the kinesthetic signals. Visual signals are directly derived from vision of the moving limb, or indirectly from the effects of the movement on external objects; kinesthetic signals are derived from movement-related mechanical deformations of the limb, through receptors located in the skin, joints, and muscles. Nonsensory sources are mainly represented by central signals originating from various levels of the action generation system. These different types of signals do not have the same status. Visual cues are of an uncertain origin: They cannot differentiate a self-generated from an externally generated visual change. By contrast, the central cues and, to some extent the kinesthetic cues, clearly relate to a self-generated movement: They are 'first-person' cues in the sense that they can only conceivably arise from the self.

Trying to examine the respective contributions of these different signals reminds the classical issue raised by the Two Williams debate alluded to in the historical section: Is consciousness of action an afferent or an efferent phenomenon? For a long time, experimenters have failed to resolve this

issue, mainly because of the methodological difficulty of isolating the two sources of information from one another. There are no reliable methods for suppressing kinesthetic information arising during the execution of a movement. Alternatively, it is possible to prevent muscular contractions in a subject who is attempting to move, for example, by using a curarizing agent (i.e., an agent that blocks neuromuscular transmission) for paralyzing one limb: If the subject reports sensations from his attempts to move his paralyzed limb, these sensations should result from outflow motor commands, not from proprioceptive inflow. The available evidence shows that no conscious perception of movement arises in this condition. However, experiments where an arm is only partially curarized (the arm is not paralyzed, but muscular force is weakened) suggest a more balanced conclusion: Subjects requested to estimate the heaviness of weights that they attempt to lift with their weakened arm report an increased perceived heaviness. This illusion was interpreted as reflecting the increase in motor outflow needed to lift the weights. This result provides indirect evidence as to the possibility for central signals to influence conscious experience.

A more direct solution to this problem is to examine patients with a pathological loss of haptic sensations (sensations from skin, joints, and muscles). One such patient, patient GL suffering a long-term sensory neuropathy, has been extensively studied by several experimenters, including Y. Lamarre and J. Paillard. Patient GL has no haptic information about the movements she performs. Thus, provided visual feedback from her movements is suppressed, the only information on which she can rely to form a phenomenal experience about her own action must be derived from central signals during the action generation processes. This point was examined by using the apparatus already described for the Fourneret and Jeannerod experiment. GL had to draw a line with her unseen hand while the line was made to deviate to the right by an angle increasing from 1 to 20 over successive trials. Like a normal subject, GL performed the task without difficulty: She was able to compensate for the deviation and to reach the target. When asked, at the end of each trial, to estimate verbally the angle by which she thought her hand had deviated to the left for bringing the line to the target, GL never explicitly reported a feeling of discordance between what she had seen and the movement that she thought she had made. Remember that, in this task, normal subjects become clearly aware of a displacement of their hand toward the left to compensate for the disturbance when the discordance exceeds a certain value. Instead, GL consistently gave responses indicating that she thought she had drawn the line in the direction of the target. In spite of expressing perplexity at the end of some trials, GL never became aware of the discordance and, consequently, of any strategy of correction she had to apply to correct for the deviation. When asked to describe her feelings, she only mentioned that she found the task 'difficult' and requiring an 'effort of concentration.'

Another experiment with the same patient also addressed the question of a possible role of the efferent processes in motor consciousness, by exploring the production and the perception of muscular force. When muscular force is applied isometrically (with no change in muscle length), kinesthetic input is limited, because there is no displacement of the limb: Thus, this condition should maximize the role of the central commands in the conscious appreciation of the exerted force. When instructed, first to isometrically apply a certain degree of force with one hand, and then to match this degree of force with the other hand, GL performed with a close to normal accuracy. This result indicates that she was able to produce accurate central commands. Yet, she was unable to report any conscious feeling from her effort, neither did she experience fatigue when a high degree of muscular contraction had to be maintained.

The central, nonsensory cues, which are still available in patient GL, appear to be of little use for consciously monitoring her own movements, except for the vague feelings of effort and difficulty that she reported in one of the tasks. However, a mere opposition between peripheral and central sources of information in providing cues for consciousness (which was the core of the Two Williams debate) may be misleading because it does not take into account the complete set of events arising during the voluntary execution of a movement. In the current model of self-generated actions, as designed by D. Wolpert and

his colleagues, the central signals are normally monitored and used as a reference for the desired action, and the reafferent sensory (e.g., kinesthetic) signals arising from the executed movement have to be matched with this reference. In patient GL, because no reafference resulted from her executed movement, this matching process could not take place. Here we propose that the conscious information about one's movements is normally derived, not directly from the reafferent signals themselves, but from the output of the matching process, for which the presence of both central and peripheral reafferent signals is required. In the case of GL, it was the lack of kinesthetic input from movement execution or isometric force generation, which severely impaired the possibility to consciously monitor the efferent signals.

The same explanation holds for the lack of conscious feelings during attempted movements in completely paralyzed subjects (e.g., produced by curarization). In spite of the generation of intense motor commands for fighting against the paralysis, the absence of corresponding kinesthetic reafferent signals prevents the matching process to take place. Conversely, when reafferent signals are present, but do not match the central signals, the matching process generates signals proportional to the degree of mismatch between the two. This would account for the conscious sensations of effort reported by subjects with an incomplete paralysis, where kinesthetic signals are still preserved and in normal subjects during lifting weights with incompletely paralyzed arms.

Can We Perceive Our Own Intentions?

The picture of consciousness that arises when it is studied in its relation to voluntary action is that of a *post hoc* phenomenon. Consciousness of action is bound to signals resulting *a posteriori* from the completion of the action itself, not to central signals that arise prior to the action. Accordingly, one should not expect that events preceding the execution of an action can come to consciousness. Is this also the case for intentions? Intentions are part of the representation of actions; they contribute to the stream of processing that flows from the early part of the representation down to action execution, when it occurs. In this section, we concentrate on experiments which have looked for the unfolding of intentions, their neural correlates, and the conditions for their conscious access.

Benjamin Libet was the first to tackle this problem. He instructed subjects to perform simple hand movements *ad libitum* and to report the instant (W) at which they became aware of wanting to move. In order to do so, subjects verbally reported the clock position of a spot revolving on a screen. An electromyogram (EMG) was recorded from arm muscles for measuring the precise onset of the movement: The 'W' judgment was found to precede EMG onset by 206 ms. In addition, electroencephalogram (EEG) potentials were recorded from the subjects' skull. The 'readiness potential,' a DC potential that appears during preparation to voluntary action, was found to anticipate 'W' by about 345 ms. This striking result shows that the intention (in the sense of 'wanting to move' or 'feeling the urge to move') can be perceived as distinct from execution itself; it also shows that the subject's declarative awareness of this phenomenon does not correspond to the actual onset of movement preparation, which starts much earlier.

The Libet's paradigm revealed the possibility (to be discussed further in the next section) of attending to one's own mental states preceding action. It also provided important information on the brain areas involved in this conscious process. A specific region of dorsal premotor cortex (the region corresponding to the supplementary motor area (SMA)) was consistently found to be activated in conjunction with the feeling of urging to move. First, the readiness potential, which in the experiments by Libet and his colleagues and by their followers, precedes the subjects' report of the feeling (the 'W' judgment), is thought to originate from SMA. Second, electrical stimulation of a small area immediately anterior to the SMA (the pre-SMA) induces a feeling of urging to move, as reported by patients during direct low intensity electrical stimulation applied to this area during neurosurgical operations. The pre-SMA and the SMA are relatively closely connected to motor cortex: Indeed, increasing the intensity of the stimulus produces muscular contractions in the body part where the feeling was resented. Finally, a

functional magnetic resonance imaging (fMRI) study of normal subjects attending to their intention showed that the 'W' judgment made by these subjects was associated with activation of the pre-SMA and the dorsolateral prefrontal cortex.

The dorsolateral cortex itself is known to be involved in willed actions. Its role in carrying intentions and motor plans, conscious or not, and in inhibiting concurrent activity has been demonstrated by neuropsychological experiments in normal subjects as well as by clinical observations (see below). Christopher Frith and his colleagues went further in monitoring brain activity in normal subjects during tasks involving a voluntary choice between two different actions (what they consider as the hallmark of free will). In one condition, the subjects were instructed to move, at their own choice, either their right index finger or their right thumb. In the other condition, they had to utter a word, for example, beginning with the letter s. Brain activity measured with positron-emission tomography (PET) was found to be specifically increased in an area corresponding to the lateral prefrontal cortex corresponding to Brodmann area 46: in the condition of finger movement, the activation was bilateral; in the condition of word choice, it was limited to the left side. The role of prefrontal cortex and other cortical regions in voluntary action is also supported by experiments measuring the time course of brain activity during conscious decisions. Recent fMRI experiments by Haynes and his colleagues in Berlin demonstrated that predictive information about the decision to move could be detected in prefrontal cortex (in the frontopolar area 10) and in a posterior parietal region up to 7–10 s prior to subject's awareness of his or her decision. This suggests that, by the time the subject's decision reaches awareness, it has been influenced by unconscious brain activity lasting several seconds.

How Do We Get the Sense of Being the Agent of Our Actions?

Actions like aimlessly moving one's finger or uttering a phoneme, such as those described in the previous section, have no real goal and little impact on the external world. Still, they represent a minimal case for voluntary actions: They result from a deliberate decision or choice from the agent. In John Searle's classification, they correspond to what he calls prior intentions, that is, intentions which are accompanied by the feeling of intending, wanting, wishing, choosing, or, in other words, of the feeling of conscious will. In classical philosophy, the concept of will is associated with the subjective impression of causing an action to appear. Thomas Reid, for example, in his *Essays on the Active Power of the Human Mind* (1788), wrote that "every man is conscious of a power to determine, in things which he conceives to depend upon his determination. To this power, we give the name of will."

For the psychologist, however, the problem is that of the relationship of this conscious sense of willing to act to the forthcoming action. As we saw earlier, execution of actions is automatic, and it is not preceded by a clear insight about what is going to happen. The brief flash of consciousness found by Libet in the 200–300 ms preceding muscular contraction is not likely to represent a significant cause for the action to appear. Indeed, Libet considered that his own results "lead to the conclusion that cerebral initiation... of a spontaneous voluntary act... can and usually does begin *unconsciously*". "The brain," he said, "'decides' to initiate or, at least, to prepare to initiate the act before there is any reportable subjective awareness that such a decision has taken place." The fact that consciousness follows, and does not precede, the process of action generation is hardly compatible with the notion of a causal role of the conscious sense of will. We are therefore faced with a paradox: whereas we feel and we strongly believe that our thoughts determine our behavior, we realize that conscious free choice and conscious will are consequences, not causes, of the brain activity that itself causes the actions to appear. The fact that brain activity is ahead of our mental states, and not the reverse, is not a surprise. The surprising thing, however, is the discordance between this very fact and our subjective experience of being the agent of the action (the sense of agency).

Before discussing the issue of the causal will, we will examine experimentally what are the cues which determine the sense of agency. One way to answer this question is to examine the responses of

subjects placed in experimental situations where an uncertainty is artificially created about the origin of an action, and where they are explicitly requested to make conscious agency judgments. In these experiments, normal subjects are instructed to execute simple finger movements without direct visual control of their hand. Instead of seeing their hand during their movements, subjects see on a screen the image of two hands: Their own hand or an alien hand (the hand of an experimenter) executing the same or different movements. Successive trials involve different degrees of conflict between the seen and the felt position of the subject's hand, and between the seen and the actually executed finger movements: The two hands can either make the same movements, different movements, or no movement at all. At the end of each trial, after the two hands have disappeared, a pointer is placed at the location of one of the two hands and the subject is asked to judge whether that hand was his own or that of the experimenter. In this condition, the subject's errors can be in two different directions: in the other-to-self direction when the subject attributes to himself the hand of the experimenter (over-attribution errors) and in the self-to-other direction when he attributes his own hand to the experimenter (under-attribution errors). According to E. van den Bos and M. Jeannerod who made this experiment, the overall pattern of results was that the subjects tended to make more over-attribution errors. This pattern was particularly striking in the condition where all the cues for discriminating between the two hands had been suppressed: In spite of responding at chance on average, the subjects still made significantly more errors by over-attribution than by under-attribution.

The conclusion that can be drawn from the above results is twofold. First, the sense of agency can be subject to misinterpretation and to illusions (and possibly even to delusions, in psychotic states such as schizophrenia). Second, attribution judgments are biased: A subject placed in ambiguous situations like those created by the substitution experiments experiences alien actions as his own. This is indeed an example of an illusion of agency: Faced with an action of an uncertain origin, this subject will tend to believe that he is the agent who caused that action to appear.

The dissociation in normal subjects, for which we already saw several illustrations, between automatic processing of an action and conscious responses about that action raises the question of the relationship between their respective underlying mechanisms. Would the sense of agency be affected by perturbation of the automatic process like those we reported in the previous sections? Indications in the literature are very scarce on this point. Nielsen had reported that his subjects, when misattributing the alien hand to themselves, felt as if they had lost control of their movements, 'as if driving on ice.' Other authors reported that subjects observing a delayed presentation of their motor performance experienced the bizarre sensation of having an 'anarchic' hand. In both examples, subjects experienced the illusion of a disturbed sense of agency when the feedback from their actions did not correspond to the expected effect.

Disorders of Volition

Pathological disorders affecting volition may shed light on the debate about the role of conscious will in behavior. The expression 'disorders of volition' refers to those pathological conditions where the ability to make choices, to express preferences, or possibly to experience pleasure and freedom in making these choices or expressing these preferences, are affected. Théodule Ribot, in his book on the diseases of the will, considered two opposite sides, the negative and the positive, of the alterations of the will.

Several terms have been used to describe such pathological conditions. On the negative side, aboulia, a pure disease of the will according to Ribot, has been defined as an impairment to execute what is in mind. In this condition, there is no paralysis, no disorder of the muscular system, there is no lack of desire, but the transition from desire to execution becomes abnormally difficult. An extreme example of aboulia has been described clinically under the term of athymhormia: Such patients show loss of drive and of search for satisfaction, lack of curiosity, lack of taste and preferences, flattened affect. This condition, clearly distinct from depression, can be observed in schizophrenia and may also be caused by lesions in the

basal ganglia. Another example is the 'chronic fatigue syndrome' (CFS), defined by persistent or relapsing unexplained fatigue. CFS patients were tested in a task where they had to imagine themselves performing an action (a motor imagery task), while their brain activity was monitored with fMRI. They were found behaviorally to be slower than controls. In addition, an area in their anterior cingulate cortex, known to be involved in error monitoring, remained inactive when they made errors, which supports the notion that CFS may be associated with an inability to build action plans and to monitor one's own performance according to these plans.

On the positive side are found conditions where the deficient will cannot block impulses to act: The power to control and to inhibit, which Ribot considered as the highest level of the will, is impaired. A well-known example of this pathological condition is that of patients with frontal lobe lesions, who are unable to consciously monitor their performance and to override their automatic responses. These patients tend to compulsively imitate gestures or even complex actions performed in front of them by another agent; when presented with common graspable objects, they cannot refrain from using them. This striking behavior, termed 'imitation' or 'utilization' behavior by F. Lhermitte, may be explained by an impairment of the normal suppression, by the prefrontal areas, of inappropriate actions triggered by external stimuli. As a consequence of this impairment, the behavior of the patient is under the dependence of external events. One such patient exemplified this typical behavior by putting pairs of goggles on his nose each time he was presented a new pair, ending with several pairs superimposed upon each other!

This type of behavior observed following frontal lesions suggests the existence of an imitative tendency, which would normally be inhibited in everyday life situations. Compulsive imitation in frontal patients would result from an impairment of this inhibitory mechanism. Following this idea, Marcel Brass and his colleagues have studied brain activity in normal subjects who were instructed to execute a simple predefined finger movement (e.g., lift one finger) in response to the onset of an observed movement executed by an experimenter. When the observed movement was congruent with the movement the subject was instructed to perform, the response was given within a short latency; in contrast, when it was incongruent (e.g., the subject had to lift his finger in response to an observed tapping movement), the latency was increased. In addition, in incongruent trials, the lateral prefrontal cortex in the middle prefrontal gyri was strongly activated. It is therefore conceivable that a prefrontal lesion involving the lateral prefrontal cortex impairs the possibility to refrain from imitating other people. Lhermitte had used the term 'environmental dependency syndrome' to designate the tendency of these patients to stick to external events, at the detriment of their own free will.

Conclusion

At this point, the problem of the consciousness of our own actions and intentions clearly merges with that of self-consciousness. The ability to identify oneself as the agent of a behavior or a thought – the sense of agency – is the way by which the self builds up as an entity independent from the external world and from other agents. The conscious sense of will arises from the belief that our thoughts have a causal influence on our behavior. While we tend to perceive ourselves as causal, we actually ignore the cause from which our actions originate. Because the conscious thought and the observed action are consistently associated, even though they may not be causally related, the narrative self tends to build a cause-and-effect story. Therefore, conscious free choice, like conscious will, may not be a direct perception of a causal relation between a thought and an action, but rather a feeling based on the causal inference one makes about the data that do become available to consciousness. In conclusion, according to Wolfgang Prinz, "There appears to be no support for the folk psychology notion that the act follows the will, in the sense that physical action is caused by mental events that precede them and to which we have privileged access." "Experimental evidence suggests that two different pathways [. . .] may be involved in action – one for the generation of physical action and one for the mental awareness of its causal antecedents. If anything, the second follows the first, and not vice-versa [. . .]."

The role of consciousness should rather be to ensure the continuity of subjective experience across actions which are – by necessity – executed automatically. Because it reads behavior rather than starting it, consciousness represents a background mechanism for the cognitive rearrangement after the action is completed, for justifying its results, or modifying the factors that have been at the origin of an action which turned out to be unsuccessful, as in learning a new skill, for example. This mechanism could have the role of establishing a declarative cognitive statement about one's own preferences, beliefs, or desires. But the fact is that what we know about our mental content does not match its actual functioning, that which determines the causal relations of our behavior. When asked questions about these causal relations, we tend 'to tell more than we can know' and to build explanations based on incomplete information. This tendency unavoidably creates a dissociation between what we believe we do and what we actually do.

See also: Brain Basis of Voluntary Control; Free Will; Habit, Action, and Consciousness; Perception, Action, and Consciousness.

Suggested Readings

Berrios GE and Gili M (1995) Will and its disorders: A conceptual history. History of Psychiatry 6: 87–104.

Brass M, Zysset S, and Von Cramon Y (2001) The inhibition of imitative response tendencies. NeuroImage 14: 1416–1423.

Fourneret P and Jeannerod M (1998) Limited conscious monitoring of motor performance in normal subjects. Neuropsychologia 36: 1133–1140.

Gandevia SG and McCloskey DI (1977) Changes in motor commands, as shown by changes in perceived heaviness, during partial curarization and peripheral anaesthesia in man. Journal of Physiology 272: 673–689.

Haynes J-D, Sakai K, Rees G, Gilbert S, Frith C, and Passingham RE (2007) Reading hidden intentions in the human brain. Current Biology 17: 323–328.

Jeannerod M (2006) Motor Cognition: What Actions Tell the Self. Oxford: Oxford University Press.

Knoblich G and Kircher TTJ (2004) Deceiving oneself about being in control: Conscious detection of changes in visuomotor coupling. Journal of Experimental Psychology. Human Perception and Performance 30: 657–666.

Lhermitte F (1983) Utilisation behaviour and its relation to lesions of the frontal lobes. Brain 106: 237–255.

Libet B, Gleason CA, Wright EW, and Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain 106: 623–642.

Liepmann H (1905) Ueber Störungen des Handelns bei Gehirnkranken. Berlin: S. Kargen.

Nielsen TI (1963) Volition: A new experimental approach. Scandinavian Journal of Psychology 4: 225–230.

Prinz W (2003) How do we know about our own action? In: Maasen S, Prinz W, and Roth G (eds.) Voluntary Action. Brains, Minds and Sociality, pp. 21–33. New York: Oxford University Press.

Ribot T (1894) Les maladies de la volonté. Paris: Alcan.

Wegner D (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.

Wolpert DM, Ghahramani Z, and Jordan MI (1995) An internal model for sensorimotor integration. Science 269: 1880–1882.

Biographical Sketch

Marc Jeannerod was born in Lyon, France. He completed his degree in medicine at the University of Lyon, with a specialization in neurology and clinical neurophysiology. After a post doctoral stay at the University of California at Los Angeles, he became professor in physiology at the Claude Bernard University in Lyon, head of the INSERM Laboratory of Experimental Neuropsychology (1978–1996), and director of the Institute for Cognitive Science (1997–2004). Marc Jeannerod has published several books devoted to the control of action. His last book is Motor Cognition (OUP, 2006).

Perception, Action, and Consciousness

M A Goodale, University of Western Ontario, London, ON, Canada
A D Milner, Durham University, Durham, UK

Glossary

Basal ganglia – Several large clusters of nerve cells (including the caudate nucleus, putamen, and the globus pallidus) located deep in the brain below the cerebral hemispheres. The basal ganglia play a role in a number of functions including motor control, cognition, and emotion.

Blindsight – Residual visual abilities in individuals with large lesions of primary visual cortex who claim that they are blind in that part of the visual field in which they show evidence of visual sensitivity. For example, patients with unilateral lesions of primary visual cortex can sometimes point accurately toward stimuli in visual field contralateral to their lesion while at the same time denying any conscious experience of those stimuli.

Brainstem – The major route by which the forebrain receives information from, and sends information to, the spinal cord and peripheral nerves. The chief divisions of the brainstem are the midbrain, pons, and medulla.

Cortical blindness – Blindness that results from damage to the primary visual areas in the cerebral hemispheres. Individuals who have cortical blindness may nevertheless show blindsight.

Dorsal stream – Visual pathway arising in early visual areas in the cerebral cortex and projecting to the posterior parietal cortex. This pathway is thought to subserve the visual control of actions, such as reaching and grasping.

Extinction – In neuropsychology and neurology, extinction (or bilateral simultaneous extinction) refers to a phenomenon in which a patient with a unilateral lesion (typically in the inferior regions of the posterior parietal cortex) fails to detect a visual (or tactile) stimulus presented contralateral to the lesion when another (similar) stimulus is presented simultaneously on the opposite side. When the contralateral stimulus is presented on its own, the patient is able to detect it.

Hypoxia – A deficiency of oxygen reaching the tissues of the body.

Inferotemporal cortex – The most ventral part of the temporal lobe. The inferotemporal cortex contains higher-order areas of the ventral stream of visual processing.

Metacognition – Awareness of one's own thinking and decision making; sometimes termed 'knowing about knowing.'

Optic ataxia – An inability to guide the hand toward an object using vision. Optic ataxia is caused by damage to the posterior parietal cortex.

Phototaxis – Movement of an organism toward (or away from) light.

Posterior parietal cortex – Cortex in the parietal lobe behind the postcentral sulcus. A prominent sulcus in the posterior parietal cortex is the intraparietal sulcus (IPS). This region of the posterior parietal cortex contains visuomotor areas that make up the dorsal stream of visual processing.

Premotor cortex – A complex mosaic of interconnected areas in the frontal lobe immediately anterior to primary motor cortex. Although there is not complete agreement about the function of different parts of premotor cortex, it has been suggested that areas in this region of cortex participate in motor planning and movement selection. Mirror cells are found in this region.

Ventral stream – Visual pathway arising in early visual areas in the cerebral cortex of the

primate brain and projecting to the inferotemporal cortex. This pathway mediates visual perception, allowing the visual recognition of objects and events. Processing in this pathway is necessary, but not sufficient, for visual awareness.
Visual agnosia – An inability to recognize visual stimuli despite spared low-level visual processing. There are several varieties of visual agnosia, all of which involve damage to some part of the ventral stream. In the case of associative agnosia, the patient is unable to recognize an object despite being able to draw a reasonably faithful representation of what he or she sees. In the case of apperceptive agnosia (or visual form agnosia), the deficit is more fundamental and the patient cannot recognize even simple shapes or discriminate between them – and is unable to copy line drawings.

Introduction

Vision is the primary route to our conscious experience of the world beyond our bodies. Although we certainly hear and feel things in our immediate environment, neither of these sensory experiences is any match for the rich and detailed representation of the world provided by our sense of sight. The majesty of a distant mountain or the angry face of an approaching enemy can be appreciated only as visual experiences. But vision does more than provide us with our perception of the world. It also allows us to move around that world and to guide our goal-directed actions. Although it is tempting to think that these different functions of vision are mediated by one and the same visual representation in our brain, it has become increasingly clear over the last two decades that the visual pathways that underlie our perception of the world are quite distinct from those that underlie the control of our actions. Indeed, the distinction between 'vision-for-perception' and 'vision-for-action' has emerged as one of the major organizing principles of the visual brain, particularly with respect to the visual pathways in the cerebral cortex. But before elaborating this distinction it is important to understand how vision began.

The Evolutionary Origins of Vision

Visual systems first evolved not to enable animals to perceive the world, but to provide distal sensory control of their movements. One clear example of this is phototaxis, a behavior exhibited by many simple organisms whereby they move toward or away from light. Some bacteria, for example, use orange light as a source of energy for metabolic activity, but must avoid ultraviolet light, which can damage their DNA. As a consequence, these bacteria have developed a differential phototactic response, whereby the system measures light intensity at different wavelengths so that they end up moving toward orange light and away from UV light. To explain the bacteria's light-sensitive behavior, it is not necessary to argue that these single-celled organisms 'perceive' the light or even that they have some sort of internal model of the outside world coded in their one or more of their organelles. One simply has to posit the existence of some sort of input–output device within the bacteria that links the intensity of ambient orange and UV light to the pattern of locomotion. As it turns out, exactly the same argument can be made about the visually guided behavior of more complex organisms, such as vertebrates. Indeed, as we shall see later, a broad range of human behavior can also be explained without reference to experiential perception or any 'general-purpose' representation of the outside world.

In vertebrates, different classes of visually guided behavior have evolved as relatively independent neural systems. For example, in present-day amphibians, such as the frog, visually guided prey-catching and visually guided obstacle avoidance are mediated by separate neural pathways arising in the retina and projecting to distinct motor networks in the brain that produce the constituent movements of these two classes of behavior. In fact, evidence from several decades of work in both frog and toad suggests that there are at least five separate visuomotor modules, each responsible for a different kind of visually guided behavior and each having neural pathways from

input to output. The outputs of these different modules certainly have to be coordinated, but in no sense are they guided by a single general-purpose visual representation in the frog's brain. There is evidence as well that the same kind of visuomotor modularity found in amphibians also exists in the mammalian and avian brain.

Nevertheless, even though there is considerable evidence for visuomotor modularity in all classes of vertebrates, the very complexity of the day-to-day living in many mammals, particularly in higher primates, demands much more flexible organization of the circuitry. In monkeys (and thus presumably in humans as well), many of the visuomotor circuits that are shared with simpler vertebrates appear to be modulated by more recently evolved control systems in the cerebral cortex. Having this layer of cortical control over the more ancient subcortical networks makes it possible for primates to have much more flexible visually guided behavior. But even so, the behavior of primates, particularly with their conspecifics, is so complicated and subtle, that direct sensory control of action is often not enough. To handle these complexities, representational systems have emerged in the primate brain (and presumably in other mammals as well), from which internal models of the external world can be constructed. These representational systems allow primates such as ourselves to perceive a world beyond our bodies, to share that experience with other members of our species, and to plan a vast range of different actions with respect to objects and events that we have identified. This constellation of abilities is often identified with consciousness, particularly those aspects of consciousness that have to do with decision making and metacognition. It is important to emphasize that the perceptual machinery that has evolved to do this is not linked directly to specific motor outputs, but instead accesses these outputs via cognitive systems that rely on memory representations, semantics, spatial reasoning, planning, and communication. In other words, there are a series of cognitive 'buffers' between perceiving the world and acting on it, and the relationship between what is on the retina and the behavior of the organism cannot be understood without reference to other mental states, including those typically described as conscious. But once a particular course of action has been chosen, the actual execution of the constituent movements of that action are typically carried out by dedicated visuomotor modules not dissimilar in principle from those found in frogs and toads.

To summarize: vision in humans and other primates (and perhaps other animals as well) has two distinct but interacting functions: (1) the perception of objects and their relations, which provides a foundation for the organism's cognitive life and its conscious experience of the world, and (2) the control of actions directed at (or with respect to) those objects, in which separate motor outputs are programmed and controlled online. These different demands on vision have shaped the organization of the visual pathways in the primate brain.

Two Visual Systems in the Primate Cerebral Cortex

In the primate brain, two 'streams of visual processing' arise from early visual areas in the cerebral cortex and project to higher-order visual areas (Figure 1). One of these projection systems, the

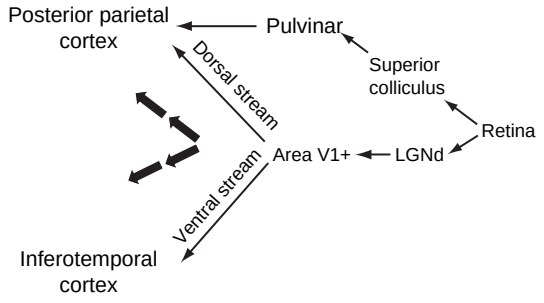


Figure 1 Schematic representation of the two streams of visual processing in human cerebral cortex. The retina sends projections to the dorsal part of the lateral geniculate nucleus in the thalamus (LGNd), which projects in turn to primary visual cortex (V1). Within the cerebral cortex, the ventral stream arises from early visual areas (V1+) and projects to regions in the occipito-temporal cortex. The dorsal stream also arises from early visual areas but projects instead to the posterior parietal cortex. The posterior parietal cortex also receives visual input from the superior colliculus via the pulvinar. On the left, the approximate locations of the pathways are shown on an image of the brain. The routes indicated by the arrows involve a series of complex interconnections.

dorsal stream, projects from early visual areas to the posterior parietal cortex, a region of the brain that is reciprocally connected to motor areas in the frontal cortex and sends projections to the basal ganglia and other (older) motor nuclei in the brainstem. The dorsal stream also receives visual input from the midbrain (via the thalamus). The other projection system, the ventral stream, projects from early visual areas to inferotemporal cortex. The ventral stream has strong connections with medial temporal areas, including the amygdala and hippocampus, as well as the prefrontal cortex. As it turns out, the functions of the two streams map quite well onto the distinction between vision-for-action and vision-for-perception discussed above. Not surprisingly, given its interconnections with motor structures, it is the dorsal stream that plays the critical role in the programming and control of actions, transforming real-time information about the location and disposition of objects into the coordinate frames of the relevant motor systems. In contrast, it is the ventral stream (together with associated cognitive networks) that mediates the construction of the rich and detailed representations of the world that allow us to identify objects, events, and actions in others, attach meaning and significance to them, and infer their causal relations. In summary, processing in the ventral stream provides the conscious visual percepts that are essential for accumulating a knowledge base about the world, knowledge that we can access for cognitive operations, such as planning and decision making. Processing in the dorsal stream does not generate visual percepts; it generates skilled actions (as part of a network of structures involved in sensorimotor control). Of course, the two streams are not hermetically sealed from one another. Indeed, they work together in controlling our behavior as we live our complex lives – but they play separate and complementary roles in the production of adaptive behavior.

Much of the evidence for this idea first came from work with neurological patients. The best known of these cases is the patient D.F., who developed a profound 'visual form agnosia' as a consequence of a hypoxic episode in which her brain was starved of oxygen. The nature of D.F.'s deficit in form vision can be understood to some extent by examining the drawings illustrated in Figure 2. Not only was D.F. unable to identify the simple line drawings illustrated in left-hand column of this figure, but she was also unable to copy them, at least in a recognizable way. A preserved ability to see fine detail allowed her to depict some aspects of the drawings, such as the dots indicating the print in the line drawing of the open book. Nevertheless, she was unable to duplicate the overall shape or arrangement of the elements of the line drawings in her copies. D.F.'s inability to copy the drawings is not due to a problem in controlling the movements of the pen or pencil; when she was asked on a separate occasion to draw an object from memory, she was able to do so reasonably well, as the drawings on the right-hand side of Figure 2 illustrate. Needless to say, when D.F. was shown any of the drawings she had done herself, whether the ones retrieved from memory or those copied from another drawing, she had no idea what they were and commented that they all looked like 'squiggles.' It is important to emphasize that D.F. retains the ability to perceive and describe the colors and other surface properties of objects, such as their visual texture. The missing elements in her perceptual experience are restricted to shape and form, and

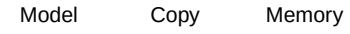


Figure 2 The patient D.F.'s attempts to draw from models and from memory. D.F. was unable to identify the line-drawings of the apple, open book, or the boat shown on the left. In addition, her copies were very poor. Note, however, that she did incorporate some elements of the line-drawing (e.g., the dots indicating the text in the book) into her copy. When she was asked on another occasion to draw these same items from memory, she produced a respectable representation of all three (right-hand column). When she was later shown her own drawings, she had no idea what they were.

thus her problems cannot be dismissed as a generalized inability to make perceptual reports.

Remarkably, even though D.F. shows no perceptual awareness of the form or dimensions of objects, she automatically adjusts her hand to the size, shape, and orientation of an object as she reaches out to pick it up. For example, even though D.F. is unable to distinguish between rectangular blocks of different dimensions, when she reaches out to pick up one of the blocks, the aperture between her fingers and thumb is scaled in-flight to the width of the object, just as it is in people with normal vision (see Figure 3a). In other words, D.F. can scale her grip to the dimensions of an object in anticipation of picking it up, even though she is unable to 'perceive' the dimensions of that object. Similarly, she will rotate her hand correctly for objects placed in different orientations, and will direct her fingers to stable grasp points on the surface of the objects, even though in other tests she fails to indicate either verbally (or in a manual matching task) the orientation or shape of those same objects. She is also able to avoid other objects in the workspace as she reaches out toward a goal, even though she cannot judge their relative locations correctly in a more perceptual task. D.F. exhibits normal visuomotor control in other tasks as well, including stepping over obstacles during locomotion, despite the fact that her perceptual judgments about the height of these obstacles are far from normal. To summarize: despite the fact that D.F. has lost all conscious perception of the form of objects, including their size, shape, and orientation, her visuomotor systems are able to make use of these same object properties to control skilled object-directed actions.

So where is the damage in D.F.'s brain? Although D.F. shows some diffuse loss of tissue throughout her cerebral cortex (consistent with hypoxia), she has prominent focal lesions bilaterally in a region of the human ventral stream that has been shown to be involved in the visual recognition of objects. It is presumably the damage to these object-recognition areas that has disrupted her ability to perceive the form of objects. But clearly these ventral-stream lesions have not interfered with her ability to use visual information about form to shape her hand when she reaches out and grasp objects. The preservation of normal visually guided grasping in the face of ventral-stream damage suggests this ability is dependent on another visual pathway, the most likely candidate being the dorsal stream. This conclusion has been supported by neuroimaging evidence showing that when D.F. grasps

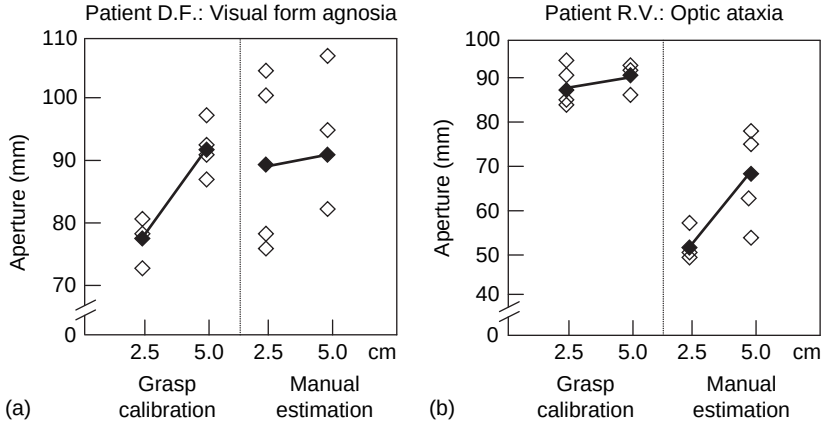


Figure 3 Graphs showing the size of the aperture between the index finger and thumb during object-directed grasping and manual estimates of object width for D.F., a patient with visual form agnosia and R.V., a patient with optic ataxia. D.F. (a) showed excellent grip scaling, opening her hand wider for the 50 mm-wide object than for the 25 mm wide object. D.F.'s manual estimates of the width of the two objects, however, were grossly inaccurate and showed enormous variability from trial to trial. In contrast, R.V. (b) was able to indicate the size of the objects reasonably well (individual trials marked as open diamonds), but her maximum grip aperture in flight was not well-tuned. She simply opened her hand as wide as possible on every trial. Reproduced from Goodale MA, Milner AD, Jakobson LS, and Carey DP (1991) A neurological dissociation between perceiving objects and grasping them. Nature 349: 154–156, with permission from Nature Publishing Group.

objects that vary in size and orientation, she shows relatively normal activity in a small region of the dorsal stream that has been implicated in the visual control of grasping in healthy individuals. It should be emphasized that although patients like D.F. are rare, there are a number of other cases in the literature in which the same striking dissociation between visual perception and visuomotor control has been documented.

But what about patients with damage to the dorsal stream? As it turns out, these patients exhibit a pattern of deficits and spared abilities that are complementary to that seen in patients with ventral-stream lesions. Thus, patients with dorsal-stream lesions typically have problems reaching toward targets placed in different positions in the visual field, particularly the periphery. This deficit is referred to clinically as 'optic ataxia.' But the failure to locate an object with the hand cannot be construed as a problem in spatial representation: many optic ataxia patients, for example, can describe the relative position of the object in space quite accurately, even though they cannot direct their hand toward it. Also, the deficit is not purely motor: these patients usually have no difficulty using input from other sensory systems, such as proprioception or audition, to guide their movements. In addition to their deficits in reaching, many patients with damage in the dorsal stream are unable to use visual information to rotate their hand, scale their grip, or configure their fingers properly when reaching out to pick up objects, even though they are able to correctly report the orientation, size, and shape of those objects (see Figure 3b). In addition, they do not take into account the positions of potential obstacles when they are attempting to reach out toward goal objects even though they can indicate the relative location of the obstacles in other ways. In summary, patients with optic ataxia exhibit neither a purely visual nor a purely motor deficit, but instead a specific deficit in visuomotor control, confirming the critical role that the dorsal stream plays in the control of skilled actions.

In addition to such work with neurological patients, there is a wealth of evidence from monkey neurophysiology and human neuroimaging supporting the idea of a ventral 'perception' stream and a dorsal 'action' stream.

Different Metrics and Frames of Reference for Perception and Action

But why did two separate streams of visual processing evolve in the primate cerebral cortex in the first place? Or, to put it another way, why couldn't one 'general purpose' visual system handle both vision-for-perception and vision-for-action? The answer to this question lies in the computational requirements of the two kinds of vision. To be able to grasp an object successfully, for example, the brain must compute the actual size of the object, and its orientation and position with respect to the grasping hand of the observer (i.e., in egocentric coordinates). The time at which these computations are performed is equally critical. Observers and goal objects rarely stay in a static relationship with one another and, as a consequence, the egocentric coordinates of a target object can often change radically from moment to moment. For these reasons, it is essential that the required coordinates for action be computed in an egocentric framework at the very moment the movements are to be performed.

The computations underlying perception are quite different. Vision-for-perception does not deliver the absolute size of objects or their egocentric locations. In fact, such computations would be counterproductive for a recognition system precisely because we almost never stay fixed in one place in the world. The problem can be easily solved by the alternative strategy of encoding the size, orientation, and location of objects relative to each other. Such a scene-based frame of reference permits a perceptual representation of objects that transcends particular viewpoints, while preserving information about spatial relationships (as well as relative size and orientation) as the observer moves around. Indeed, it has been suggested that if the perceptual machinery had to deliver the real size and distance of all the objects in the visual array, the computational load would be prohibitive. The products of perception also need to be available over a much longer timescale than the visual information used in the control of action. We may need to recognize objects we have seen minutes, hours, days – or even years before. In short, the perceptual information is lodged in memory. To achieve this, the coding of the visual information has to be

somewhat abstract – transcending particular viewpoint and viewing conditions. By working with perceptual representations that are object- or scene-based, we are able to maintain the constancies of size, shape, color, lightness, and relative location, over time and across different viewing conditions. Although there is much debate about how this is achieved, it is clear that it is the identity of the object and its location within the scene, not its disposition with respect to the observer that is of primary concern to the perceptual system. Object recognition occurs when current perception concurs with stored information about previously encountered objects. Thus, the ventral stream provides the perceptual foundation for the off-line control of action, projecting action into the future, and incorporating stored information from the past into the control of current actions.

Perception, Action, and Consciousness

As outlined in the section titled 'Different metrics and frames of reference for perception and action,' the ventral and dorsal streams play different but complementary roles in the service of behavior. The ventral stream (together with associated cognitive machinery) permits the brain to identify goals and plan appropriate actions; the dorsal stream (in conjunction with related circuits in premotor cortex, basal ganglia, and brainstem) programs and controls those actions. Ultimately then, both systems transform visual information into motor output. In the dorsal stream, the transformation is quite direct: visual input and motor output are essentially 'isomorphic' with one another. In the ventral stream, however, the transformation is indirect: the construction of a perceptual representation of the visual world enables a more 'propositional' relationship between input and output, taking into account previous knowledge and experience.

The neuropsychological evidence from patients such as D.F. suggests that there is a close relationship between ventral-stream processing and consciousness. It is not that D.F. is simply unable to describe the form of objects; she seems to have absolutely no conscious appreciation of their dimensions, shape, or orientation. For example, even though D.F.'s hand automatically conforms to the dimensions of an object when she reaches out to grab it, she cannot indicate the width of the object by using an explicit manual 'matching' response, using her index finger and thumb to show how wide she thinks it is. The loss of form perception is only one example of how ventral-stream damage can affect visual consciousness. For example, patients with damage to other ventral-stream areas can lose all conscious experience of color or visual texture. Such patients may continue to see the boundaries or edges between adjoining patches of color, even though they have no appreciation whatsoever of the colors determining those boundaries. Patients with damage to primary visual cortex, which is the major input to the ventral stream, typically report seeing nothing in the visual field contralateral to their lesion. They are cortically blind. But some of these patients show evidence of spared visuomotor control in their blind field, a phenomenon sometimes called blindsight. In other words, they can make eye movements or point to targets that they cannot consciously see, and may even show evidence of anticipatory shaping of the hand when they reach out to grasp objects placed in their blind field. It is thought that blindsight of this kind may rely on projections to phylogenetically ancient visuomotor structures in the midbrain and brainstem, which are in turn connected with the visuomotor systems in the dorsal stream. All of this evidence, as well as a wealth of other data, suggests that the ventral stream is necessary (though not sufficient) for visual consciousness.

From an evolutionary perspective, the visual phenomenology that arises from ventral-stream processing must confer some kind of advantage to those that possess it. In other words, the capacity for conscious perception must give an organism an edge in natural selection. One can only speculate as to what that advantage might be. It is possible that conscious representations of the world are the only representations that can (eventually) enter our long-term visual knowledge. As already discussed, by retrieving information from long-term memory, we can manipulate this information in working memory (together with information from current percepts) for flexible 'off-line' control of

behavior. Moreover, when an animal is aware of what it sees, it can use this information to decide between different actions (including making no action), to plan future actions, and to communicate what it sees (or has seen) to its conspecifics.

Being conscious of visual processing in the dorsal stream would confer no such advantages on the animal. Because the transformations that are carried out on visual input in the service of action involve 'just-in-time' computations based on the particular disposition of the goal object with respect to the actor, there would be no value and even a real cost in allowing this information to be accessible to conscious experience. Indeed, if such information were conscious, it would more often than not stand in real contradiction to the scene-based representations of the world offered by the ventral stream, which transcend particular viewpoints and thereby retain their utility over time. Far better then, that dorsal-stream computations should take place as automatically and unconsciously as those of the vestibular system in its efforts to help us maintain an upright and steady posture. In fact, there is now empirical evidence that the visual information that guides our movements is indeed unconscious. Normally, of course, we can consciously see our own arm as we reach toward a target, and we can consciously see other objects in the region of the target that could interfere with our reaching movements. But patients who show 'extinction' (typically following damage to the right parieto-temporal brain region) will often report not seeing a brief stimulus on the left when it is accompanied by a stimulus on the right (i.e., the presence of a stimulus on the right 'extinguishes' the perception of a similar stimulus on the left). Nevertheless, such patients will avoid colliding with an obstacle on the left when reaching toward a goal, even though they report seeing nothing in that part of space.

Although it is the ventral stream that delivers the contents of our visual consciousness, this does not mean that the workings of the dorsal stream play no role at all in the determining our awareness. For one thing, by virtue of helping to direct our eyes (and shift our attention) between different objects and locations in the environment, the area in the dorsal stream that mediates the visual control of eye movements (and covert shifts of attention) causes changes in the information that can be processed by the ventral stream, and thus the content of our visual awareness. The dorsal stream also participates in consciousness in other ways as well. People are typically aware of the actions they perform and, as has already been emphasized, the dorsal stream plays a critical role in the visual control of those actions. But there are some important distinctions to be made here. Although a person is often aware of the visual stimulus to which their action is directed, that visual awareness is mediated by processing in the ventral, not the dorsal, stream. The actual visual information used by the dorsal stream to specify and control the constituent movements of a goal-directed action (including an eye movement) is inaccessible to consciousness. Yet the compelling nature of visual consciousness makes it difficult to resist the intuition that it is one's perception of the goal object that is guiding one's action.

The neuropsychological studies of D.F. and other patients provides compelling evidence for the idea that ventral-stream processing is necessary for visual awareness. There is now more direct evidence, however, for the role of the ventral stream in giving rise to visual awareness, deriving from single-unit and lesion studies in the monkey, and neuroimaging studies in humans.

Although human subjects can readily report verbally on what they see, monkeys have to make their reports in other ways, having first been trained to do so. Monkeys, for example, can be trained in a detection task to press a touch screen to indicate whether or not a stimulus was presented somewhere on the screen. What is interesting is that even though monkeys with unilateral removal of primary visual cortex can point reasonably accurately toward a visual stimulus presented in their blind field, they will categorize the same stimulus as a blank (stimulus-absent) trial when faced with it in such a detection task. This result provides a compelling confirmation of the reality of blindsight, showing that visuomotor control can remain virtually intact despite the absence of visual perception. This dissociation provides strong empirical support for the thesis that visual perception depends on inputs to the ventral stream from primary visual cortex.

But perhaps the most convincing evidence for the relationship between awareness of a visual

stimulus and neural activity in the ventral stream has come from studies using the phenomenon of binocular rivalry, in which different visual stimuli (say a face and a cloudburst pattern) are simultaneously and independently presented to the left and right eye. When human observers are presented with such stimuli, they typically report fluctuations in what they see, sometimes reporting seeing a face and sometime a cloudburst pattern. Only very rarely do they see a 'blend' of the two stimuli. This technique has been used with monkeys by training them to make a manual response to signal a switch between seeing stimulus A (the cloudburst pattern) and stimulus B (the face). In a sense, the monkey was being asked to report on the content of its visual experience. As it turns out, single neurons in the inferior temporal cortex, a higher-order area in the ventral stream, that were tuned to one stimulus (say the face) fired more rapidly when the monkey reported seeing that particular stimulus. This correlation between perceptual report and firing rate was not nearly so strong for neurons in earlier visual areas and was completely absent in primary visual cortex. These striking results demonstrated for the first time that there is a direct link between the activity of particular neurons in the ventral stream and what an animal perceives.

More recent research with humans shows exactly the same thing. It has now become possible to record from single neurons in humans, typically in patients with epilepsy who have had electrodes implanted in their brains to localize the site where the seizure originates. Using a variant of the binocular rivalry paradigm, researchers have shown that visual neurons in the medial temporal lobe respond only to the perceived stimulus and never respond to the suppressed stimulus presented to the other eye. Similar effects have been observed in functional magnetic resonance imaging (fMRI) studies. For example, when volunteers in the fMRI scanner are presented with rivalrous images of a face and a building to the left and right eye, the activity in two areas in the ventral stream that are differentially selective for faces and scenes fluctuate in a reciprocal fashion with what the volunteers report seeing. Thus, when they report seeing a face, the 'face area' is more active and when they report seeing the building, the 'place' area is more active. The relationship between visual awareness and brain activity is not seen in the dorsal stream; even when observers are unaware of visual stimuli because of interocular suppression, the fMRI activation elicited by those stimuli is just as great as when those stimuli are consciously perceived.

Although all these studies (and many others) provide compelling evidence that visual consciousness is closely correlated with activity in the ventral stream, this does not mean that we are conscious of everything that is processed by this system. If this were the case, we would be overwhelmed with information, and consciousness could serve no function. In fact, even in the binocular rivalry experiments, activity was never completely abolished when the monkeys or humans reported not seeing the stimulus – it was just reduced. This may suggest that active inhibitory and/or facilitatory mechanisms are at work – and that for visual awareness to occur activity must exceed some sort of threshold. The neural origins of these inhibitory and/or facilitatory mechanisms are unknown, although regions in prefrontal cortex, medial temporal cortex, and the inferior parietal lobule (regions associated with the switching of visual attention) may play a significant role. But even though ventral-stream areas are subject to such modulatory effects from other brain regions, it is ventral-stream activity itself that determines the content of our visual awareness.

Perception, Action, and Illusions

Although much of the evidence for the neural organization of the two visual systems model has come from human neuropsychology and neuroimaging, as well as from work with nonhuman primates, evidence for fundamental differences in the metrics and frames of reference used by vision-for-perception and vision-for-action has also come from studies in normal human observers. These latter studies have also provided compelling evidence with respect to the dissociation between conscious visual experience and more 'automatic' visuomotor control.

One of the most striking examples of dissociations between vision-for-perception and vision-for-action in normal observers has come from work with pictorial illusions. For example, when

people are asked to pick up a target in the context of a size-contrast illusion, such as the Ebbinghaus Illusion (see Figure 4), their grip aperture is typically scaled in flight to the real not the apparent size of the target. Although grip scaling escapes the influence of the illusion, the illusion does affect performance in a manual matching task in which people are asked to open their index finger and thumb to indicate the perceived size of a disk. Thus, the aperture between the finger and thumb is resistant to the illusion when the vision-for-action system is engaged (i.e., when the participant grasps the target) and sensitive to the illusion when the vision-for-perception system is engaged (i.e., when the participant estimates its size).

This dissociation between what people say they see and what they do underscores once more the differences between vision-for-perception and vision-for-action. The obligatory size-contrast effects that give rise to the illusion (in which different elements of the array are compared) normally play a crucial role in scene interpretation and object identification, a central function of vision-for-perception. In contrast, the execution of a goal-directed act, such as manual prehension, requires metrical computations that are centered on the target itself, rather than on the relations between the target and other elements in the scene. In fact, if our visually guided movements were based on the relative rather than the absolute size of objects, then many of our everyday actions, from driving a car to picking up a wine glass, would be subject to critical errors. As it turns out, the true size of a target (for grasping at least) can be computed from the retinal-image size of the object coupled with an accurate estimate of distance based on reliable cues such as vergence of the eyes. Such computations would be quite insensitive to the kinds of pictorial cues that distort perception when familiar illusions are presented.

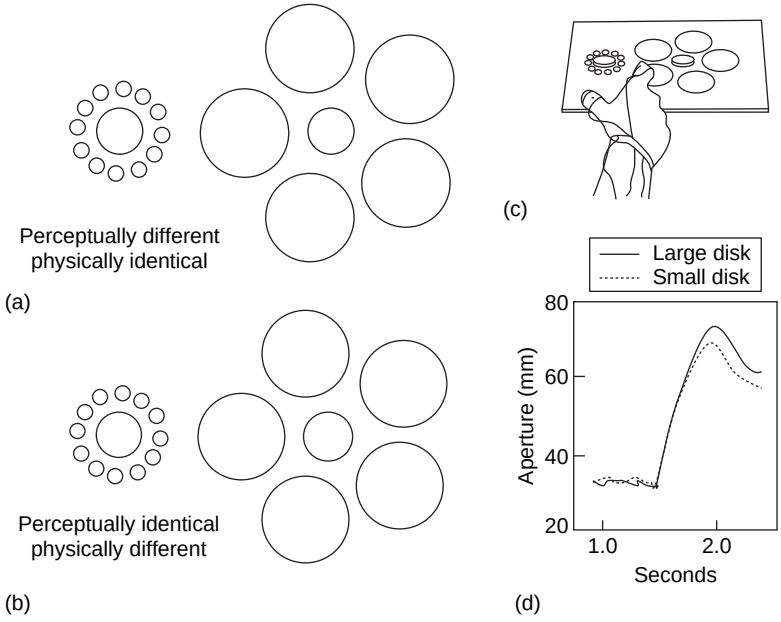


Figure 4 The effect of a size-contrast illusion on perception and action. (a) The traditional Ebbinghaus illusion in which the central circle in the annulus of larger circles is typically seen as smaller than the central circle in the annulus of smaller circles, even though both central circles are actually the same size. (b) The same display, except that the central circle in the annulus of larger circles has been made slightly larger. As a consequence, the two central circles now appear to be the same size. (c) A 3-D version of the Ebbinghaus illusion. People are instructed to pick up one of the two 3-D disks placed either on the display shown in panel A or the display shown in panel B. (d) Two trials with the display shown in panel B, in which the same person picked up the small disk on one trial and the large disk on another. Even though the two central disks were perceived as being the same size, the grip aperture in flight reflected the real not the apparent size of the disks. Reproduced from Aglioti S, DeSouza JFX, and Goodale MA (1995) Size-contrast illusions deceive the eye but not the hand. Current Biology 5: 679–685, with permission from Elsevier.

There has been considerable debate in the literature about whether or not grasping and other visuomotor responses (such as saccadic eye movements) are refractory to pictorial illusions. In some instances, particularly with unpracticed or awkward movements (where there is a lot of cognitive supervision), grip scaling appears to fall victim to size-contrast illusions like the Ebbinghaus. But of course the fact that actions are sensitive to illusory displays under certain conditions can never by itself refute the idea of two visual systems, which is securely based on a much larger body of evidence ranging from neuroimaging to neurophysiology. Indeed, it is unsurprising that perception affects our motor behavior, even within the context of the two-visual-systems model. After all, there are a number of situations, such as picking up a hammer or a cup of coffee, where our perception of the goal object will determine the kind of grip posture we adopt. The real surprise (at least for monolithic accounts of vision) is that there are a number of situations where visually guided action is unaffected by pictorial illusions. The actions that fall into this category tend to be rapid skilled actions, usually with the preferred hand. Nevertheless, it is fair to say that the claim that actions can be resistant to pictorial illusions is still regarded as controversial, particularly amongst those who favor a more monolithic account of visual processing.

Most of the pictorial illusions that have been used to dissociate vision-for-perception and vision-for-action distort perceived size by only a few millimeters. There are other illusions, however, which are not only much larger but which also show an actual reversal of depth. One particularly striking example is the hollow face illusion shown in Figure 5. In this illusion, knowledge about what faces look like impels observers to see the inside of a mask as if it were a normal protruding face, and the illusory face is perceived to be located several centimeters in front of the actual surface of the hollow mask. Despite the fact that observers cannot resist this compelling illusion, when they are asked to reach out quickly and flick off a small bug-like target stuck on the face, they unhesitatingly reach to the correct point in space (i.e., inside the mask). In other words, despite the presence of a strong hollow-face illusion, people direct rapid movements to the real, not the illusory positions of the targets. To do this, the visuomotor system must have access to a different source of visual information from that driving the illusion. In either case, people seem to be unaware of the veridical depth information they are using to control their flicking movements – and furthermore the use of this information does not 'break' the illusion. Again, this provides compelling evidence that one's conscious perception of a visual

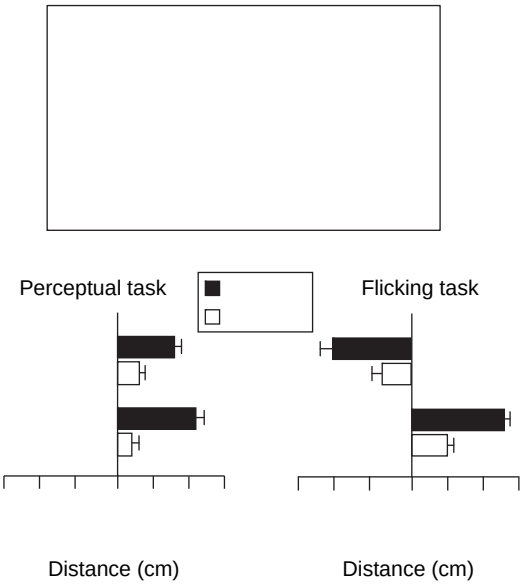


Figure 5 Perceptual judgments and visuomotor control with the hollow-face illusion. (a) A small magnet was placed on either the cheek or forehead of the normal face (left) or the hollow mask (right). Participants were required either to flick the magnet from the normal or illusory (actually hollow) face or to estimate its distance psychophysically. Inset shows a photograph of bottom-lit hollow face, in which the illusion of a normal convex face is evident. (b) (Left) The mean psychophysical (perceptual) judgments of the apparent position of the magnets on the illusory and normal face with respect to the reference plate from which the two displays either protruded or receded. Note that participants perceived the hollow face as protruding forward like the normal face. (Right) The mean distance of the hand at the moment the participant attempted to flick the target off the cheek or forehead of the illusory (actually hollow) or the normal face. In the case of the illusory face, the end points of the flicking movements corresponded to the actual distances of the targets, not to consciously seen distances. Error bars indicate the standard error of the mean. Reproduced from Króliczak G, Heard P, Goodale MA, and Gregory RL (2006) Dissociation of perception and action unmasked by the hollow-face illusion. Brain Research 1080: 9–16, with permission from Elsevier.

stimulus does not control visuomotor responses directed at that same stimulus.

Time and the Two Streams

As was discussed earlier, the visuomotor systems in the dorsal stream works in real time, using information provided in a 'bottom up' fashion from the retina. Thus, movements directed to remembered objects (objects that were present, but are no longer visible) might be expected to differ from movements directed to objects in real time. In fact, this is exactly the case. For example, when people reach out to grasp objects that were visible only a few seconds earlier, their grip scaling is now susceptible to pictorial illusions that were present in the display. Such sensitivity is to be expected, of course, if the programming and control of a delayed grasping movement is dependent not on processing in the dorsal stream but on memories derived from perceptual processing in the ventral stream. This conclusion is supported by the fact that the visual agnosic patient, D.F., who has ventral-stream lesions, demonstrates extremely poor size scaling of her grip when she attempts to grasp a target object after it has been removed from view – even though she shows excellent real-time control of her grasping. In one experiment, when a 2 s delay was introduced between viewing the object and initiating the grasp, there was no correlation at all between the size of the object and the aperture of her grasp in flight. An even more surprising result has been obtained in experiments with patients with lesions of the dorsal stream. Even though these individuals have great difficulty scaling their grasp when reaching out to grasp visible objects immediately after the objects are presented, they show a paradoxical improvement in performance if they are required to wait for 5 s before initiating their movement. Again these findings support the idea that the programming and control of delayed actions depends on information derived from earlier perceptual processing in the ventral stream.

Summary

There appear to be two ways in which visual information can influence and guide behavior. One is immediate and direct. For example, visual information about the size, shape, and disposition of an object with respect to the observer can be automatically transformed into the required egocentric coordinates for the programming and online control of a smoothly executed grasping movement. This kind of visual guidance, which is mediated by the dorsal stream of visual projections, needs to be quick and accurate, and evolution has ensured that it is. The visual information used by the dorsal stream is not accessible to consciousness – even though the actions controlled by that information clearly are. The other way in which vision can influence behavior is much less direct, and depends upon the construction and storage of visual representations that are initially processed in the ventral stream and reflect the structure and semantics of the scene facing the observer. The nature and intentions of subsequent actions will to varying degrees depend on the retrieval, and mental manipulation of, these representations. It is these representations that make up the visual contents of consciousness. The division of labor between the two streams is associated with fundamental differences in the metrics and frames of reference used by vision-for-action and vision-for-perception. Although both streams process information about the structure of objects and about their spatial locations, they use quite different modi operandi to do this. The operations carried out by the ventral stream use scene-based frames of reference and relational metrics; those carried out by the dorsal stream use egocentric frames of reference and absolute metrics. The two streams work together in the production of goal-directed behavior. The ventral stream (together with associated cognitive machinery) identifies goals and plans appropriate actions; the dorsal stream (in conjunction with related circuits in premotor cortex, basal ganglia, and brainstem) programs and controls those actions. Thus, a full understanding of the integrated nature of visually guided behavior will require that we specify the nature of the interactions and information exchange that occurs between the two streams of visual processing.

See also: The Neural Basis of Perceptual Awareness; Neuroscience of Volition and Action.

Suggested Readings

Bruno N, Bernardis P, and Gentilucci M (2008) Visually guided pointing, the Müller-Lyer illusion, and the functional interpretation of the dorsal-ventral split: Conclusions from 33 independent studies. Neuroscience and Biobehavioral Reviews 32: 423–437.

Carey DP (2001) Do action systems resist visual illusions? Trends in Cognitive Sciences 5: 109–113.

Goodale MA (2008) Action without perception in human vision. Cognitive Neuropsychology 25: 891–919.

Goodale MA and Milner AD (2004) Sight Unseen: An Exploration of Conscious and Unconscious Vision. Oxford: Oxford University Press.

Logothetis NK (1998) Single units and conscious vision. Philosophical Transactions of the Royal Society of London. B Biological Sciences 353: 1801–1818.

Milner AD (2008) Visual awareness and human action. In: Weiskrantz L and Davies M (eds.) Frontiers in Consciousness Research, pp. 169–214. Oxford: Oxford University Press.

Milner AD and Goodale MA (2006) The Visual Brain in Action. 2nd edn. Oxford: Oxford University Press.

Milner AD and Goodale MA (2008) Two visual systems re-viewed. Neuropsychologia 46: 774–785.

Biographical Sketch

Professor Goodale was born in England but was educated entirely in Canada. After completing his PhD in psychology at the University of Western Ontario, he returned to the United Kingdom where he worked as a postdoctoral fellow at the University of Oxford with Professor Larry Weiskrantz. After two years at Oxford, Professor Goodale accepted a position in the School of Psychology at the University of St Andrews before returning to Canada in 1977, where he now holds the Canada Research Chair in visual neuroscience at the University of Western Ontario. Professor Goodale is best known for his work on the functional organization of the visual pathways in the cerebral cortex, and was a pioneer in the study of visuomotor control in neurological patients. His recent research uses functional magnetic resonance imaging (fMRI) to look at the activity in the normal human brain as it performs different kinds of visual tasks. Professor Goodale serves on the editorial board of a number of journals including *Experimental Brain Research* and *Neuropsychologia*, and is the past-President of the *Association for the Scientific Study of Consciousness*. He has been recognized for his distinguished scientific achievements by the Canadian Society for Brain, Behaviour, and Cognitive Science, receiving the D.O. Hebb Award in 1999. He was elected as a fellow of the Royal Society of Canada in 2001.

David Milner was born in Leeds, England. After taking his first degree at the University of Oxford, he undertook training in clinical psychology, and then a PhD in behavioral neuroscience, both at the Institute of Psychiatry at the University of London. In 1970, he moved to an academic position at the University of St Andrews, Scotland. It was there that Dr. Milner first collaborated with Dr. Goodale – at that time on research into midbrain visual function in animals. In 1988, he first came across the now-famous patient D.F., who had just suffered catastrophic brain damage due to carbon monoxide poisoning. It was the remarkable early results from testing D.F. that stimulated the development of the idea that the two main visual pathways in the primate brain might play complementary roles in visual perception and visuomotor control, respectively. These ideas were presented in detail in the monograph *The Visual Brain in Action* (Oxford University Press, 1995). In parallel with this work Dr. Milner has been carrying out a program of research into the clinical condition of visuospatial neglect over the past 20 years. David Milner was elected in a fellow of the Royal Society of Edinburgh in 1992, and served as Editor-in-chief of the journal *Neuropsychologia* from 2000 to 2005. He has been director of the cognitive Neuroscience Research unit at Durham University, England, since 2000.

Perception: Subliminal and Implicit

M Snodgrass, University of Michigan Medical Center, Ann Arbor, MI, USA
E S Winer, University of Illinois, Chicago, IL, USA
N Kalaida, University of Michigan Medical Center, Ann Arbor, MI, USA

Glossary

Conscious perception index – Task used to measure conscious perceptual influences; usually direct.
Direct task – Requires discriminative judgments (e.g., identification) about relevant aspects of ostensibly unconscious stimuli themselves.
Exclusiveness problem – Apparent requirement that objective threshold approaches make strong assumption that direct measures are insensitive to unconscious influences; else, unconscious effects at objective thresholds seem intrinsically weak or impossible.
Exhaustiveness problem – Difficulty of ensuring that direct tasks actually assess all conscious perception relevant to putatively unconscious perceptual effects.
Indirect task – Either assesses influences of ostensibly unconscious stimuli on other stimuli (e.g., priming effects) or on judgments regarding irrelevant aspects of the unconscious stimulus itself (e.g., stimulus preference when only prior exposure was manipulated).
Null sensitivity problem – Difficulty of ensuring that all relevant conscious perception has indeed been eliminated; compounded by direct task measurement error.
Objective threshold – Operationalizes consciousness as stimulus intensity where direct task performance is at chance; more stringent than subjective thresholds.
Subjective threshold – Operationalizes consciousness as stimulus intensity where observers deny awareness; this occurs while direct performance still exceeds chance.
Unconscious perception index – Task used to measure unconscious perceptual influences; usually indirect, but not necessarily so.

Introduction

Implicit and/or subliminal perception paradigms attempt to study the unconscious effects of sensory stimuli (usually visual), utilizing exposure conditions that preclude their conscious representation. The goal is to prevent the occurrence of any potentially confounding conscious perceptual influences altogether, so that any obtained effects can be said with confidence to reflect purely unconscious perceptual processes. In typical unconscious perception paradigms, two tasks are used: one intended to index conscious perceptual influences and a second to index unconscious influences. The conscious perception index is usually a direct discrimination task such as stimulus detection (e.g., ''Has stimulus X been presented or not?'') or identification (e.g., ''Was stimulus X or Y presented?''). Presumably, then, if stimuli are consciously perceivable, observers should perform above chance on such tasks, whereas demonstrating chance performance suggests the absence of any relevant conscious perception. Because these tasks straightforwardly request intentional reports on stimulus dimensions relevant to ostensibly unconscious effects, they are generally regarded as the most sensitive conscious perception indexes. Conversely, the unconscious perception index is frequently an indirect task such as priming, wherein the unintended influence of an initial stimulus on the processing of a later stimulus is examined. Because unconscious influences are frequently thought to not require conscious intention to

manifest, indirect tasks are generally regarded as the best unconscious perception indexes. The usual objective is to obtain indirect effects despite chance performance on the direct task, thereby demonstrating unconscious perceptual influences on the former.

If successful, unconscious perception paradigms can not only inform us about the nature and scope of unconscious mental processes, but moreover have fundamental implications for the nature of consciousness itself. For example, it is widely held that more complex and flexible mental processes are possible consciously than unconsciously. By investigating what happens when stimuli are consciously versus unconsciously perceived, we can empirically examine these and other crucial questions. To clearly establish the presence versus absence of consciousness turns out to be much harder than it might initially seem, however. In particular, it is surprisingly difficult to convincingly demonstrate that putatively unconscious effects are not, instead, merely weakly conscious effects in disguise. This fundamental problem is not limited to unconscious perception; indeed, it continues to bedevil all fields that attempt to distinguish conscious versus unconscious influences (e.g., implicit/unconscious memory and/or learning). Now, given that unconscious perception is currently widely accepted, one could easily assume that this core methodological dilemma had somehow been settled long ago. Despite the apparently sanguine state of the field, however, this is clearly not the case – which is particularly unfortunate, given that the 120-year plus history of unconscious perception essentially consists of cyclically alternating periods of skepticism versus acceptance. Accordingly, the current positive consensus is superficial and runs the risk of simply continuing the boom-and-bust cycle unless further progress can be made on these foundational issues.

Perhaps even more importantly, careful consideration of the nagging problem of how to distinguish conscious and unconscious influences is not only methodologically essential but fundamentally informative in its own right, because it forces us to consider fundamental issues such as how to define consciousness, how conscious and unconscious processes interact, and the role of conscious control and intention. Indeed, there remains essentially no consensus on these essential issues, even among unconscious perception proponents, let alone remaining skeptics. Moreover, these continuing disagreements have vital implications regarding which data are valid and how they should be interpreted. Accordingly, great caution is advisable when assessing virtually all unconscious perception research, as dramatically different conclusions would ensue depending on exactly how their ostensibly unconscious status is understood.

How Should Consciousness Be Defined?

It is generally agreed that conscious perception covaries with stimulus intensity, usually manipulated by varying either stimulus exposure duration, the intensity of masking (initially presented stimuli become less consciously perceivable when immediately followed by a second, masking, stimulus), or both. Not surprisingly, strong stimuli are clearly visible, whereas conscious perception diminishes as stimulus strength is reduced. Finally, conscious perception appears to disappear entirely when stimuli are weak enough, and the point where this occurs is taken as the threshold for consciousness. But how exactly should this point be operationalized? Notably, there are two basic alternatives.

Subjective Threshold Methods

Perhaps the most straightforward approach is to simply ask people whether or not they can see the stimuli. These methods, derived from classical psychophysics, rely on self-reports of phenomenal states (i.e., subjective experience), and are now usually called subjective threshold methods. Such methods date from the very beginning of scientific psychology, and from their inception have yielded a striking effect: Performance on direct forced-choice discrimination tasks (e.g., choosing which letter is on a distant card; selecting which of two weights is heavier) remains above chance, even when observers convincingly deny any relevant conscious experience. These are canonical subjective threshold effects, and reflects how they were originally discovered in the late nineteenth century. At the same time, such findings illustrate that unconscious perception indices can also be direct

tasks, even though indirect tasks are typically used for this purpose.

Subjective threshold methods continued to flourish throughout the 1940s and 1950s, particularly in so-called New Look paradigms, which often combined now-unfashionable psychoanalytic hypotheses with experimental psychological methods. For example, it was shown that taboo (i.e., socially unacceptable, at least at that time) words had higher perceptual thresholds than less objectionable words, apparently demonstrating a simple form of unconscious defense. Strikingly, however, virtually all unconscious perception research essentially halted as modern psychophysics emerged with the development of signal detection theory (SDT). Although psychophysicists had long known that self-reports might not exclusively reflect observers' actual perceptions, but were perhaps also influenced by their willingness to respond in certain ways as well, there had heretofore been no clear method to disentangle these processes. Crucially, SDT finally succeeded in rigorously demonstrating that observers' discrimination reports did not, as it might intuitively seem, simply convey the contents of their conscious perceptual states, but instead were indeed the joint product of independent, separable perceptual and decision processes. With this fundamental insight in mind, it became clear that subjective thresholds might not demarcate some qualitative boundary between conscious and unconscious perception, but perhaps instead reflect differing levels of confidence associated with varying degrees of conscious perception. In other words, denying awareness might simply indicate very low confidence, not the complete absence of awareness. After all, very weak stimuli clearly produce remarkably faint, indistinct, and unclear conscious perceptual experiences that, not surprisingly, individuals are quite uncertain about. Accordingly, SDT demonstrated that subjective threshold phenomena might well be simply very weakly conscious after all. With taboo words, for example, people might simply want to be a little more sure before reporting salacious or profane words than more innocuous material. In the language of SDT, then, this apparently dramatic effect could simply reflect different response criteria, rather than actual differences in how conscious the stimuli were.

It is important to emphasize that, perhaps contrary to initial appearances, the SDT criterion artifact critique does not somehow contradict or deny observers' subjective reports. Rather, because SDT demonstrates that all reports are influenced by both perceptual and response criteria, alternative explanations in terms of the latter become straightforwardly possible. For example, even when neutral or innocuous stimuli are used, and hence there are no extraneous (e.g., embarrassment-related) influences on observers' response criteria, a very weak conscious perception will plausibly resemble no stimulus at all much more closely than it does some positive response option. This could easily produce denials of awareness on a subjective detection task, but yet could clearly arise from weak, below-criterion conscious perception. Taken together with the classic above-chance discrimination performance characteristic of subjective threshold effects, the SDT criterion artifact critique convinced most investigators that subjective threshold approaches were likely simply invalid, and unconscious perception research all but ceased for the next 15 years.

Objective Threshold Methods

With the advent of SDT and the clear problems with criterion-based methods, the revival of unconscious perception research in the mid-1970s and early 1980s instead arranged stimulus conditions such that observers not only denied awareness, but moreover could not perform above chance on direct discrimination tasks. Because such performance is behaviorally observable, these are called objective thresholds. Further, in contrast to subjective threshold approaches, requiring chance performance on direct tasks appeared to render them unusable as unconscious perception indexes. Accordingly, objective threshold paradigms rely heavily on indirect tasks (usually priming) to index unconscious perceptual influences. Notably, to arrange objective threshold conditions requires reducing stimulus intensity even more than when subjective thresholds are used; hence, objective thresholds are more rigorous. Unfortunately, this greater rigor seemed to yield much more mixed findings than those obtained using subjective threshold methods, leading many to conclude that genuine objective threshold results did not exist at all.

In the late 1980s, further important methodological criticisms emerged. For example, it is very difficult to ensure that true discrimination performance does not exceed chance (the null sensitivity problem), because any obtained performance value inevitably contains measurement error. Accordingly, even when observed performance is at chance, the true underlying state of affairs might reflect some small, but nonetheless genuine, smidgen of conscious perception. To make matters even more difficult, it is essential that the particular conscious perception index in question indeed taps whatever conscious perception that could conceivably be responsible for the ostensibly unconscious effects (the exhaustiveness problem). In other words, the conscious perception index must validly assess the right type or kind of conscious perception; otherwise, demonstrating zero discrimination ability, even if it were possible, is useless. Unless both these problems are resolved, then, it is still conceivable that weakly conscious perceptual processes could still account for putatively unconscious effects, even using ostensibly more rigorous objective threshold methods.

Importantly, however, these difficulties are just as problematic for subjective threshold methods. After all, they too must utilize exhaustively sensitive conscious perception indices. Moreover, as the SDT criterion artifact critique suggests, observers in subjective threshold paradigms might actually possess small but nonzero confidence, despite overall denials of awareness. Here again, then, true conscious perception might actually exceed zero, despite subjective reports. Quite recently, subjective methods have begun to take this problem much more seriously, and have increasingly turned to Type II SDT methods as a possible solution. In contrast to more typical Type I SDT tasks, wherein observers discriminate stimulus states in the external world (e.g., whether or not a stimulus has been presented), in Type II tasks observers make a further discrimination which arguably concerns their internal phenomenal state – namely, how confident they are in the correctness of their response. Although Type II methods are important, a deeper grasp of them requires technical analysis beyond the scope of this article.

So Which Methods Are Valid?

It seems that unconscious perception researchers face a Goldilocks dilemma: Perhaps subjective methods are too lax, whereas objective methods may be too stringent. Even more importantly, however, a closer look at subjective and objective methods suggests that they embody fundamentally different models of unconscious perceptual influences. Recall, for example, that the canonical subjective threshold effect is above chance performance on direct discrimination tasks, even when awareness is denied. Obviously, in subjective models, this above chance performance must be due to unconscious perceptual influences. But if this is so, requiring that direct discriminations not exceed chance, as objective methods do, would apparently reduce or even completely eliminate all influences – not only conscious, but unconscious too. Subjective threshold models inherently predict, then, that objective threshold effects should simply not occur or, at best, be weaker and less interesting versions of their more robust subjective brethren. In contrast, for objective threshold methods to be viable, they must apparently assume that direct discrimination tasks exclusively tap only conscious, not unconscious, perceptual influences – and, in turn, that unconscious perceptual influences manifest more freely on indirect methods such as priming. If so, this would apparently indicate that direct tasks are indeed not influenced by unconscious perceptual processes. This would, in turn, crucially imply that subjective threshold effects on direct tasks must actually be due to conscious influences after all, and hence that subjective methods are invalid. Consequently, subjective and objective methods are not just more or less stringent, as many assume, but actually imply fundamentally incompatible models, and hence cannot both be valid.

Modern Unconscious Perception Models

Given the above, any adequate unconscious perception model must not only address these formidable methodological difficulties, but deal meaningfully with their theoretical implications

as well. To date, three major attempts have been made to address these issues. Before considering these modern unconscious perception models, however, it is helpful to recall that any such model must, as its ultimate objective, falsify the null hypothesis, that is, the single process conscious-perception-only model. Moreover, because this model is simpler, it is preferable unless clearly refuted. Accordingly, greater specification of this model is important. Fortunately, certain properties of the conscious-perception-only model are widely agreed upon – namely, as noted earlier, that the degree and complexity of conscious perceptual processes are positively related to stimulus intensity. See Figure 1(a). This alternative null hypothesis model is kept in mind as the three recently proposed unconscious perception models are described.

Philip Merikle and Associates' Subjective Threshold Model

As with all subjective threshold approaches, this model holds that subjective effects are indeed

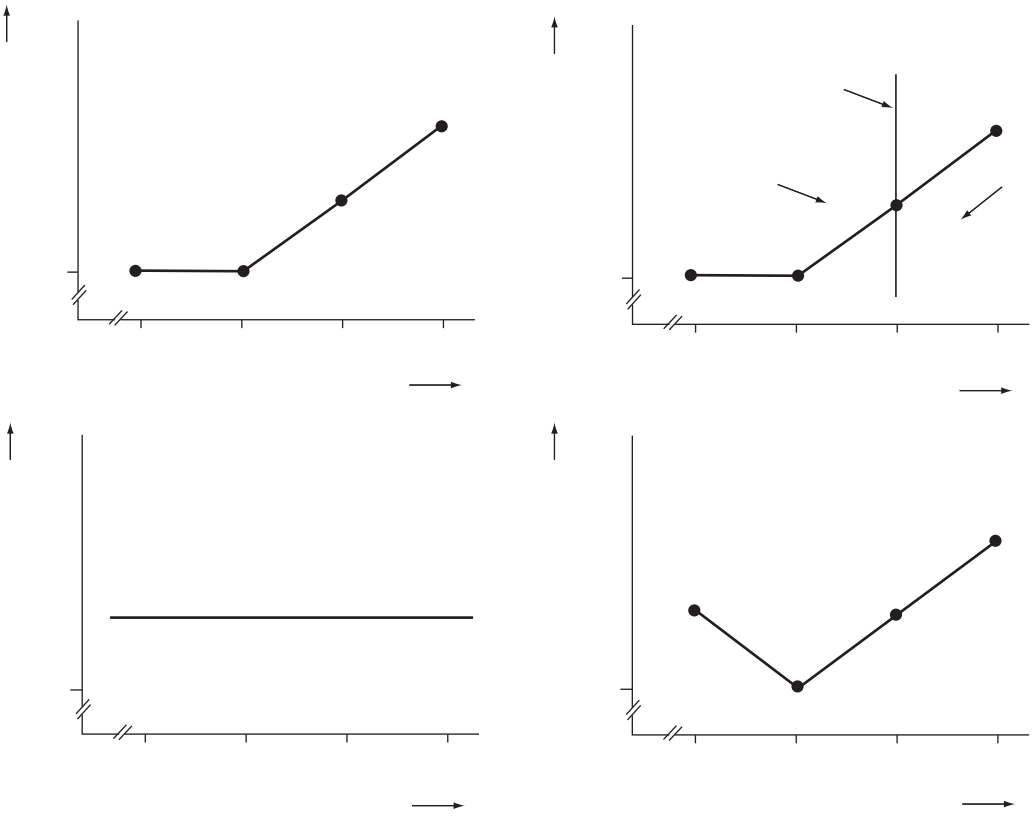


Figure 1 (a) The single-process conscious perception model's monotonic prediction for the relationship between the conscious and unconscious perception indices. Unconscious index performance remains flat and does not exceed zero until both the objective detection (ODT) and objective identification (OIT) are surpassed; then, the relationship becomes positive; (b) The subjective threshold model's monotonic prediction for the conscious versus unconscious perception index relationship. Although identical to the single-process conscious perception model's prediction, the subjective threshold model nonetheless holds that stimuli below the subjective identification threshold (SIT) are unconscious, whereas stimuli above this threshold are conscious; (c) The objective threshold/rapid decay model's monotonic prediction for the conscious versus unconscious perception index relationship. Here, unconscious index performance exceeds zero but remains flat throughout the entire stimulus intensity range, provided very rapid responses are obtained; (d) The objective threshold/nonmonotonic model's prediction for the conscious versus unconscious perception index relationship. This relationship is negative as stimulus intensity increases from the ODT to the OIT, reaching zero at the OIT. Beyond the OIT, the relationship becomes positive. Adapted by permission from Snodgrass M, Bernat E, and Shevrin H (2004) Unconscious perception: A model-based approach to method and evidence. Perception & Psychophysics 66: 846–867. Original: Figure 3, p. 856.

unconscious, and that alleged objective threshold effects are ultimately artifactual – thus explaining the latter's apparent empirical unreliability. The very considerable challenge, then, for this model is to overcome the SDT criterion artifact critique and convincingly rule out alternative weakly conscious explanations in some more definitive fashion. This is particularly difficult because, similar to conscious perceptual influences, subjective threshold effects are positively related to stimulus intensity (see Figure 1(b)). To meet this burden, these researchers have persuasively argued that subjective (and, for that matter, objective) threshold models must additionally demonstrate qualitative differences between putatively unconscious and conscious perceptual processes to provide essential convergent evidence for the claimed distinction. Further, requiring demonstration of qualitative differences seems especially reasonable given that they presumably do differ in important ways. After all, if not, why bother to distinguish them at all?

Qualitative differences in subjective paradigms. Notably, a variety of qualitative differences between subjective threshold and clearly conscious stimuli have indeed been reliably demonstrated. These follow a central and theoretically important theme – namely, illustrating various ways in which consciously controlled strategies are indeed applied when stimuli are clearly consciously perceived, but not when subjectively unconscious. Perhaps most compellingly, it has been repeatedly demonstrated that subjective threshold stimuli produce exclusion failure (i.e., responding with just-presented stimuli despite being instructed not to), whereas observers have little difficulty successfully excluding responding with stimuli that are clearly consciously perceived. This qualitative difference is particularly convincing, given the widely popular belief that complex cognitive processes, such as selective control over responding, can only occur when stimuli are consciously perceived.

Despite its intuitive appeal, however, exclusion failure and its cognates can also be criticized as a possible SDT criterion artifact. For example, exclusion is plausibly a criterion-dependent decision process; that is, observers may wish to be reasonably confident that they perceived the initial stimulus in order to exclude it as a response. If so, just as in the standard criterion artifact account, exclusion failure could simply reflect low confidence rather than genuine unconscious perceptual processes. Accordingly, these undeniably impressive qualitative differences might simply indicate two plausibly conscious processes, that is, conscious perception proper and criterion-dependent strategic exclusion processes. If the SDT-based critique of exclusion-related paradigms is correct, subjective threshold models need to demonstrate additional, more definitive qualitative differences; otherwise, the single-process conscious perception model is preferable. Perhaps in part due to their strong intuitive appeal, however, exclusion-related qualitative differences are still currently widely accepted.

Further, this alternative account illustrates a little-noticed but important weakness of the otherwise appealing qualitative differences approach: Demonstrating qualitative differences alone does not demonstrate that one or the other process is unconscious, but rather merely that they are distinct. With this in mind, it is probably necessary to independently demonstrate that the relevant stimuli are in fact unconsciously perceived in order for qualitative differences to truly carry any inferential weight. Despite this underappreciated problem, however, the qualitative differences approach is important and generative, and has proved to be widely influential. Indeed, in some fundamental sense it is necessary, because such differences must exist in some important sense for conscious and unconscious processes to really be distinct at all.

Blindsight and subjective threshold models. Blindsight is a striking neuropsychological syndrome in which rare individuals suffering from certain forms of brain damage perform quite well on simple direct discrimination tasks (e.g., position discrimination) despite vigorous denials of visual awareness, thus yielding a particularly dramatic subjective threshold effect. Further, blindsight effects differ qualitatively from what simply very weak conscious perception would produce. On the other hand, confidence ratings in blindsighted observers predict performance, in line with weak conscious accounts. Overall, the status of blindsight is unclear.

Although many researchers seem convinced that blindsight is genuinely unconscious, they may instead be qualitative different but nonetheless weakly conscious. After all, given the brain damage intrinsic to blindsight, one would expect any residual visual capacities to qualitatively differ – whether conscious or unconscious.

Anthony Greenwald and Associates' Objective Threshold/Rapid Decay Model

Apparently convinced by the SDT criterion artifact critique, these researchers assume that subjective threshold effects are indeed weakly conscious and hence that more stringent objective threshold approaches are necessary. They further posit that objective threshold phenomena are inherently very short lived; if so, this could explain why objective paradigms seemingly produce mixed results, alleviating this important concern. Moreover, if objective effects are indeed intrinsically very short lived, then their reliable measurement requires paradigms that engender very rapid responding – such as Greenwald and associates' response window procedure, typically implemented in conjunction with response conflict priming paradigms. Here, there are typically two stimulus categories, and the primes and targets either match or mismatch on each trial. Using such approaches, large and readily repeatable objective threshold effects have been obtained, apparently overcoming their alleged unreliability and supporting the hypothesis that such effects are indeed quite short lived. Strikingly, in these paradigms, typically no relationship is found between the direct and indirect measures (see Figure 1(c)). This pattern, which clearly differs from the single-process conscious perception model's (cf. Figure 1(a)), may indicate that conscious and unconscious perceptual influences are not only distinct but essentially independent. This conclusion, however, requires strong assumptions that the conscious and unconscious perception indexes are exclusively sensitive to conscious and unconscious perceptual influences, respectively. Alternatively, perhaps only conscious influences are involved, and the lack of relationship between the two indexes might result simply from severe time pressure on one task but not the other – thus accounting for the lack of a relationship.

Notably, response window effects seem largely primitive and unsophisticated (i.e., nonsemantic), instead largely driven by part-word analysis. This feature may not be intrinsic to unconscious perceptual processes, however, but instead possibly result from the severely limited processing time imposed by the response window procedure itself. Further, because part-word effects could be caused by partial stimulus identification, it may be that classification tasks (the usual conscious perception index in these paradigms, e.g., "Was the word positive or negative?"), which require considerably more information, may not be exhaustively sensitive to such less sophisticated information. If so, these priming effects may be weakly conscious.

The regression approach and the null sensitivity problem. Methodologically speaking, Greenwald and associates' regression approach has been widely influential. Here, regression analysis techniques are used to model the unconscious perception index (e.g., priming effects) as a function of the conscious perception index (e.g., word valence classification performance), with the former on the *y* axis and the latter on the *x*-axis. Analogous to regression analysis generally, here, the *y*-intercept estimates the value of the putatively unconscious effect when performance on the conscious perception index is zero (i.e., chance). Accordingly, obtaining significant *y*-intercepts apparently demonstrates unconscious perception; if instead only conscious perceptual processes were present, the *y*-intercept should equal zero.

Because *y*-intercepts can be generally obtained regardless of the overall mean on the *x*-axis variable (here, direct discrimination performance), the regression approach appears to avoid the null sensitivity problem. Specifically, demonstrating true overall chance performance now seems unnecessary, because one can simply ascertain the *y*-intercept. Notwithstanding this intuitively appealing framework, however, unfortunately the direct measure – here, the predictor on the *x*-axis still contains measurement error, the source of the null sensitivity problem to begin with. Such error in the predictor violates fundamental regression assumptions, and in single-predictor situations is

known to artifactually flatten regression slopes. In many common situations, such flattened slopes will in turn produce invalidly inflated y-intercepts. Although corrective adjustments have been proposed by Greenwald and associates, it is thus far unclear whether these modified procedures definitively solve the problem.

Moreover, examining y-intercepts is a special case of the general use of regression equations to predict one variable given knowledge of another. When the x and y variables are unrelated as in response window paradigms, using the regression equation to predict any value of y given x, including y-intercepts (i.e., predicting y when $x = 0$), may simply not be meaningful. Despite these potentially serious problems, however, the regression approach is an important advance which moreover draws valuable attention to the relationship between conscious and unconscious influences, a theoretically crucial but often ignored matter. The regression approach has been widely influential, and much recent research routinely presents such analyses.

Valence classification and other response conflict paradigms. Finally, in an important convergence, it is becoming increasingly clear that various other widely employed objective threshold priming paradigms, such as number classification (e.g., primes and targets are numbers above and below five) and motor activation (e.g., primes and targets are left- and right-pointing arrows) are response conflict paradigms, as is Greenwald and associates' valence classification paradigm. Notably, these kinds of priming tasks yield similar processing characteristics such as the lack of relationship between the conscious and unconscious perception indices, producing the typically flat regression slopes noted above. Moreover, these other priming paradigms, similar to valence classification, have importantly demonstrated that consciously intended response strategies (e.g., to classify targets in the instructed way) further affect how the unconsciously presented primes are processed as well. This is important theoretically, as it contradicts long-standing assumptions that unconscious processes work completely independent of, and are unaffected by, conscious intentions. Instead, such evidence suggests that conscious intentions indeed moderate unconscious perceptual influences, although not in the optional, flexible way generally observed with clearly consciously perceivable stimuli.

Snodgrass and Associates' Objective Threshold/Nonmonotonic Model

Snodgrass and associates suggest that objective threshold effects' apparent unreliability is in fact moderated by stimulus intensity, which in turn depends on the particular discrimination task used to determine the objective threshold. In particular, because detection tasks require less information than identification tasks, objective detection thresholds require more stringent exposure conditions than those necessary for objective identification thresholds. Because either task is plausibly exhaustively sensitive given typical unconscious perception indices, however, these task type differences in stimulus intensity have heretofore gone relatively unnoticed. Surprisingly, however, it turns out that stimulus intensity strongly moderates these effects: Under objective detection threshold conditions, unconscious effects are both sizeable and reliable; in contrast, under objective identification conditions, such effects are frequently absent.

In contrast, as stimulus intensities increase still more, beyond the objective identification threshold and into subjective threshold regions, the relationship becomes positive, then matching the familiar pattern characteristic of subjective and/or clearly conscious stimuli. It has been repeatedly demonstrated, for example, that direct identification and/or classification performance is strongly and positively related to priming effects when objective thresholds are clearly exceeded. Putting it all together, this constitutes a nonmonotonic relationship (first negative, then positive), see Figure 1(d).

Methodological implications. Collectively, these findings suggest a methodologically powerful negative relationship between the conscious and unconscious perception indices when stimulus intensities are between the objective detection and identification thresholds. This directly contradicts the positive (or at least zero) relationship predicted by the single process conscious perception model, providing considerable warrant for inferring a second, unconscious perceptual process. Along the same lines, the negative relationship

rebuts analogous predictions made by the exhaustiveness and null sensitivity problems. Further, the regression approach benefits when negative relationships are obtained, because slope flattening now underestimates rather than overestimates y-intercepts. Consequently, the presence of measurement error with negative relationships actually increases, rather than decreases, confidence in any obtained results. Moreover, the presence (vs. absence) of a relationship between the conscious and unconscious perception indices allows clearly valid use of the regression equation to estimate the y-intercept, whereas such procedures are dubious when the indices are unrelated.

Finally, negative versus positive relationships have important implications for evaluating the probative force of various qualitative differences. In particular, if ostensibly unconscious effects are weaker, less complex, or less controlled than conscious effects, such qualitative differences – however plausible – are questionable because one would expect this same pattern with weak versus strong conscious perception as well. Unfortunately, many qualitative differences are vulnerable to this alternative interpretation (e.g., Greenwald and associates' partial word effects; see above). On the other hand, obtaining the opposite pattern provides considerably stronger evidence for unconscious perception. For example, finding more complex effects under objective than subjective conditions (e.g., semantic vs. structural) would be harder to explain with solely conscious influences, given the latter's stronger stimulus conditions.

Process implications. The nonmonotonic relationship further suggests that conscious and unconscious influences may be functionally exclusive, such that when conscious perception is present, it typically overrides unconscious perceptual influences. At first, when only barely detectable, weak conscious perception is useless and hence interferes with unconscious effects, producing the negative relationship region. As the objective identification threshold is exceeded, however, conscious perception complex enough to drive typical unconscious perception index effects begins to emerge, and the relationship becomes positive. In contrast, the subjective and objective/rapid decay models assume that conscious and unconscious perceptual influences are either positively related or unrelated, respectively. Finally, the nonmonotonic model further implies that subjective and objective threshold effects index qualitatively distinct processes, rather than the latter simply being weaker versions of the former, as is often assumed.

Crucially, however, for the nonmonotonic relationship to be genuine, objective detection thresholds must indeed actually be below objective identification thresholds. If, instead, these thresholds were the same or even reversed, the nonmonotonic model would be weakened or refuted outright. Although available evidence is supportive thus far, further tests are clearly necessary. Moreover, it is important to note that objective detection threshold studies in general are relatively uncommon in recent years, probably because such thresholds are difficult to attain using typical computer monitors. Overall then, at present, the objective threshold/nonmonotonic model's empirical base is significantly less extensive than that of the subjective and objective threshold/rapid decay models, and must be seen as relatively provisional.

Unresolved Issues and Current Controversies

Various other fundamental issues are currently unresolved; these also have important methodological and theoretical implications.

The Role of Direct versus Indirect Tasks

In the vast majority unconscious perception research, the conscious perception index is a direct task, whereas the unconscious perception index is typically indirect. Although rarely discussed, this practice implicitly assumes that unconscious influences manifest more readily on indirect than direct measures – regardless of whether objective or subjective threshold methods are used. If reliable results in such designs are indeed obtainable, it implies that direct tasks indeed tap mostly or entirely conscious influences (cf. the exclusiveness problem). On the other hand, recent findings reliably suggest that unconscious perceptual processes do influence direct task performance after all, but in an unexpected fashion: Individual differences, sometimes interacting with task set, moderate the

actual direction of unconscious influences such that some individuals perform above chance (i.e., facilitate) while others actually perform below chance (i.e., inhibit). Such bidirectional effects do not affect the overall mean; hence, they are easily missed unless individual differences are examined, leading to the erroneous inference that unconscious influences are simply absent on direct measures. In contrast, indirect measures seem more frequently to produce primarily unidirectional, facilitative effects – with certain important exceptions (e.g., response conflict priming paradigms in certain situations). It is currently unknown why direct versus indirect task effects would differ in this way.

Is Attention Necessary for Consciousness?

Some findings suggest that similar findings are obtained when stimuli are presented outside of focal attention, on the one hand, and when they are attended but rendered unconscious using typical masking techniques. This has led some to suggest that attention is necessary for consciousness – although not sufficient, because other findings suggest that attention can be directed to stimuli that nonetheless remain unconscious. Crucially, however, these parallels have been obtained using subjective threshold methods; whether unattended versus objectively unconscious stimuli would also produce similar findings is not yet known. Consequently, the unattended/subjectively unconscious parallels could alternatively mean that both reflect weakly conscious, not unconscious, processes.

Unconscious Perception and Brain Processes

Not surprisingly, neuroimaging and other physiological measures are increasingly employed to investigate the brain correlates of unconscious perceptual processes. Findings from such investigations, typically indicating that unconscious processes are less complex and/or relatively localized, are consistent with currently influential Global Workspace theories which link consciousness to widespread frontoparietal brain activations. However, as with other qualitative differences, it is unclear whether this truly reflects differences between conscious and unconscious processes or may instead simply contrast weak versus strong conscious processes.

Implicit, Unconscious, and Subliminal: Synonymous or Distinct?

Frequently, these three terms are used interchangeably to indicate unconscious perceptual processing. Increasingly, however, it is clear that they differ in at least one important sense. For example, in implicit effects, the relevant stimuli themselves are clearly conscious – it is their influence on some subsequent task that is ostensibly unconscious. For example, in implicit memory paradigms, observers are perfectly conscious of initially presented word lists, but may not realize that these previously seen words influence how they complete word fragments later on. In contrast, unconscious perceptual paradigms – even when using implicit terminology – invariably attempt to prevent conscious perception of the critical stimuli proper, not just their subsequent influences on other tasks. Accordingly, unconscious and implicit perception may index distinct processes. For example, the former may assess phenomenal consciousness itself (e.g., visual qualia), whereas the latter may assess reflective (i.e., metacognitive) consciousness.

Finally, although subliminal is universally used synonymously with unconscious, its original derivation can create erroneous impressions. Literally, subliminal means below the limen, which in classical psychophysics reflected the stimulus intensity at which subjects would report a stimulus 50% of the time. If assessed, discrimination performance easily exceeds chance at such intensities. Indeed, from the perspective of SDT, the limen is closely related to the criterion, and hence subjective threshold approaches. In modern usage, however, subliminal terminology is often used in objective threshold paradigms, which do not have the limen implication.

Does Masking Technique Make a Difference?

Different masking methods are frequently used in unsystematic ways. Usually, some form of pattern

masking is used, where a particular kind of nonsense shape either immediately follows or precedes the primary stimulus, yielding backward or forward masking, respectively. Alternatively, other important techniques such as metacontrast masking and binocular rivalry are sometimes used. Notably, although masking theory suggests that the various techniques should produce importantly different effects, such differences are not yet completely clear, and certain techniques formerly believed to differentially limit unconscious processes apparently do not. Even so, as further investigations attempt to specify key features of unconscious processes, possible masking type effects should be carefully explored, both as potential confounds and as substantively interesting moderators in their own right.

Conclusions

Notwithstanding the longstanding and perhaps especially difficult methodological issues in unconscious perception research, in recent years important advances (including qualitative differences, the regression approach, and the potential usefulness of negative relationships) have emerged which may allow substantial progress in this notoriously controversial area. Nonetheless, there remains sufficient uncertainty that a distinguished minority is still unconvinced that genuinely unconscious perception occurs in any form, especially involving complex semantic processing. With this in mind, considerable additional work is essential to finally enable definitive progress. Along the way, however, it is likely vital to distinguish subjective and objective methods, because they may index fundamentally distinct processes, rather than simply being more or less stringent.

See also: Implicit Learning and Implicit Memory; Intuition, Creativity, and Unconscious Aspects of Problem Solving; Perception: Unconscious Influences on Perceptual Interpretation; Psychoactive Drugs and Alterations to Consciousness; Unconscious Cognition.

Suggested Readings

Abrams R and Greenwald A (2000) Parts outweigh the whole (word) in unconscious analysis of meaning. Psychological Science 11: 118–124.

Cheesman J and Merikle PM (1984) Priming with and without awareness. Perception & Psychophysics 36: 387–395.

Draine S and Greenwald A (1998) Replicable unconscious semantic priming. Journal of Experimental Psychology: General 127: 286–303.

Dulany D (1997) Consciousness in the explicit (deliberative) and implicit (evocative). In: Cohen J and Schooler W (eds.) Scientific Approaches to Consciousness, pp. 179–212. Mahwah, NJ: Erlbaum.

Holender D (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioral & Brain Sciences 9: 1–23.

Marcel A (1983) Conscious and unconscious perception: Experiments in visual masking and word recognition. Cognitive Psychology 15: 197–237.

Merikle P and Joordens S (1997) Parallels between perception without attention and perception without awareness. Consciousness and Cognition 6: 219–236.

Merikle P, Smilek D, and Eastwood J (2001) Perception without awareness: Perspectives from cognitive psychology. Cognition 79: 115–134.

Naccache L and Dehaene S (2001) Unconscious semantic priming extends to novel unseen stimuli. Cognition 80: 223–237.

Reingold E and Merikle P (1988) Using direct and indirect measures to study perception without awareness. Perception & Psychophysics 44: 563–575.

Reingold E and Merikle P (1990) On the inter-relatedness of theory and measurement in the study of unconscious processes. Mind & Language 5: 9–28.

Snodgrass M (2002) Disambiguating conscious and unconscious influences: Do exclusion paradigms demonstrate unconscious perception? American Journal of Psychology 115: 545–580.

Snodgrass M, Bernat E, and Shevrin H (2004a) Unconscious perception: A model-based approach to method and evidence. Perception & Psychophysics 66: 846–867.

Snodgrass M, Bernat E, and Shevrin H (2004b) Unconscious perception at the objective detection threshold exists. Perception & Psychophysics 66: 846–867.

Snodgrass M and Shevrin H (2006) Unconscious inhibition and facilitation at the objective detection threshold: Replicable and qualitatively different unconscious perceptual effects. Cognition 101: 43–79.

Biographical Sketch

Michael Snodgrass is a senior research associate and a codirector of the Laboratory for Conscious and Unconscious Processes in the Department of Psychiatry at the University of Michigan Medical Center. Dr. Snodgrass' primary research focus is unconscious perception, especially theoretical and methodological issues.

Natasha Kalaida is a project coordinator at the Laboratory for Conscious and Unconscious Processes in the Department of Psychiatry at the University of Michigan Medical Center. Her research interests include unconscious perception, especially translational applications investigating the causes and improved treatment of anxiety disorders.

E. Samuel Winer is a graduate student at the University of Illinois at Chicago, where he is completing his PhD in clinical psychology. His primary research focus is on unconscious perception in relation to individual differences.

Perception: The Binding Problem and the Coherence of Perception

T Schmidt, University of Giessen, Giessen, Germany

Glossary

Bálint–Holmes syndrome – A clinical condition following bilateral parietal brain damage, characterized by eye movement and reaching problems, severe spatial disorientation, and simultanagnosia.

Binding problem – The task of bringing together individually coded features so that they can be integrated into a single coherent object representation. Sometimes binding has to occur across space (grouping), time (object constancy), or sensory modalities (crossmodal binding). Solving the binding problem requires explaining how binding is established, signaled, and read out by the cognitive system.

Cardinal cell – A single cell signaling a distinct feature, object, or scene.

Feature conjunction – A stimulus that is defined by more than a single feature. Sometimes, illusory feature conjunctions are perceived.

Feature Integration Theory – Proposes that visual attention is used to link independently represented features to a particular spatial location in order to form an object representation.

Neuronal synchronization – Two or more cells are synchronizing when they tend to fire their action potentials at the same time.

Population coding – Groups of cells jointly signaling a feature, object, or scene. More efficient than coding by cardinal cells.

Recurrent processing – Occurs when information going out from one area keeps reentering that area by feedback loops. In contrast, the fast feedforward sweep is the first one-way pass of activity through the visual system before feedback information becomes effective.

Simultanagnosia – A clinical condition that can be described as an inability to perceive more than one object at a time.

Synesthesia – A benign variant of normal perception where the perception of certain objects is reliably accompanied by additional sensations, possibly in a different sensory modality.

Visual search – Psychophysical task where participants have to search a target item among a set of distractor items. Often search is assumed to be serial when target positions have to be searched one at a time, and parallel when search times do not depend on the number of distractors. In the latter case, the target seems to ‘pop out’ immediately.

Introduction

Most visual objects we encounter possess a multitude of properties. Looking out at a parking lot, you can see a number of cars, each one defined by its particular color, shape, and its unique place on the lot. Visually, these objects can be characterized as collections of visual features, which must be tied together in the correct way. ‘Binding’ refers to the process whereby separately analyzed features are combined to form a perceptual object. The ‘binding problem’ arises when features have to be assigned unambiguously to different objects without creating false combinations.

Figure 1 gives an example. The stimulus that the system tries to encode here consists of a

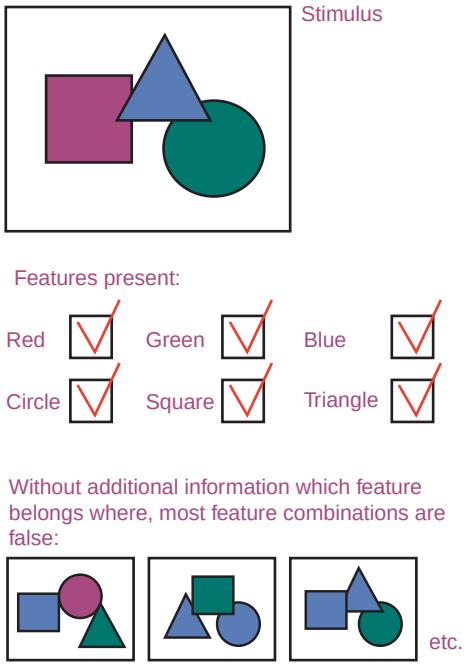


Figure 1 A binding problem occurs whenever a stimulus is dismantled into separate features that have to be recombined at a later stage. If the stimulus consists of a red square, a blue triangle, and a green circle, simple feature detectors are only able to signal that a square, a triangle, and a circle must be present, and that there are the colors red, blue, and green. Correctly conjoining these features to the correct stimulus objects and locations requires a binding mechanism. Without binding, the interpretation of the stimulus remains ambiguous.

red square, a blue triangle, and a green circle, each shape at a particular position. Assuming that the system is clear about the relative positions of all objects, there are still many ways to combine color and shape features with those positions: Without prior knowledge which feature belongs where, there are no less than 36 possibilities here, even with only three colors, shapes, and positions. Considering larger collections of objects, such as those cars in the parking lot, immediately shows that the number of false combinations is typically immense. If our visual system was only able to check which features are present, but not how they combine, there would be little chance of picking the correct combination. This suggests that the binding problem must be solved before accurate perception is possible.

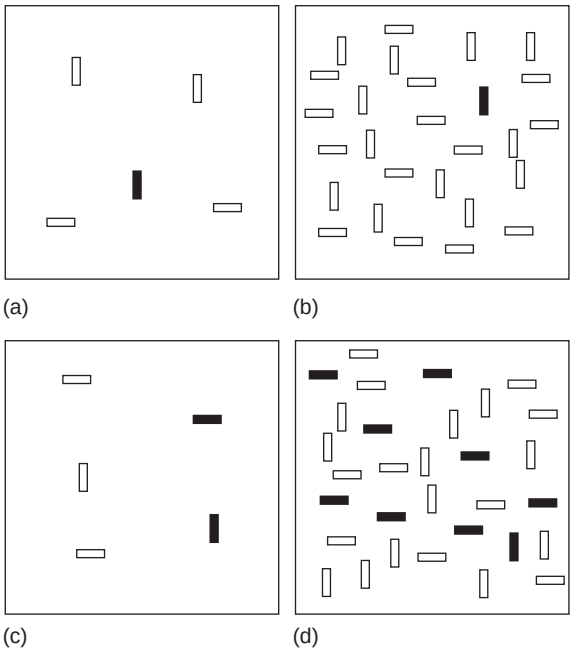


Figure 2 In these visual search displays, the target element is a vertical black bar embedded in a context of distractor elements. The participants' task would be to signal as quickly as possible whether the target is contained in a display. In panels (a) and (b), the target is easy to find because it is defined by a single feature (blackness) that is not shared by any distractor. As a result, the target seems to 'pop out' and can be found immediately, independent of how many distractors are present. In panels (c) and (d), the target is more difficult to find because it differs from the distractors by a conjunction of features (blackness and verticality), while each of these two features can occur in the distractors. FIT holds that in such a conjunction search, visual attention has to scan the display element by element, leading to search times that sharply increase with the number of distractors.

Empirical Evidence for Binding Problems in Visual Perception

Psychophysical Evidence for Binding Problems in Vision

There is solid evidence from psychophysical research that the system does face a binding problem, and sometimes fails to solve it. Much of this evidence comes from Anne Treisman's paradigm of 'visual search.' In visual search tasks, participants try to find a target item amongst a set of distractor items, and indicate as quickly as possible whether the target is present or absent. To get a

feeling for this type of task, take a moment to search for the black vertical bar in each of the panels of Figure 2.

You may have noticed that the target is very easy to find in panels (a) and (b). This is because the target differs from all the distractors in one distinctive feature: It is the only black item present and seems to 'pop out' of the display. Because search time is constant, no matter how large the number of distractors, this type of search is often called 'parallel search.' In contrast, search is difficult in panel (c) and especially in panel (d). The reason for this is that you have to search for a black vertical bar among black horizontal and white vertical bars. Because of this, you cannot use a single feature to distinguish the target from the distractors; instead, the target is defined by a 'conjunction' of features. Search times for conjunction targets are not only longer than pop-out searches, they also increase strongly with the number of distractors. This finding suggests that observers have to engage in a 'serial search,' where items are scanned one at a time until the target is found. This also agrees with the finding that search times increase twice as steeply with the numbers of distractors when the target is absent from the display. When the target is present, it is on average found halfway through the search, but when it is absent, all positions have to be searched before the search can be called off. This is not the case for parallel search: Here, observers are able to decide very quickly that 'nothing pops out,' and to declare the target absent. The fact that conjunction targets are more difficult to find than single-feature targets suggests that binding features into conjunctions is difficult for the visual system. As we will see below, Treisman's Feature Integration Theory (FIT) postulates that visual attention is needed to combine single features into a coherent object representation.

Treisman assumed that single features are represented independently of their spatial position, so that in the absence of binding, the features might become disattached from any specific position. Evidence for such 'free-floating' features comes from experiments where observers view brief displays of various objects under conditions of diverted or reduced attention. Various experiments showed that under such conditions, participants often report feature combinations that were not actually there, which suggests that the features of the presented objects had been combined erroneously. For instance, after a brief presentation of red circles, green circles, and green squares, subjects may report a red square; and after presentation of a collection of letters 'O' and 'R,' they may report having seen a 'P' or a 'Q ,' falsely conjoining the R's squiggle to one of the O's. These mistakes are called 'illusory conjunctions'; they are a clear indication that the visual system sometimes fails to bind features together in the correct way. Illusory conjunctions are frequent in patients with hemineglect (XXX), further indicating that attentional deficits may lead to binding problems. However, it is not entirely clear whether this failure occurs during actual perception or during the reconstruction from memory of what has been seen.

The Big Challenge: Binding Over Space and Time

The binding problem as classically formulated involves cases where a single, multi-featured object occupies a specific location in the visual field. In such cases, binding might be comparatively easy to explain because all the object features refer to the same spatial position. More difficult problems occur when binding has to be extended over space or time: these problems are known as 'grouping' and 'object constancy' and can be considered the toughest challenges in binding research.

For example, imagine an octopus passing by an unsuspecting diver. The octopus might undergo marked changes in shape and color in addition to its change in position; yet these changing features must continue to be bound to the same animal. In this case, location does no longer serve as an unambiguous cue to object coherence: The different parts of the octopus's body have to be assigned to the same animal, even though they occupy different spatial positions. When the octopus passes behind some partly occluding object, or some parts of its body occlude other parts, we still have to group the visible parts together. And when the moving animal is completely occluded for a stretch of time, we have to use memory to bind together the visual experiences of the animal before and after it passed the occluder. Young infants do not master this feat of object constancy over time. It also seems to

challenge the enemies of octopuses, whose most effective defensive weapon is an occluding ink cloud.

Single-Case Evidence for Binding Problems: Synesthesia and the Bálint–Holmes Syndrome

One instructive approach to the binding problem is to study populations of people where binding goes awry. 'Synesthesia' is a not-too-uncommon condition where people's perception of certain events or objects is reliably accompanied by the experience of additional sensations, sometimes in a different modality. The most common type of synesthesia seems to be grapheme-color synesthesia, where the viewing of letters or digits is accompanied by the experience of colors. These combinations are usually one-to-one, so that each grapheme goes with a specific color, and are stable over time. In other synesthetes, digits might be associated with certain sounds; or the spoken names of certain persons might go with different tactile sensations. Synesthetes differ in the way these concomitant experiences intrude into ongoing perception: Some of them perceive digits or letters as downright colored, some only see a weak colored halo around them, and others report 'knowing' which colors belong to the digits but do not report experiencing them phenomenally.

Synesthesia is a benign variant of human perception that bears no obvious disadvantages for the observer. To the contrary, some artists have used synesthetic experiences to inspire their art (e.g., color-sound synesthesia in painter Wassily Kandinsky, composers Olivier Messiaen and Duke Ellington, and electronic music avantgardist Aphex Twin). Some synesthetes with digit-color synesthesia actually use the additional color information when calculating numbers. Color-digit synesthesia can aid in finding a target digit in a visual search display, even though it is disputed whether synesthesia can elicit a full-fledged pop-out effect. Many studies show that synesthetic experiences can produce perceptual conflicts: For example, digits printed in their idiosyncratic synesthetic colors can be identified faster than digits printed in incongruous colors. Such results indicate that at least some types of synesthesia involve early visual processing. It has also been shown that synesthetically experienced colors activate brain areas implicated in color processing. Even though it is still unclear why this happens, the binding of extra features to certain perceptual objects is stable over time and seems to be hard-wired, perhaps by permanent connections between cells that lead to reliable coactivations between different brain areas.

In contrast to synesthesia, the Bálint–Holmes syndrome is a severely disabling clinical condition that is often seen after bilateral damage to the posterior parietal cortex (at the junction between the parietal and occipital lobes). It is disputed whether Bálint's syndrome really qualifies as a distinct diagnostic category, because it is a rather loose collection of symptoms that may occur after different types of lesion, not necessarily involving the parietal cortex. Patients with Bálint–Holmes syndrome have difficulty directing their gaze from one visual object to the next, and are impaired in performing visually directed hand and arm movements (optic ataxia) even though they are not afflicted with paralysis of the affected limbs. Some patients suffer from severe spatial disorientation: In extreme cases, they may be unable to find their way from their bed to the sink in the same room.

Importantly in our context, many of these patients also suffer from a condition called 'simultanagnosia,' which can be described as an inability to perceive more than one object at a time. For instance, when looking at the overlaid shapes in Figure 3, patients would only be able to consciously perceive one of them at a time. Other patients have difficulty reading because they tend to skip words and lines, and generally have to read sentences in a word-by-word

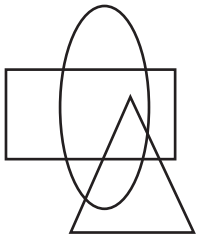


Figure 3 In this example, several contour shapes are superimposed, and we have to trace the contours across space to tell the shapes apart. The spatial position of object parts is not sufficient for assigning them to the correct objects here. Patients with simultanagnosia, a condition often associated with the Bálint–Holmes syndrome, have difficulty distinguishing more than one object at a time in displays like this.

fashion. Many of these problems may stem from the fact that the patients have difficulty localizing different objects or object parts, changing their direction of gaze, or shifting their focus of attention from one aspect of the scene to another.

In sum, synesthesia can be conceived as a bias to bind certain features to objects independently of whether this is required by the stimulus. In contrast, simultanagnosia is a condition where binding breaks down to a degree where only one coherent object can be maintained at a time. Both conditions show that binding is not a trivial feat for the visual system.

A Binding Problem in Early Vision?

The binding problem was first recognized in computer science and became a prominent research topic in psychology and neuroscience during the 1980s. One major force that gave the binding problem some urgency in cognitive neuroscience was the increasingly popular notion that the visual system disassembles incoming stimuli into different processing streams, each dealing with a separate feature such as color, orientation, or motion, in specialized regions of the cortex.

The most influential model of this sort was put forward by Margaret Livingstone and David Hubel in 1988. These authors proposed that different visual features (specifically, color, motion, orientation, and retinal disparity) were analyzed in anatomically separate processing streams. These streams would originate in different types of ganglion cells in the retina, run through different types of layers in the visual part of the thalamus, and enter different layers of the primary visual cortex (V1). Staining patterns for cytochrome oxidase (CO), an enzyme whose presence is associated with increased metabolic demand, were supposed to provide the anatomical scaffold for further segregation in V1 and the adjacent area V2. For instance, color would be processed in the CO-rich 'blobs' of V1 and then passed on to the 'thin stripes' of V2 and finally the putative 'color area' V4. Similarly, the motion stream would run through the 'thick stripes' of area V2 and on to the putative 'motion area' MT. Early single-cell studies largely supported this model.

In the Livingstone–Hubel model, different visual features are segregated into different anatomical compartments, and each feature is processed by a specialized subset of cells. The visual system is thus viewed as disassembling incoming visual information into their component features. This model has been massively popular, and its central tenets are still taken for granted by many researchers. Accordingly, many reviews of the binding problem use the disassembly metaphor as a point of departure: If the system dismantles incoming information into its component features, how does it later reassemble the processed features into perceptual objects?

However, more recent research has undermined many of the central tenets of the Livingstone–Hubel model. Careful anatomical studies indicate that signals from different thalamic layers do not remain segregated while filtering through the different layers of V1, but mix early and extensively. While it has been confirmed that CO-rich compartments in V1 tend to be linked to those in V2 (albeit not as exclusively as originally proposed), the evidence for a strict segregation of specialized cells into these compartments is mixed. For example, even though color-selective cells are more prominent in the 'thin' stripes of V2 than in the 'thick' or 'interstripes,' most of them are tuned to stimulus orientation as well.

The principal conclusion from this evidence is that neurons in V1 and V2 should not be regarded as simple feature detectors at all: It is unlikely that simple visual features are ever encoded completely independently of each other in early visual processing. Instead, cells are typically broadly tuned to more than one feature without being neatly segregated into different cortical compartments, which raises doubts about the disassembly metaphor in its strong form.

Other Types of Binding

Multisensory Integration

As noted above, the processing of different visual features is not strictly segregated in early cortical areas. In contrast, the processing of different sensory modalities, like vision, audition, and touch, clearly is, with each one originating in a well-defined primary sensory area of the cortex. To bring different modalities into register so that they refer to a single multisensory object, such as a honking car or an insect you see tickling your

foot, some type of binding has to occur. This integration of information from different modalities is called 'crossmodal binding.'

There are many behavioral indicators of crossmodal binding. For example, people can respond faster to simultaneous stimuli in two different modalities than to either stimulus alone (the 'redundant target effect'). Facilitating effects from redundant stimuli in different modalities are especially strong if these stimuli occur in close temporal and spatial proximity to the stimulus in the target modality. For instance, sensitivity in the detection of a visual stimulus in one modality can be enhanced by simultaneously presenting an auditory stimulus at the same location. Some effects of crossmodal binding are so compelling that they may lead to crossmodal illusions: For example, a ventriloquist's voice may appear to come from the puppet rather than the actual speaker, even though the two are clearly spatially separate.

Anatomically, widespread areas of the brain are involved in the analysis of sensory stimuli without being devoted to any sensory modality in particular. These include the premotor, cingulate, and prefrontal cortex. Other regions are especially important for combining input from two or more sensory modalities, for example, the superior temporal gyrus (vision and audition), many areas in the parietal cortex (vision, audition, and somatosensory information) as well as many subcortical structures (e.g., the midbrain's superior colliculus, which contains different retinotopic maps for vision, audition, and motor output, all superimposed in different layers).

Binding Problems in Action Control and Monitoring

Other types of binding problems occur in the control and monitoring of motor actions. The great majority of motor tasks require an accurate mapping of stimuli to responses. For example, when driving a car, you are expected to respond reliably to special visual cues like red traffic lights or stop signs. Laboratory experiments often involve arbitrary mappings between stimuli and responses. Some of these stimulus–response mappings are more difficult to perform than others, which can be due to so-called 'stimulus–response compatibility' effects. For example, responding with a left keypress to a square and with a right keypress to a circle is easier when the square is presented on the left-hand side of the display while the circle is presented on the right-hand side (this particular compatibility phenomenon is known as the 'Simon effect'). Other binding problems occur when simple motor actions must be integrated into more complex actions or action sequences.

Binding effects might also play a role in the monitoring of self-generated actions. Patrick Haggard and coworkers showed that when participants performed keypresses at times of their own choosing to elicit tone signals, they tended to perceive the time of action and the time of the action's effect as closer than they really were. The authors argue that this misjudgment reflects a binding together of actions and action effects.

Possible Solutions to the Binding Problem in Visual Perception

Crossmodal, motor, and sensorimotor binding effects pose intriguing challenges for any theory of binding, but they have been rarely addressed so far. Most theories of binding focus on visual processing. In what follows, we will assume that the visual system is confronted with an essentially infinite set of possible objects or scenes, and has to deal with this diversity given a finite number of visual cells that can encode various stimulus features. The system's task is to use information from the feature detectors to establish a representation of the visual scene where each detected feature is assigned unambiguously to the correct object. (As a caveat, remember our conclusion that most cells in early visual areas do not qualify as simple feature detectors, but are somewhat selective for several features.)

A number of different approaches to solving the binding problem have been proposed. To evaluate these proposals, it must be noted that a complete theory of binding has to account for three types of problems. First, there is the 'establishment problem,' which is essentially the binding problem proper: How does the system determine which features belong together and should be bound

into a single object? Second, there is the problem how binding is signaled in the brain: How does the representation of bound stimulus features differ from that of unbound features? Third, it must be explained how this information is read out by the rest of the system: How does the brain recognize and retrieve an integrated stimulus representation where features have been bound together? In addition, a convincing theory of binding should be able to explain grouping and object constancy in addition to the binding of stimulus features at a single location (none of the current theories has been designed to deal with binding problems in multisensory integration or action control). Table 1 reviews to what extent the different theoretical approaches address these issues.

Binding by Convergence

The simplest way to bind features together is to have all the detectors of the simple features converge on a single 'cardinal cell.' This cell would then serve as a detector of the compound object. This solution has been proposed in the 1950s by Horace Barlow and is known as the 'neuron doctrine.' It was propelled forward by David Hubel's and Torsten Wiesel's proposal that cells with complex stimulus selectivity (e.g., for motion in a specific direction) could be formed by convergence of cells with simpler properties, and by their continuous discovery of such cells since the 1960s. More recently, neurons in the inferotemporal cortex have been discovered that are amazingly selective for even very complex objects: They are especially sensitive to complex shapes or faces, or even respond selectively to pictures containing Bill Clinton (that's no joke!).

However, the notion of cardinal cells was soon caricaturized as postulating ''grandmother cells'' or ''yellow Volkswagen cells'', which would only fire when your grandma or her old car was in your field of view. Indeed, a convergence model of this sort runs into plausibility problems very quickly. The number of distinct objects or visual scenes that can be encountered is essentially unlimited, and while the number of distinguishable scenes is likely to be much smaller, it is still vast. Now the number of visually responsive neurons is vast as well. But even if the number of neurons was sufficient to represent each distinguishable scene on a one-to-one basis, it would still be unclear how this system could ever classify a completely novel object, say, a blue camel. If you encountered a blue camel for the first time, how could you expect to possess the appropriate blue-camel neuron to recognize it? And if you had such a detector, wouldn't that imply that you likely had green, red, and yellow detectors for other animals as well? Obviously, it seems implausible that the brain would harbor specialized detectors for objects it was likely to never encounter. A further problem occurs at read-out: If a given cardinal cell fires, how is the rest of the system to know which stimulus it is encoding, and in general which cells have to be accessed to extract some information of interest? While some grandmother (or Bill Clinton) neurons do seem to exist in the visual system, coding by cardinal cells cannot serve as a general coding strategy.

Binding by Population Coding

Even though some researchers continue to argue for a simple convergence theory of binding, most

Table 1 Aspects of the binding problem addressed by different theories

	Addresses establishment problem	Addresses signaling problem	Addresses readout problem	Addresses grouping or object constancy
Convergence coding	No	Yes	No	Yes
Population coding	No	Yes	No	Yes
Synchrony	No	Yes	No	Yes
Attention	Yes	Yes	Yes	No
Recurrent processing	Yes	Yes	Yes	Yes

authors dismiss this approach as implausible because it requires a vast number of detector cells to encode a combinatorial nightmare of possible feature conjunctions. A simple way to work around this difficulty is to consider the possibility that different cells jointly signal different objects. For example, consider a network of cells where each cell can be in one of two states, ON (= 1) or OFF (= 0). To represent 1024 objects, we could use 1024 cardinal cells, only one of which would be ON for any given object presented. But alternatively, we could assign one object to each possible state of the entire network. In that case, only 10 ON/OFF cells would be sufficient to represent the objects because they allow for $2^{10} = 1024$ states of the network (e.g., 1001011100 could represent one object and 0111010001 another). This scheme is known as 'population coding.' Of course, it is not confined to cells with binary states.

However, problems remain even if cardinal-cell coding is replaced by population coding. Just like the cardinal-cell model, the theory is able to tell us how binding is signaled, but not how it is established (i.e., how these cell properties were generated in the first place) nor how the relevant information can be read out (i.e., how the rest of the system knows which cell or population state is signaling what). Thus, both the convergence and the population coding theories remain confined to explaining object representation rather than being a solution to the actual binding problem. On the other hand, they are able to deal with binding across space and time, because cells coding different object parts or positions could readily converge onto cells that would signal an object independent of its particular state or position.

Binding by Synchrony

In the early 1980s, Christoph von der Malsburg proposed a different way how neurons could signal that they are encoding the same visual object. He proposed that cells encoding properties of the same object could fire in 'synchrony,' that is, emit their spikes at about the same time (say, within a time window of 5–10 ms). In a coding scheme like this, a small number of objects could be represented at the same time, each one signaled by a distinct firing frequency of the neurons encoding it. This theory has been playfully likened to the working of a transistor radio, which tunes into different stations by switching from one frequency band to the next.

Much of the empirical work on neural synchronization comes from Wolf Singer and his coworkers. In their studies, these researchers could convincingly show that synchronization does occur in the brains of cats and macaque monkeys. In a particularly striking study, recordings were made from pairs of cells in area MT of macaque monkeys, a visual area that contains many cells sensitive to motion in a specific direction. The receptive fields of the two cells were partly overlapping, so that a single moving bar could cover both receptive fields. Cells synchronized their spiking activity when a single bar stimulated both receptive fields, but not when two unconnected bars stimulated the receptive fields simultaneously – even if the unconnected bars were actually more effective than the single bar in evoking responses from the two cells. Results like this are consistent with the idea that synchrony between two cells serves to signal that both cells are dealing with the same object.

Even though the evidence for neuronal synchronization is solid, its interpretation has been controversial. For example, it is still unclear how synchronization comes to pass: Are cells actively synchronized by a top-down signal, is synchronization elicited by the visual stimulus, or do cells somehow synchronize each other? Obviously, if there was some such mechanism that could bind single-neuron feature detectors together by synchrony, it would be that synchronizing mechanism, not the synchronization per se, that solves the binding problem.

A similar shortcoming exists on the read-out side: How could neurons detect that some of their inputs are synchronous? Incoming electrical potentials sum more effectively when they arrive together, which means that inputs from synchronous cells might have more impact on the target neuron than inputs from asynchronous cells. But typical visual neurons receive input from several thousand cells, and it takes at least hundreds of incoming spikes to elicit a single spike from the target cell. This basically means that each incoming spike will always be simultaneous to many other incoming

spikes, even without synchronization, and that the advantage gained from active synchronization of some of these spikes might be small.

The synchrony model might be described as a population-coding model with the possibility of neuronal synchronization added. For this reason, it is at least as powerful as the population-coding approach, including its potential capacity to deal with binding across space and time. However, it also inherits some of its shortcomings: Just like the cardinal-cell and population-code models, the synchrony model is a theory of how binding is signaled, not how it is established or read out. The mechanisms that generate and decode synchrony have remained elusive so far, and it is still unclear to what extent neuronal synchrony plays a role in signaling binding.

Binding by Attention

Anne Treisman's Feature Integration Theory (FIT), developed in the context of visual search tasks, postulates that the correct binding of object features requires visual attention. Treisman starts from the assumption that the visual system is equipped with a set of 'feature maps,' with a different map for every conceivable feature (e.g., 'red,' 'green,' 'square,' 'circular'). These feature maps are connected to a 'master map of locations,' which encodes the exact positions of objects in visual space (Figure 4). When the search target is defined by a single feature, for example, 'red,' one only has to monitor the respective feature map and trace its activation to the target's position on the master map. This is what occurs in parallel search when the target seems to 'pop out' of the display (see 'Empirical evidence for binding problems in visual perception' section). However, when the target is defined by a conjunction of features, it is necessary to employ visual attention. Each object location on the master map has to be scanned by a 'spotlight' of attention, and for each of these positions the single features have to be collected from the feature maps and bound into an 'object file' that can be compared with a target template stored in working memory. As soon as the current object file matches the target template, the object has been found, and the search is terminated. Because the sequential scanning of object positions is a time-consuming process that depends on the number of positions to be covered, this model can explain serial search.

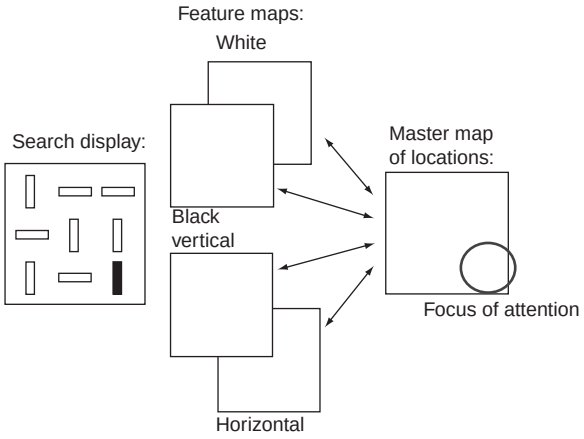


Figure 4 In FIT, stimuli are disassembled into their component features and represented in multiple 'feature maps,' which all feed into a 'master map of locations.' Visual attention must be directed to a position in the master map to extract all the features corresponding to this position and bind them together in a single object representation. In the original theory proposed by Anne Treisman and G. Gelade, location information was exclusively confined to the master map. Later variants and extensions of the theory differ in the way they allow positional information to be encoded within the feature maps themselves, making them more plausible physiologically.

Subsequent research has produced some challenges for Treisman's theory that led to refinements and variants. More recent theories extending Treisman's work (e.g., Jeremy Wolfe's 'guided search theory') view the master map as a map of object 'salience,' that is, the degree to which an object differs in its features from the remaining objects. They stress the possibility of controlled preactivation of feature maps, which allows searching for single features as well as feature conjunctions. Consistent with that idea, many studies showed that the distinction between parallel and serial search was not all-or-none: Sometimes, single-feature targets would be difficult to detect; sometimes conjunction targets would be surprisingly easy. Furthermore, it is unlikely that all the stimulus distinctions that can be used in visual search experiments are independently coded in the brain in separate feature maps: While this is conceivably so for the color or orientation of simple stimuli, it is less likely for features such as stimulus size, 'circularity,' or 'squareness.' Finally, the original notion that

features are encoded independently of stimulus location has become physiologically implausible, because many identified feature maps in the brain either show some retinotopy themselves, or are tightly linked and cross-activated by retinotopic maps.

In FIT, the deployment of visual attention is necessary to bind single features into coherent objects. A major advantage of FIT is that it can explain all stages of the binding process, not just the way binding is signaled. The trick is to use the 'spotlight' of visual attention to solve the binding problem sequentially at small portions of the image, where all the assembled features are simply assumed to belong to only one local object – in such a small, local image patch, there is simply no binding problem left to solve. The flipside of this strategy is that the theory cannot solve binding problems where the different features are at different locations, as is the case in grouping and image segmentation.

Binding by Recurrent Processing

Generally, visual areas receiving input from earlier areas also send massive feedback connections back to the originating areas. These feedback connections allow for 'recurrent activity' between different areas in the visual hierarchy. Victor Lamme and Pieter Roelfsema argue that recurrent processing is critical for the development of visual attention, visual awareness, binding, and grouping. Analyzing the response latencies of various cortical areas to a sudden visual stimulus, these authors argued that each new stimulus creates a wave of activation traveling from posterior to anterior areas, reaching most cortical areas within about 150 ms, including prefrontal and primary motor cortices. The authors estimated that this leaves cells with only about 10 ms time to pass their own activation on to later areas, which is about the duration of a typical interspike interval. Therefore, if most cells have to pass on their activation with the next spike fired, there is little or no time for them to integrate feedback from other cells. Based on this, Lamme and Roelfsema suggested that the first wave of visual activation travels through the system as a 'fast feedforward sweep' whose wave front is essentially free of intracortical feedback information.

Naturally, the only binding processes possible in this feedforward activation phase would be binding by convergence or population coding. However, Lamme and Roelfsema argue that feedforward processing is generally not sufficient for solving the binding problem or for generating visual awareness. Along with several other authors, they propose that conscious perception is possible only with recurrent processing of stimuli. Evidence for this view comes from studies indicating that visual awareness of a stimulus is suppressed if feedback loops from higher visual areas through primary visual cortex are disrupted at critical points in time, for instance, by a visual masking stimulus or by transient magnetic stimulation.

Recurrent processing is not only viewed as a precondition for visual awareness here, but also for fully solving the binding problem. In particular, the theory assumes that feature detectors (or populations) activated by the feedforward sweep send reentrant projections back to earlier areas, so that activation in any area can be readily linked to earlier topographic maps. In this way, features of a single object can be processed in separate areas but would all the time refer back to a single spatial location by recurrent interactions with a common spatial map. The analogy to Treisman's FIT is obvious; and indeed Lamme regards visual selective attention as a special form of recurrent activity.

Based on the distinction between feedforward and recurrent processing, Pieter Roelfsema has suggested a model that is also suitable for grouping and image segmentation. He suggests that grouping occurs in two steps: At first, there is only 'base grouping,' which is confined to the stimulus conjunctions that can be encoded by convergent or population coding during the feedforward phase of processing. This step is followed by 'incremental grouping,' whereby neurons encoding the same object are labeled with enhanced activity. Incremental grouping depends on recurrent processing and is able to actively bind different parts of an object together across space. As an example, let's take the problem of telling apart the three shapes in Figure 3. Incremental grouping would start as an activity enhancement somewhere on an outline contour and then spread along the contour until the entire shape was labeled. On its way around the contour, the process would take advantage of the fact that visual cells selective for similar contour orientations tend to wire together and coactivate each other.

Binding and Awareness

What is the relationship between visual binding and visual awareness? Some authors are convinced that solving the binding problem is the key to solving the consciousness problem. Indeed, we have the subjective impression of a world where all visual objects are seamlessly coherent, and we have seen that it takes sophisticated laboratory experiments to demonstrate that a visual binding problem exists at all. Also, some of the theories of binding also address visual awareness (e.g., the recurrent activation model).

However, the relationship between binding and awareness is less than clear. One major question is whether awareness starts before or after binding has been accomplished. For example, reports of illusory conjunctions suggest that we are sometimes phenomenally aware of features that have not yet been bound into a single object. Consistent with that observation, Victor Lamme and Ronald Rensink have both proposed theories where vision starts from a transient, phenomenally conscious representation where binding might still be incomplete, and is transformed by visual attention or recurrent processing into a more durable representation where tightly bound objects can be accessed for further cognitive processing. Note that in those theories, binding is not a necessary condition for awareness. At least for the moment, it therefore seems best to treat binding and awareness as separate problems.

Conclusions

'Binding' refers to a process whereby separately analyzed features or object parts are combined to form a single perceptual object. The 'binding problem' arises when features have to be assigned unambiguously to different objects without creating false combinations. Although it is unclear to what extent binding problems already arise in early neural coding, there is solid psychophysical evidence that they can occur in vision, multisensory integration, and action control. Moreover, binding breaks down in Bálint's syndrome, and can occur in unexpected ways in synesthesia. Proposed solutions include convergent cell specialization, population coding, temporal synchronization, attentional selection, and recurrent activation. While all of these approaches can explain how binding is signaled in the brain, some of them have difficulty explaining how it is detected, or how it is established in the first place. Because the relationship between visual binding and visual awareness is still unclear, it currently seems best to treat both problems separately.

See also: The Neural Basis of Perceptual Awareness; Neurobiological Theories of Consciousness.

Suggested Readings

Calvert GA and Thesen T (2004) Multisensory integration: Methodological approaches and emerging principles in the human brain. Journal of Physiology 98: 191–205.
Ghose GM and Maunsell J (1999) Specialized representations in visual cortex: A role for binding? Neuron 24: 79–85.
Gray CM (1999) The temporal correlation hypothesis of visual feature integration: Still alive and well. Neuron 24: 31–47.
Haggard P and Cole J (2007) Intention, attention and the temporal experience of action. Consciousness & Cognition 16: 211–220.
Hubbard EM and Ramachandran VS (2005) Neurocognitive mechanisms of synesthesia. Neuron 48: 509–520.
Lamme VAF (2003) Why visual attention and awareness are different. Trends in Cognitive Sciences 7: 12–18.
Lamme VAF and Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences 23: 571–579.
Rensink RA (2000) The dynamic representation of scenes. Visual Cognition 7: 17–42.
Rizzo M and Vecera SP (2002) Psychoanatomical substrates of Bálint's syndrome. Journal of Neurology, Neurosurgery, and Psychiatry 72: 162–178.
Roelfsema PR (2006) Cortical algorithms for perceptual grouping. Annual Review of Neuroscience 29: 203–227.
Shadlen MN and Movshon JA (1999) Synchrony unbound: A critical evaluation of the temporal binding hypothesis. Neuron 24: 67–77.
Sincich LC and Horton JC (2005) The circuitry of V1 and V2: Integration of color, form, and motion. Annual Review of Neuroscience 28: 303–326.
Wolfe JM and Cave KR (1999) The psychophysical evidence for a binding problem in human vision. Neuron 24: 11–17.

Suggested Listening

Gentle Giant (1972) Octopus. Hamburg, Germany: Repertoire Records.

Biographical Sketch

Thomas Schmidt graduated in psychology at the University of Braunschweig, Germany, in 1997. He received his doctoral degree in natural sciences in 2002 and his habilitation degree in 2007, both from the University of Goettingen. Since 2005, he is working as a senior researcher at the Department of General and Experimental Psychology at the University of Giessen, Germany. He has recently accepted a call for a professorship in Psychology at the Technical University of Kaiserslautern, Germany. His major research interests include unconscious perception and visuomotor priming effects, basic visual perception and attention, visual short-term memory, and spatial cognition. He is currently running a project funded by the German Research Foundation, entitled 'Response priming, attention and awareness.' His website: www.allpsych.uni-giessen.de/thomas.

Perception: Unconscious Influences on Perceptual Interpretation

B G Breitmeyer, University of Houston, Houston, TX, USA

Glossary

Blindsight – A neurological visual deficit, produced by destruction of the primary visual cortex, in which a patient reports no phenomenal awareness of stimuli that nonetheless can affect his/her behavior.
Extrastriate cortex – Also known as peri- or circumstriate cortex, the secondary and tertiary areas of cortical visual processing.
Masked priming – The ability of a stimulus whose visibility is suppressed or masked to prime the perceptual response to a probe stimulus.
Primary visual cortex – Also known as striate cortex, the initial site of cortical visual processing.
Qualia – The subjectively experienced aspects of the contents of visual cognition, for example, the color, shape, or motion of a stimulus.
Subcortical visual processing – The processing of information anywhere below the level of the visual cortex. This includes areas such as the lateral geniculate in the thalamus and the superior colliculus and pulvinar in the midbrain.
Visibility suppression – The suppression of the visibility of a stimulus via experimental procedures that include the use of various psychophysical techniques as well as the application of magnetic pulses at cranial points overlying specific areas of the visual cortex.

Introduction

Prior to the 1980s, few if any studies published in scientific journals or books dealt explicitly with consciousness. The scientific study of consciousness and of its companion, the unconscious, began during the latter half of the nineteenth century. In vogue in these early years of academic psychology and psychoanalysis, it subsequently fell into disfavor, especially in the Anglo-American academy, with the advent of behaviorism and its emphasis on overtly measurable or observable variables. It remained in this unfavorable state until the latter part of the twentieth century when the study of consciousness accelerated noticeably, especially in the past 25 years. However, historical events predating and anticipating this resurgence of interest were already evident in the decade following World War II. Theoretical developments in the study of information and communication, artificial intelligence, and neuroscience coincided and converged in the 1950s with what can be called the cognitive revolution in experimental psychology. One of the key features of this revolution was the study of human information processing and its control via attention, a phenomenon that since William James's *The Principles of Psychology* was viewed as intimately related to consciousness. For that reason, from the 1950s until the 1980s, very many investigations of psychological and cognitive phenomena implicitly focused on topics that currently are at the forefront of consciousness research. Since the 1980s and especially the 1990s, the Decade of the Brain, the number of scientific and scholarly studies and publications on consciousness has run into hundreds and perhaps thousands, most of them published in recently established journals dealing explicitly with consciousness. Moreover, these journals are affiliated with several professional organizations devoted to the scientific and scholarly study of consciousness.

The flipside of this resurgent interest in the scientific and scholarly study of consciousness has been a renewed interest in the study of

unconscious information processing, sometimes known by its misnomer subliminal perception. Subliminal (or unconscious) perception has been considered a misnomer for two reasons. First, along with the taboo of consciousness studies occasioned by the ascendancy of behaviorism, the very existence of subliminal perception and its putative influence on human action has been viewed extremely skeptically, if not denied outright, on theoretical, methodological, and empirical grounds. Initially much of the interest in subliminal perception was motivated by psychodynamic approaches. While they advocated a theory-driven hermeneutics which allowed interpretation of manifest behaviors or symptoms in terms of a hidden unconscious, for the more positivist-minded scientists such interpretations appeared much too facile and post hoc, reflecting the lack in these approaches of testable hypotheses clearly linking theoretical concepts to experimental outcomes. The collateral damage of this severe skepticism prevailed at least until the 1980s. Second, according to the definition adopted by some leading cognitive psychologists perception is the conscious registration of sensory activity. Thus one can make a distinction between vision and visual perception, audition and auditory perception, and so on, with the former of each pair of terms being more inclusive than the latter. This definition or perception is adopted in the following discussions and perforce excludes talk of subliminal or unconscious perception. Such exclusion of course leaves as a remainder a good portion of visual information processing that is not perceptual and that can influence human cognition and action, so to speak, below the conscious radar.

By definition, we cannot be consciously aware of the contents of these unconscious processes. Hence, any unconscious influences on perceptual interpretation cannot be assessed by conscious introspection. Fortunately they can be inferred from the results of studies that use a variety of objective procedures, including psychophysical, behavioral, electrophysiological, and brain-imaging techniques. The inferences drawn from these findings have been subject to ongoing theoretical, methodological, and empirical controversies. Although the following deals exclusively with unconscious visual information processing, the major issues discussed also apply to the study of unconscious processing in other sensory modalities.

Conceptual, Empirical, and Methodological Issues in Studying Unconscious Information Processing

Defining Unconscious Information Processing in Terms of Phenomenal Consciousness

Any discussion of unconscious influences on perceptual interpretation presupposes some understanding of unconscious information processing. Establishing such a conceptual handle requires arriving at workable and acceptable definitions of consciousness. Through our individual private experiences, all of us know what consciousness is; however, most of us have only vague or conceptually confused ideas about how to define it. Arriving at clear conceptual distinctions and definitions is therefore a preliminary but necessary exercise. As a starter, a most useful distinction is that between consciousness as a state and consciousness as a trait. Consciousness as a state globally characterizes our waking hours, when we are (more or less) alert, and depends primarily on the overall level of arousal or activation of the central nervous system. It therefore is something that we lack entirely when we are in a deep, dreamless sleep, in deep anesthesia, or in a coma. When awake there can be various gradations in the global state of consciousness that depend on effects of diurnal activation cycles, fatigue, disease, or drugs. It is possible that unconscious information processing occurs during deep dreamless sleep or anesthesia; however, it is difficult to assess such processing. The definition of consciousness as a state may arouse little controversy, since it can be tied to quantifiable indices of activation in the central nervous system. Consciousness as state can interact with cognitive contents. For example, a driver who hears a siren briefly becomes more aroused and alert to the ongoing traffic pattern.

Cognitive contents can be characterized by a number of features. For instance, some cognitive contents differ according to their specific modality such as audition, touch, or smell. Similarly

consciousness can be viewed as a trait or feature of cognitive contents. For that reason its definition is a bit more controversial. When cognitive contents are conscious they are deemed to possess qualia that confer on them a feely aspect, which is like knowing or having something from a first-person subjective perspective. One can be in a highly alert state of consciousness yet, due to a neurological disorder or an experimental manipulation, the experiential field can be lacking some qualia characterizing a particular object of cognition, for example, the shape or the color of an object and so on. Most current research in unconscious information processing, since it uses human observers who are in a fully alert state, explores this trait aspect of consciousness. Thus one can, by attenuating or suppressing certain qualia from experience, explore the remaining types and levels of information processing. When qualia are suppressed, we can speak of types and levels of unconscious information processing.

Even for the fully alert state, one can distinguish among various types of consciousness. Distinctions abound; and because they are highly controversial and debated among cognitive scientists, they cannot be covered exhaustively. To illustrate a sample of the controversies and problems that have arisen, we can focus on four categories of consciousness that recently have captured the attention of cognitive scientists and philosophers: phenomenal, access, reflective, and narrative consciousness. Each of these presupposes that one is in a conscious state.

Phenomenal consciousness

This type of consciousness can be considered as the totality of the subjective contents, that is, of experienced qualia.

Access consciousness

A second type of consciousness has been defined more inclusively in terms of control of cognitive behavior expressed through reasoning, recognition, identification, verbal reporting, or motor response. Hence, access consciousness, at least in limited form, can be considered as a consciousness that is separable from phenomenal consciousness or simply includes it as one of its particular forms. On the one hand, the control of some behaviors such as forced-choice discrimination of the location of objects or their direction of motion in cases of blindsight, where phenomenal consciousness (of qualia) is lacking, could qualify as access consciousness. On the other hand, it has been noted since the beginnings of experimental psychology, that the phenomenal visual field at any one time can contain much greater content than can be accessed through verbal (or other means of) report. For example, numerous studies of short-term visual store, also known as iconic memory, have demonstrated that one can be (briefly) aware of the extensive contents of the visual field but have only limited access to it (imagine having to report what is seen when a lightning flash briefly illuminates a richly varied landscape at night). Hence, phenomenal awareness is neither a necessary condition (e.g., blindsight) nor a sufficient one (iconic memory) for defining access consciousness.

These characteristics of access consciousness, while presumably establishing its independence from phenomenal consciousness, are puzzling to many cognitive scientists who find it hard to talk of consciousness without phenomenal awareness. Below it will become apparent why the positing of access consciousness without a phenomenal counterpart would seriously question an entire research program on unconscious information processing. For that reason, most investigators in this program subsume access consciousness under phenomenal consciousness rather than vice versa.

Narrative consciousness

As long as we are in a conscious state, we experience a flow or stream of contents passing through our awareness. These changing contents are viewed as the contents of a temporally evolving narrative.

Reflective consciousness

This type of consciousness is defined in terms of a monitored self-awareness of phenomenal consciousness. Such monitoring is most often the exception than the rule (although it may be a highly polished skill among practiced meditators). For example, a person may be phenomenally aware at any one time of a particular cognitive content, and his or her cognitive system may have access to this content for control of behavior, but s/he

may be unaware of the fact that s/he is having these experienced contents or that they change over time. In this sense, reflective consciousness can be regarded as a distinct form of, or separable from, phenomenal, access, and narrative consciousness.

Yet, at the moment that the reflective person becomes aware of the changing contents that is, flow or stream of his or her narrative consciousness, this awareness is itself a new content of narrative consciousness and, having been recognized as a new content, is also an expression of access and of phenomenal consciousness. For these reasons and until (1) debates among cognitive scientists converge on a settlement regarding controversies surrounding these (and other) concepts of consciousness and (2) distinctions among consciousnesses are sufficiently refined, the following schema is adopted: All types of consciousness are regarded as specific variants of phenomenal consciousness. Another important reason is that if we allow behaviorally relevant information somehow accessed in, say, blindsight to qualify as a form of consciousness, we have every reason to reject many, if not all, influences on human action or perceptual interpretation that are phenomenally unconscious. Thus, in the following discussions, lack of phenomenal awareness, that is, lack of experienced qualia, is taken as the *sine qua non* of unconscious information processing.

Methodological Issues Concerning the Study of Unconscious Information Processing

Given this specific definition of unconscious information processing, establishing or exploiting conditions which adequately comply with this definition poses several problems. In studies of unconscious visual information processing, either (1) an observer with a known neurological deficit, for example, blindsight, is used to study effects of a (total or partial) deficit-induced suppression of the visibility of stimuli or else (2) a normal observer is subjected to experimentally controlled conditions that (totally or partially) eliminate the visibility or phenomenal awareness of stimuli. In both cases, the problem is how to establish that phenomenal awareness of the stimulus is absent. Before dealing with the specific methodological rationales adopted in such studies, it pays to make some important distinctions among the types of stimulus information an observer has and lacks in her visual awareness.

The issue of criterion content

Criterion content refers to the information used by an observer to indicate a specific visual property of a stimulus. For instance, a stimulus is either present or absent in display; if present, it has a particular location in the display, a shape or form, luminance-contrast or lightness, color, direction of motion, speed of motion, and so on. Given the many dimensions or attributes that can specify a visual object, it is not surprising that an observer can be blind to, that is, phenomenally unconscious of, some stimulus dimensions without being blind to others. For example, a neurological patient suffering from achromatopsia or cortical color blindness would be phenomenally lacking hue distinctions but would retain phenomenal distinctions of luminance contrast, form, motion speed, motion direction, and so on. In other words, a stimulus may be visible in some parts or respects and invisible in others. Another example is that of a normal observer who, participating in a visual masking study, is unable to report the shape or the color of a masked target stimulus, but is able to report its presence or location in the visual field. Both examples illustrate that a stimulus may be visible according to one criterion content but not according to another. Thus, one can specify the existence of unconscious visual information processing along one stimulus dimension, for example, shape or color, while retaining conscious information processing of another dimension, for example, the location of the stimulus in the visual field.

This point needs to be made repeatedly, since many investigators have ruled out and continue to rule out the presence of unconscious information processing simply because some information can be used either (1) to detect the mere presence or location of a stimulus even when an observer is phenomenally blind to its other qualia such as color, shape, speed of motion, depth, and so on, or (2) to correctly infer the presence of a stimulus feature that is not phenomenally experienced. For example, in studies of backward visual masking, in

which a briefly flashed target is followed by a briefly flashed mask, the shape of the target, which is suppressed from phenomenal awareness by the aftercoming mask, nevertheless can be determined indirectly by attending to subtle apparent-motion cues that arise from the spatio-temporal target-mask sequence. Based strictly on behavioral report, the observer did make correct shape discriminations and thus can be said to have seen the shapes of the targets. However, based on phenomenal report, the observer did not at all see the shape of the target, but was able to cleverly infer it on the basis of another, actually irrelevant, phenomenal dimension of motion. This illustrates that, in studies of unconscious visual information processing, appropriate weight be ascribed to subjective phenomenal reports as well as to objective behavioral responses.

Indexing the absence of phenomenal consciousness

Research into unconscious information processing rests on specific assumptions about what constitutes conscious and unconscious processing. One approach relies on the notion of a threshold for conscious awareness. The assumption here is that each observer has some internal level of (neural) activation, t, above which processed information registers in consciousness and below which it does not. The idea is that conscious experience simply relies on more of the same activity that is present in unconscious processing. The concept of a threshold has a long history, and threshold measurements are ubiquitous in psychophysical and perceptual research. Since consciousness is characterized by subjective, first-person experience, one way to determine threshold for consciousness is to ask observers to report on their subjective experiential state. For example, when a stimulus is presented in the affected part of the visual field of a blindsight patient, s/he reports whether or not the stimulus is experienced; or when, in a normal observer, the visibility of a target stimulus is suppressed by a mask stimulus, the observer reports whether or not s/he has seen the target. However, measurements of subjective thresholds pose problems, since, as noted, it is hard to determine what criterion an observer uses to render his binary yes/no report. To deal with this problem, researchers can adopt what are presumably criterion-free or objective threshold measuring procedures. One way to accomplish this, for example, is to use a multiple-alternative (usually a two-alternative) forced-choice procedure, in which an observer is required to report, and guess if need be, during which of several possible cued time intervals the target stimulus was presented. Another example is to require the observer to indicate which of two (or more) equally likely shapes characterize the target stimulus. Other stimulus attributes such as location, color, size, luminance contrast, and so on, could also be used.

Although objective threshold measures may be preferable to subjective ones, other approaches to measuring stimulus detection reject the entire notion of fixed internal thresholds. The theory of signal detection is such an approach. The notion here is that an observer s sensory system processing a stimulus event is characterized not by some internal threshold but rather by the following assumed features: (1) the existence of (intrinsic or externally introduced) noise in the detecting system, (2) the level of noise activation varies randomly (with a Gaussian or normal probability distribution), (3) a stimulus, that is, a signal, adds a fixed amount of activity, indexed by a sensitivity parameter known as d' (d prime), to the noise response of the system, (4) the existence of an adjustable criterion level of activity, indexed by a parameter known as b. When an activation level produced by the signal (plus the ever present noise) is below b, the system decides that a stimulus has not been presented; when above b, it decides that a stimulus has been presented. Finally, (5) the criterion level b can be changed by contingencies based on the relative frequency of a signal occurrence or on payoffs and penalties regarding correct and incorrect decisions and the expectancies. Within this framework, conscious processing reflects a detecting system in which a stimulus produces a $d' >> 0$; whereas unconscious processing reflects a detecting system in which a stimulus produces a $d' \square 0$.

Two problems characterize this approach. One is statistical in nature. Determining that $d' \square 0$ is equivalent to confirming the null hypothesis, a procedure that demands very strong statistical power so as to maximally reduce the probability

of a Type 2 error, that is, of falsely concluding that a failure to report a stimulus indicates absence of its conscious registration. The other problem is that whether or not a stimulus event is detected depends not only on d' and b but also, as noted above, on the criterion content; that is, on the particular stimulus information used by the observer to make his responses. For instance, the information content of interest may be the shape of a stimulus. While its shape may not register at all in consciousness, its location or presence may. Here one could wrongly conclude on the basis of one stimulus attribute or criterion content, for example, its location, that another stimulus feature, its shape, was visible when in fact it was not. This is equivalent to committing a Type 1 statistical error, that is, rejecting the valid null hypothesis that no shape was seen, when in fact it was not. The solution to this problem is to note that detection of each stimulus attribute, A, can in principle be characterized by its own sensitivity and criterion level, d'_A and b_A. This can be illustrated by considering the shape or form attribute of a stimulus. As a particular example, two forms, a diamond and a square, serve as signal presented on half the trials (signal trials), and an eight-sided star (combination of diamond and square) serves as a neutral nonsignal in the other half of the trials (noise trials). On each trial an observer is asked to decide if the stimulus is one of the signals (a diamond or a square) or else the neutral star. Here, on each trial an observer may be able to see that some stimulus was presented but nonetheless not be able to see which stimulus shape it was. Thus, the d' for shape would be close to 0.

The problem of exhaustive exclusion

The above also addresses a related problem relevant to both the threshold-determining and the signal detection approaches. In order to assure that a stimulus is not consciously processed, it is deemed essential according to some current methodological accounts that measures used to assess stimulus detectability be exhaustive. That is to say, according to the exhaustiveness criterion, ideally only if all possible measures of detectability fail to yield evidence of conscious registration of a stimulus can one conclude that its processing is unconscious. The problem here is that one can never be assured that all such measures have in fact been used. One can circumvent this problem by noting that one need not eliminate the visibility of all aspects of a stimulus but rather focus only on those aspects that are relevant. For example, recalling that if one is interested in studying unconscious processing of color, say, in an achromatopsic patient, conscious detectability of the stimulus along another attribute such as shape, surface contrast, or location need not be eliminated, unless such an irrelevant attribute provides information that can be used to infer the (unseen) color of a stimulus.

Indirectly indexing the presence of unconscious information processing

While the presence of conscious processing can be assessed directly by subjective report, the presence of unconscious processing, by definition, cannot. Hence, indirection is a key methodological strategy for studying unconscious processing. To realize this strategy one relies on what are known as dissociation phenomena. As an example, an achromatopsic patient who has no subjective experience of color nonetheless can, under appropriate experimental conditions, render better-than-chance wavelength discriminations. In other words, the objectively measured presence of wavelength-discrimination abilities is dissociated from a lack of subjective color experience, indicated by his inability to name the color of a stimulus. Similarly, in a normal observer a stimulus or one of its properties can be rendered invisible by some experimental procedure; nonetheless, the unconscious processing of that stimulus or property can affect the response to a subsequently presented visible probe stimulus. Here, there is dissociation between the lack of visibility of a stimulus or of one of its properties and the presence of an influence it has on the processing of another stimulus.

Under some special circumstances one can create what are known as double dissociations between the visibility of one, the inducing stimulus and its effects on another, the probe stimulus. Here, on the one hand varying the visibility of an inducing stimulus is not associated with concomitant variations of its ability to influence the visibility of, or response to, the other, probe stimulus.

That is, changes of the inducing stimulus s visibility are dissociated from its ability to influence the processing of the probe stimulus. On the other hand, it is possible to obtain variations of the visibility of the probe stimulus without also varying the visibility of the other, inducing stimulus. That is, changes of an inducing stimulus s influence on the visibility of the probe stimulus are dissociated from the visibility of the inducing stimulus. For instance, the physical luminance contrast of the inducing stimulus can be varied and thus produce variable effects on the visibility of, or response to, the probe stimulus, even though the visibility of the inducing stimulus, regardless of its physical luminance contrast, is suppressed by a mask.

The use of normal versus neurologically impaired observers

The influences of unconsciously processed visual information on perceptual interpretation can be measured by studying either neurologically impaired or normal observers. When using the former observers it is, of course, important to diagnose as thoroughly as possible the nature of the neurological impairment so as to assess which visual functions are impaired and the extent to which they are impaired. Although there may be overlap of cortical damage and impairments that manifest among different classes of patients, the classes can nonetheless be distinguished. For example, blindsight patients suffer from cortical damage that differs from that of achromatopsic (color-blind) patients, whose damage in turn differs from that of akinetopsic (motion-blind) patients, whose damage in turn differs from form-agnosic patients, and so on. Correspondingly, blindsight patients manifest blindness symptoms ranging over an extensive set of stimulus properties, whereas achromatopsic patients are blind only or primarily to color, akinetopsic patients to motion, form-agnosic patients to shapes of objects, and so on. It should be obvious that such patients can be very useful to studies of unconscious and conscious visual information processing.

However, one need not rely solely on study of neurologically impaired patients. Normal, neurologically intact observers also can be used. Through application of clever experimental techniques or paradigms developed and tested over the years, one can temporarily render the visual system blind to a stimulus or to one or more of its properties. Most of these techniques, some of which have been in use for and refined over decades, rely on psychophysical procedures in which one stimulus renders another one invisible. Other, more recent techniques render the cortical visual system transiently blind by producing magnetic pulses that pass through the cranium and briefly disrupt the underlying cortical visual processes. The methods are described below.

Methods of rendering stimuli invisible

Psychophysical methods rely on the use of experimental procedures that promote the perceptual disappearance of a stimulus, resulting in its being processed unconsciously. The influence of the suppressed stimulus on the perception of visible probe stimuli can then be assessed.

Binocular rivalry

When two similar images are presented separately to the two eyes the observer typically reports having a single fused percept that is characterized by a vivid sensation of stereo depth. However, when the two images are too different or disparate, they do not allow perceptual fusion. Although this can results in diplopia, commonly known as double vision, the result is that the visibility of one of the images often is suppressed. This phenomenon can be amplified by presenting very disparate images to the two eyes, for example, a vertical grating to the left eye and a horizontal grating to the right eye, as shown in the first, top panel of Figure 1. The resulting percept is not a fused grid but rather rivalrous perceptual fluctuations. Here, the image presented to, say, the left eye dominates perceptually for a short period while simultaneously the image to the right eye is perceptually suppressed; then the image to the right eye dominates while that to the left is suppressed, and so on. While the image to one eye is suppressed, a test stimulus presented to that eye will also be perceptually suppressed.

Perceptual fading of semistabilized images

In order to maintain visibility of a stimulus, its image on the retina must regularly shift over the

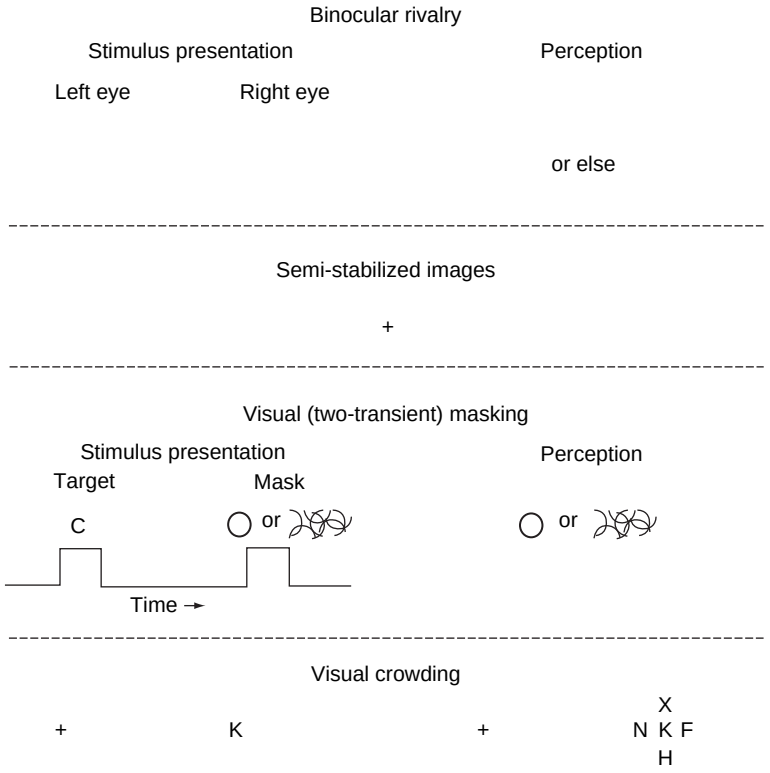


Figure 1 First or top panel: on the left is a schematic representation of two typical stimuli presented to the separate eyes to induce binocular rivalry; on the right: depiction of alternating eye-dependent percepts produced during binocular rivalry. Second panel: depiction of a visual display used to produce semistabilized retinal images. The observer should fixate the cross as rigidly as possible and, without moving his or her eyes, note what happens to the two faint stimuli to the left and right of fixation. Third panel: schematic depiction of typical visual stimuli and their time sequence used to produce visual masking. The first, target stimulus, a letter C, can be followed by the second, mask stimulus which can either be a surrounding stimulus such as a larger letter O or a spatially overlapping stimulus composed of random – in this case – arc-like elements. Fourth or bottom panel: on the left a depiction of a visual display consisting of a fixation cross and an isolated target letter, here a K, to the right of fixation, and on the right the same display but now with four other letters crowding the target letter.

retinal receptors; otherwise it fades and disappears from sight. One can exploit this phenomenon by, for instance, having the observer maintain his visual fixation as constantly as possible on, say, a tiny cross in the center of his visual field while two other stimuli are present at some distance to the right and left from the fixation cross, as shown in the second panel of Figure 1. What typically occurs when one fixates as rigidly as possible on the cross is that eventually one or both of the two eccentric stimuli will perceptually fade from view for a while, then reappear, then fade from view again, and so on. This type of fading as well as the aforementioned binocular-rivalry suppression can be amplified or prolonged by several additional methods listed below.

Motion-induced blindness

When, an array of, say, three stationary yellow dots, located at the vertices of a notional triangle is presented within in a larger array of coherently moving blue dots, the three yellow dots will, after a small while fade either singly, pairwise, or as a unit.

Flash suppression

As with binocular-rivalry suppression, the visibility of semistabilized images fades and reappears according to a more or less random, stochastic schedule. For that reason the exact timing of the fading cannot be determined. The suppression of the visibility of semistabilized as well as rivalrous images can, however, be temporally triggered and

prolonged by transiently flashed stimuli. For instance, in procedure known as continuous flash suppression, the rivalrous stimulus presented to one eye can be suppressed continuously by a stimulus flashed repeatedly to the other eye. Similarly, the fading of an eccentric semistabilized image can be triggered and prolonged by a briefly presented nearby stimulus. The visual stimulus transients presented either as a single flash or as a train of repeated flashes act as visual masks that suppress the visibility of a rivalrous or a semistabilized image that is physically present for a fairly long duration, sometimes on the order of a minute or so.

Visual (two-transient) masking

However, there also is a way to suppress the visibility of a stimulus presented briefly, say, on the order of a few tens of milliseconds. The typical procedure, as shown in the third panel of Figure 1, is to briefly present one stimulus, designated as the target, which is followed at brief intervals by a second stimulus, known as the mask. This procedure is known as backward masking. The mask can either spatially overlap the target or surround it. In either case, by choosing appropriate stimulus parameters, the effect of the second, mask stimulus is to suppress the visibility of the first, target stimulus.

Crowding

Another masking phenomenon is produced by visual crowding. This is illustrated in the fourth panel of Figure 1. Here a target stimulus is presented, usually for a relatively long duration of several seconds, at some distance from central fixation. When the stimulus is presented in isolation, that is, without being crowded by nearby, adjacent stimuli, it is clearly recognizable. However, when surrounded by other visual stimuli it is not easily or at all recognizable. It is important to note, however, that its visibility as such is not suppressed. Although some thing is seen, its identity cannot be determined.

Attentional limits

It is known that the effectiveness of visual (two-transient) masking and crowding is modulated by attentional factors. The visibility or the recognizable identity of a target stimulus is more readily masked when attention is withdrawn from it. Inattentional blindness demonstrates that an otherwise easily detectable target fails to be detected when attention is diverted to another, simultaneous task. Several other procedures that rely on the withdrawal or diversion of attentional resources can be used to suppress the identification or the visibility of a stimulus. One exploits a phenomenon known as the attentional blink. Here, several stimuli such as alphanumeric characters are presented at the same display location one at time in a rapid sequence, each stimulus flashed on the order of 100 ms. Embedded within a sequence of numeric characters are two letters of the alphabet that serve as targets. The observer is required to report the identity of the two letter targets. What typically happens is that if the first letter is identified correctly, the second letter, when presented, say, 200 400 ms later, tends not to be identified. Presumably, the attention devoted to the processing of the first stimulus after it is presented is thereby effectively withdrawn for a couple of 100 ms from the processing of the second target. Another expression of attentional limits manifests in a phenomenon known as change blindness. For example, imagine viewing a series of two alternating depictions of a scene, such as an urban street scene, in which there is a noticeable difference between the two depictions. The change or difference can be a deletion of some significant object in, or some portion of, the scene, an addition to it, or both. As long as the scene is fairly complex, that is, contains many visual elements and features, attention cannot be focused continuously on all of its contents. Hence changes introduced into the scene, such as a series of substitutions or deletions and reinsertions of elements, will go undetected in a large proportion of observers.

Transcranial magnetic stimulation

Finally a recent technique for rendering stimuli invisible that is similar to the two-transient visual masking technique or to continuous flash suppression, is to apply a magnetic pulse or a longer lasting series of magnetic pulses to the cranium overlying a particular part of the visual cortex.

This technique, known a transcranial magnetic stimulation (TMS), causes a disruption of neural activity in cortical tissue located directly below the cranial point of application. For instance, when applied to the midoccipital region of the cranium at the back of the head, TMS can produce a blindness near the foveal part of the visual field by disrupting neural activity in that area of the underlying primary (or secondary) visual cortex responsible for processing information received from the near-foveal region of each eye s retina.

Findings and Their Theoretical Interpretations

It is generally agreed among neuroscientists that conscious perception of a stimulus requires that some informational content about it be processed at relatively high cortical levels. For instance, to perceive its shape, activity in the lateral occipital cortex, which supports object recognition, must be present; while to perceive its motion, activity in the medial temporal region of the visual cortex must be present. Both of these areas of visual cortex are separated from the primary visual cortex by one or more synaptic junctures. This does not mean that lower levels of visual processing are unimportant. As noted above, blindsight, the absence of visual object qualia in the phenomenal field, results from damage to the primary visual cortex, the first anatomic site of cortical visual processing. By extension, elimination of conscious vision also results from damage to sites, such as the lateral geniculate nucleus in the thalamus, anatomically located even lower in the visual system. As a consequence there are many loci in the human visual system where unconscious information processing occurs. As an obvious and trivial example, the retinas of the two eyes also are sites where unconscious processing occurs. However, there the unconscious neural activity as such (i.e., without activating higher levels of processing) has no effect on perception or behavior, and thus cannot be assessed. The trick then is to find levels of processing in the human visual system that are unconscious and that can affect perception or behavior.

Delimiting the Sites of Postretinal Unconscious Processing in the Human Visual System

It is noted above that there is no one and only way to render a stimulus or a part of it inaccessible to consciousness. Perhaps more than coincidentally, at postretinal levels of processing there is no single site of unconscious processing of stimuli. However, in some cases the relation between method of rendering stimuli invisible and the nature of unconscious processing may be specified. Thus, it is possible, at least in principle, that two (or more) different ways of rendering a stimulus or its attributes inaccessible to consciousness can be associated with two (or more) different types or levels of postretinal processing. The following, besides giving examples of such associations between method and finding, explores the different levels and types of unconscious visual processing.

Unconscious visual processing in neurological patients

Blindsight

Patients with damage to the primary visual cortex, also known as striate cortex, fail to spontaneously report the phenomenal presence of most visual stimuli in the defective part of the visual field. Yet, when studied carefully they exhibit two methodologically distinct types of visual function. In the first type, one visual stimulus presented in the blind area can prime the response to a second stimulus presented in the sighted area of the visual field. For instance, a colored stimulus presented in the blind field can prime the response to a second colored stimulus presented in the sighted part of the field. In the second type, significantly better-than-chance performance is obtained when stimuli are presented within the visual field defect and the patients are required to guess whether a stimulus was presented, where a stimulus was presented, or which of several stimuli was presented.

Although they lack experiences of qualia (except occasionally a coarse phenomenal sense of motion), when pressed to make discriminative responses these patients are capable of residual unconscious vision. Among indices of preserved visual functions are (1) discriminating the presence of a briefly flashed stimulus from its absence,

(2) localizing, via saccades or hand pointing, a stimulus presented at variable locations in the blind area of the visual field, (3) discriminating among speeds and directions of moving stimuli, (4) rudimentary wavelength discrimination, (5) the ability of an unseen stimulus to cue temporal and spatial shifts of attention, (6) coarse orientation and low-spatial frequency pattern discrimination, (7) appropriate responses to a variety of facial emotional expressions, and (8) report of phenomenal awareness of chromatic and figural properties of the afterimage of stimuli presented into the visual field defect. Exactly where and how in the visual system these residual abilities are expressed is not certain, although it appears that direct pathways from the lateral geniculate nucleus to cortical extrastriate cortex, thus circumventing the primary visual cortex, and the subcortical pathways including the superior colliculus and pulvinar are involved.

Form blindness

Some patients have damage at poststriate or extrastriate stages of cortical visual processing. In particular, a number of patients have damage to an area of extrastriate cortex called the lateral occipital complex. This area is crucial to the perception of the form or shape of objects. Although these patients fail to report phenomenal awareness of the form of these stimuli, the form can nevertheless control motor behavior. While they fail to report the orientation of a rectangular elongated slot or aperture, they can adjust their hand appropriately when asked to insert a card held in their hand into the slot. For instance, if the length of the slot has a vertical orientation, the patient will attempt to insert the card by holding it vertically; whereas, if the orientation of the slot is changed to, say, a left oblique one, the orientation of the card held in the hand will be adjusted accordingly. Such patients also can adjust the size of their hand s grip in anticipation of grabbing an object that fails to appear in phenomenal awareness. For instance, grabbing a pencil will result in a narrower grip than grabbing a dictionary. Many more such residual unconscious controls of behavior have been documented in such patients.

There are many other types of neurological patients too many to exhaustively describe who lack specific phenomenal qualia yet who have residual visual function. An example is the aforementioned achromatopsic patient who fails to report any phenomenal awareness of color, yet can, if prompted, discriminate among different wavelengths of light. Cases of blindsight, form blindness, achromatopsia, and so on, indicate why, as suggested above, the notion of access consciousness is problematic. These patients, although phenomenally blind to a stimulus attribute, have access to residual visual information about that stimulus dimension that can control their behavioral responses. Thus, by definition they have access consciousness of the stimuli. However, it remains puzzling as to the way they can be conscious of the accessed stimulus information without being phenomenally conscious of it. Most researchers in visual cognition fail to stretch the definition of consciousness beyond phenomenal awareness, and they regard access consciousness of the type described above as examples of unconscious processing.

Unconscious visual processing in normal observers

Effects of binocularly rivalry suppression

Despite reports of global or high-level perceptual-grouping factors modulating binocular rivalry suppression and amplification of rivalry suppression along the ascending cortical processing pathways, increasing evidence indicates that binocular rivalry is by and large eye-specific and is initiated powerfully at relatively low levels of cortical processing. This suggests that the visual information that survives the suppressive effects is processed unconsciously at low levels. Thus, although some information giving rise to the detection of apparent motion, to orientation-specific adaptation effects, and to the production of afterimages survives binocularly rivalry suppression, no semantic word priming, picture priming, or color priming can be produced by stimuli presented to an eye during its suppressed phase. On the other hand, responses in the emotional centers of the brain can be obtained when pictures of emotionally expressive faces are presented to the suppressed eye. In these respects binocular rivalry has effects similar to blindsight, as would be expected since blindsight also is caused by disruption of neural activity at low cortical levels of processing.

Effects of motion-induced blindness

The unconsciously processed information of stimuli that are rendered invisible during motion-induced blindness can produce several perceptual effects. For one, the visually suppressed stimuli are still able to produce visible afterimages. Moreover, when changes are introduced to invisible stimuli during motion-induced blindness, these unseen changes can nevertheless exploit perceptual Gestalt grouping factors which update subsequently seen object representations by influencing not only how stimuli reenter into phenomenal awareness but also what objects reenter. Regarding the former effect of afterimage formation, the residual effects of unseen stimuli during motion-induced blindness are similar to the effects of unseen stimuli in blindsight.

Effects of continuous flash suppression

Continuous flash suppression can yield effects distinct from standard binocular-rivalry suppression. For instance, whereas standard methods producing alternating periods of binocular-rivalry dominance and suppression had little or no effect on a suppressed stimulus s ability to produce afterimages, continuous flash suppression, perhaps by deepening and prolonging the effects of binocular-rivalry suppression, does measurably affect the suppressed stimulus s ability to produce perceivable afterimages.

Effects of visual masking

A stimulus, called the prime, whose visibility is suppressed by a subsequent stimulus, called the mask (backward masking), can, despite its being not seen, affect the perceptual response to a subsequently presented probe stimulus. In particular, a masked, unconsciously processed prime can affect the response to a probe s color, form, and meaning. These results are interesting when considered in relation to the unconscious processing during binocular-rivalry suppression, since a stimulus that is suppressed during binocular rivalry fails to yield priming by its shape, color, or meaning. This suggests that the suppressive effects of masking are realized at a level of processing that is subsequent to, or higher than, than the level of processing at which the suppressive effects binocular-rivalry suppression occur. This is consistent with the finding that the mechanism responsible for the suppressive effects of backward masking can in turn be suppressed during binocular rivalry.

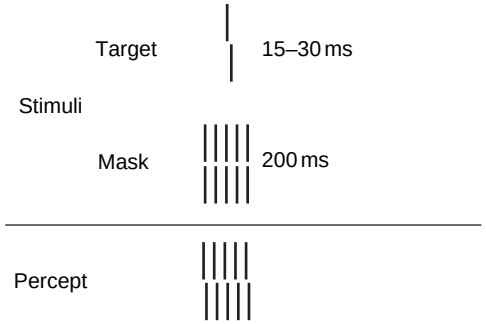


Figure 2 A schematic of typical stimuli and their temporal sequence used in studies of feature inheritance.

An interesting phenomenon that can accompany backward masking is what is known as feature inheritance. Here, as shown in Figure 2, the visibility of a target stimulus composed, say, of two misaligned vertical elements is suppressed by an array or grating composed of aligned vertical elements presented about 30 ms later. However, although the grating contains aligned vertical elements, it is perceived as if the elements in it had been misaligned in accordance with the actual misalignment of the suppressed target. Here, the unconsciously processed misalignment feature of the target is perceptually inherited by the aligned mask grating. It is possible that such feature inheritance may contribute to the priming effects produced by a masked prime.

Effects produced by TMS

Magnetic stimulation applied to the occipital pole locally disrupts neural activity in the underlying primary visual cortex. In this respect, it can be regarded as transiently inducing a state similar to blindsight. In fact, its effects are also similar in that the wavelength and orientation of a stimulus whose visibility is suppressed can still be discriminated. Moreover and also similar to blindsight, stimuli whose visibility is suppressed by TMS can nonetheless direct saccadic eye movements to their unseen location in the visual field. It is possible that other effects of unconsciously processed information obtained in blindsight patients will also be obtained with suppression produced by TMS.

Effects of attentional limits

As noted above, limits on attentional resources can be experimentally imposed in a variety of ways. It has been shown that stimuli that fail to access conscious report during the attentional blink can nevertheless prime the perceptual processing of subsequently presented probe stimuli. Priming effects can also be produced during change blindness by both the prechange stimuli and the postchange stimuli that are unavailable to conscious report. Moreover, blindness to the prechange stimuli is associated with the limited duration of iconic memory and thus, as noted above, may reflect a failure of access consciousness.

The Role of Attention in Unconscious Visual Processing

It was noted above that blindsight patients can direct attention to the location of a stimulus presented in the visual field defect. It is also becoming clear that attention can modulate the effects produced by unconsciously processed stimuli. Such attentional modulation has been found in the unconscious processing of stimuli whose visibility is suppressed during backward masking and during visual crowding. Nonetheless some visual stimuli, such as natural scenes or faces and their emotional expressions, whose visibility is suppressed by an aftercoming mask seem to require little if any attention in order to be processed unconsciously. Natural scenes and faces occur very often in our normal everyday commerce with the visual world. In addition faces conveying emotional expressions are particularly salient social stimuli. For any of these reasons such stimuli might be processed automatically with little or no need of attentional modulation. On the other hand, the unconscious processing of abstract and nonecological stimuli typically used in laboratory experiments may be subject to attentional modulation since they are both less familiar and less salient.

The Role of Feedforward and Reentrant Feedback Connections

In current approaches to visual cognition and neuroscience, prominence is given to the distinction between feedforward and reentrant or feedback connections in cortical neural networks. According to some theoretical approaches, the multisynaptic feedforward transmission and transformation of visual information along successive levels of cortical processing is not sufficient to give rise to a conscious percept. Also deemed necessary are the feedback connections from higher levels of processing to lower ones. Accordingly, much or most of the unconscious visual information processing can theoretically be accomplished during the feedforward sweep of cortical neural activity. This can give rise to, for example, the unconscious priming of the processing of stimulus attributes such as form, color, and even semantic content. Suppression of the feedforward flow of information at low levels, such as the suppression produced during binocular rivalry, would therefore leave less information processing intact than suppression produced at higher levels of the feedforward flow, such as the suppression produced by visual backward masking. Moreover, if the feedforward sweep of cortical processing is indeed largely responsible for unconscious processing, it follows that neural activity at higher levels of cortical processing will code more complex, relational stimulus properties than do the lower levels. In addition, some unconsciously processed information may proceed along an alternate, subcortical feedforward route, since there are direct connections from the retina to the superior colliculus of the midbrain and hence, via the pulvinar, to cortical extrastriate regions and to the emotional centers in the limbic system including the amygdala. Such feedforward pathways could be responsible not only for the residual motion perception in blindsight patients but also the activation of emotional centers by stimuli rendered invisible either in blindsight or by psychophysical methods.

Conclusions

Much evidence has accumulated over the past three decades indicating the existence of unconscious information processing in the human visual system. The evidence, however, must be weighed in light of important methodological and conceptual distinctions. The most important methodological issues concern the adoption of appropriate

criterion contents and of adequate rationales and procedures used to assure the suppression of the conscious registration of visual stimuli. Different neurological conditions and different experimental methods tap into different types and levels of unconscious information processing, and one of the tasks of future research is to work out the relationship among these different conditions and methods and the correspondingly different types and levels of unconscious processing.

See also: Attention: Change Blindness and Inattentional Blindness; Attention: Selective Attention and Consciousness; Concepts and Definitions of Consciousness; Implicit Learning and Implicit Memory; Perception, Action, and Consciousness; Perception: Subliminal and Implicit; Phenomenology of Consciousness; Unconscious Cognition.

Suggested Readings

Alais D and Blake R (2005) Binocular Rivalry. Cambridge, MA: MIT Press.

Breitmeyer BG and Ögmen H (2006) Visual Masking: Time Slices through Conscious and Unconscious Vision. Oxford: Oxford University Press.

Breitmeyer BG, Ro T, and Ögmen H (2004) A comparison of masking by visual and transcranial magnetic stimulation: Implications for the study of conscious and unconscious visual processing. Conscious and Cognition 13: 829–843.

Holender D (1986) Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. Behavioral and Brain Sciences 9: 1–66.

Kihlstrom JF (1996) Perception without awareness of what is perceived, learning without awareness of what is learned. In: Velman V (ed.) The Science of Consciousness, pp. 23–46. London: Routledge.

Kim C-Y and Blake R (2005) Psychophysical magic: Rendering the visible 'invisible'. Trends in Cognitive Neuroscience 9: 381–388.

Koch C (2004) The Quest for Consciousness. Englewood, CO: Roberts & Co.

Kouider S and Dehaene S (2007) Levels of processing during non-conscious perception: A critical review of visual masking. Philosophical Transactions of the Royal Society B 362: 857–875.

Mack A and Rock I (1998) Inattentional Blindness. Cambridge, MA: MIT Press.

Merikle PM and Daneman M (1998) Psychological investigations of unconscious perception. Journal of Consciousness Studies 5: 5–18.

Neumann O and Klotz W (1994) Motor responses to nonreportable, masked stimuli. Where is the limit of directparameter specification. In: Umilta C and Moscovitch M (eds.) Attention and Performance XV, pp. 123–150. Cambridge, MA: MIT Press.

Ögmen H and Breitmeyer BG (2005) The First Half Second: The Microgenesis and Temporal Dynamics of Unconscious and Conscious Visual Processes. Cambridge, MA: MIT Press.

Simons DJ and Rensink RA (2005) Change blindness: Past, present, and future. Trends in Cognitive Sciences 9: 16–20.

Stoerig P (1996) Varieties of vision: From blind responses to conscious recognition. Trends in Cognitive Neurosciences 19: 401–406.

Weiskrantz L (1997) Consciousness Lost and Found: A Neuropsychological exploration. Oxford: Oxford University Press.

Biographical Sketch

Bruno G Breitmeyer is a professor of psychology at the University of Houston. He studied mathematics at the University of Illinois, Champaign-Urbana, and received his PhD in psychology from Stanford University. He did his postdoctoral research at the Bell Laboratories, Murray Hill, NJ and, as an Alexander von Humboldt fellow, at the Neurological Clinic of Freiburg University, Germany. His research interests are in visual cognition and neuroscience. Primary research activities include investigations of the temporal dynamics of visual masking, object perception, and attention. In recent years, his research, in collaboration with his University of Houston colleague, Haluk Öğmen, has focused on the use a variety of masking techniques to study the types and levels of conscious and unconscious visual information processing. He has published numerous scientific journal articles and book chapters and coauthored three books dealing with these topics. He has also organized major workshops, served on various journal editorial boards and has received several awards, among them the Science Citation Award from the Institute of Scientific Information, the Alexander von Humboldt Fellowship, and a Distinguished Alumnus Award from the University of Illinois.

Phenomenology of Consciousness

D Zahavi, University of Copenhagen, Copenhagen, Denmark

Glossary

Externalism – The view that what is thought and experienced is essentially dependent on aspects of the world external to the mind of the subject.
Intentionality – The directedness or of-ness or aboutness of consciousness; when one perceives or judges or feels or thinks, one's experience is about or of something.
Internalism – The view that a subject's beliefs and experiences are wholly constituted by what goes on inside the mind of that subject.
Intersubjectivity – The relation between experiencing subjects.
Lived body – The body understood as an embodied first-person perspective.
Prereflective self-consciousness – The ongoing first-personal manifestation of experiential life.
Transcendental – A philosophical reflection on the conditions of possibility of experience and cognition.

Introduction

Until recently, many scientists considered the study of consciousness to be inherently unreliable due to its subjective nature and thus unsuitable for scientific research. Since the early 1990s, however, a marked change has taken place, one occasionally described in terms of an ongoing consciousness boom. Many new journals devoted to the study of consciousness have been established and currently questions pertaining to the nature of phenomenal consciousness, the structure of the first-person perspective, and the status of the self are once again considered crucial for a scientific understanding of consciousness.

Given the recent interest in the subjective dimension of consciousness, it is no wonder that many have started to emphasize the importance of phenomenology. Thus, when studying consciousness rather than, say, deep-sea ecology, we should provide a phenomenologically sensitive account of the various structures of experience, since an important and nonnegligible feature of consciousness is the way in which it is experienced by the subject.

The aim of the following contribution is to outline and discuss some central features of a phenomenological account of consciousness. In this context, however, phenomenology will not refer to a form of psychological self-observation, but rather to a specific philosophical tradition a tradition inaugurated by Husserl (1859 1938), and comprising among its most well-known champions, philosophers like Max Scheler, Martin Heidegger, Aaron Gurwitsch, Maurice Merleau-Ponty, and Jean-Paul Sartre.

Phenomenology counts as one of the important philosophical traditions in the twentieth century, and has made significant contributions to most areas of philosophy, including philosophy of mind, social philosophy, philosophical anthropology, aesthetics, ethics, philosophy of science, epistemology, theory of meaning, and formal ontology. Within the study of consciousness, phenomenology has in particular provided analyses of such topics as intentionality, perception, embodiment, self-consciousness, sociality, and temporality.

Like any other tradition, phenomenology spans many differences, and has developed as a heterogeneous movement with many branches. For the same reason, it will be impossible to do justice to all aspects of phenomenology in a short overview like the present. Rather, the focus will be on a few selected topics; topics deemed to be of particular importance for contemporary discussions in philosophy of mind and cognitive science. Furthermore, rather than spending time articulating the

differences between the various figures in phenomenology differences that have frequently been overstated the emphasis will be on some of the overarching concerns and common themes that have united its proponents.

Methodological Considerations

Edmund Husserl s maxim for phenomenology was Back to the things themselves! By this he meant that phenomenology should base its considerations on the way things are experienced rather than by various extraneous concerns which might simply obscure and distort that which is to be understood. In short, phenomenology should not start with theory; it should be critical and nondogmatic, shunning metaphysical and theoretical prejudices, as much as possible, and seeks to be guided by that which is actually experienced, rather than by what we expect to find given our theoretical commitments. It asks us not to let preconceived theories determine our experience, but to let our experience inform and guide our theories. To the extent that phenomenology stays with experience it is said to take a first-person approach. That is, the phenomenologist is concerned to understand the experience in terms of the meaning it has for the subject. It does not seek to disclose the subpersonal mechanisms that might enable us to experience the way we do. To put it differently, phenomenology is concerned with attaining an understanding and proper description of the experiential structure of our experiential life; it does not attempt to develop a naturalistic explanation of consciousness, nor does it seek to uncover its biological genesis, neurological basis, psychological motivation, or the like.

To fully appreciate the philosophical import of the phenomenological account of consciousness, it is, however, important to understand that the goal of phenomenology is not to describe each and every idiosyncratic experience here and now, this is what I experience rather, it aims to capture invariant and universal structures of experience; structures that are intersubjectively valid and accessible. For the very same reason, its analyses are open for corrections and control by any (phenomenologically tuned) subject. Furthermore, it is important to realize that phenomenology notwithstanding various other differences to said tradition is in some respects firmly situated within a certain Kantian or post-Kantian framework. One way to interpret Kant s revolutionary Copernican turn in epistemology is by seeing it as amounting to the claim that our cognitive apprehension of reality is more than a mere mirroring of a preexisting world, for which reason a philosophical analysis of reality, a reflection on what conditions something must satisfy in order to count as real, cannot ignore the contribution of consciousness. Thus, and this pinpoints a main difference to at least a good part of recent analytic philosophy s preoccupation with consciousness, the phenomenological interest in the first-person perspective is not primarily motivated by the relatively trivial insight that we need to include the first-person perspective if we wish to understand mental phenomena. Rather, the phenomenologists focus on the first-person perspective is as much motivated by an attempt to understand the nature of objectivity, as by an interest in the subjectivity of consciousness. Indeed, rather than taking the objective world as the point of departure, phenomenology asks how something like objectivity is possible in the first place. What are the primitive modes of understanding that precede our scientific understanding? How is objectivity constituted? In phenomenological texts the term constitution is a technical one. Constitution must be understood as a process that allows for the manifestation or appearance of objects and their signification, that is, it is a process that permits that which is constituted to appear, to manifest and present itself as what it is. And this process is something that in significant ways involves the contribution of consciousness. Objects are constituted, that is, experienced and disclosed in the ways they are thanks to the way consciousness is structured. Phenomenologists consequently reject the suggestion that consciousness is merely one object among others in the world, on par with though possibly more complex than volcanoes, waterfalls, ice crystals, gold nuggets, rhododendrons, or black holes, since they would consider it to be a necessary (though not sufficient) condition of possibility for any entity to appear as an object in the way it does and with the meaning it has.

Husserl occasionally speaks of the phenomenological attitude as consisting in a reflective move that takes a step back from our naïve and unexamined immersion in the world in order to focus on the way in which the world manifests itself to us. In other words, once we adopt the phenomenological attitude, we are no longer primarily interested in what things are in their weight, size, chemical composition, etc. but rather in how they appear, and thus as correlates of our experience. Indeed to put it slightly paradoxically, phenomenologists are not interested in consciousness *per se*. They are interested in consciousness because they consider consciousness to be our only access to the world. They are interested in consciousness because it is world-disclosing.

Contrary to certain misunderstandings, the phenomenological reflection does not amount to an inward turn, as if its aim is to make us ignore the world in favor of consciousness. On the contrary, it permits us to investigate the world we live in from a new reflective attitude, namely in its significance and manifestation for consciousness. Although this reflective investigation differs from a straightforward exploration of the world, it remains an investigation of reality; it is not an investigation of some otherworldly, mental realm. If we wish to describe the experiential difference between tasting wine and tasting water, between perceiving and imagining a rhinoceros, we do not sever our intentional link with the world. Rather, we discover these differences, and we analyze them descriptively by paying attention to how worldly objects and states of affairs appear to us.

When we perceive, judge, or evaluate objects, a thorough phenomenological examination will lead us to the experiential structures and modes of understanding to which these types of appearance are correlated. We are led to the acts of presentation the perception, judgment, or valuation and thereby to the experiencing subject (or subjects) in relation to whom the object as appearing must necessarily be understood. By adopting the phenomenological attitude we pay attention to how public objects (trees, planets, paintings, symphonies, numbers, states of affairs, social relations, etc.) appear. But we do not simply focus on the objects precisely as they appear; we also focus on the subjective side of consciousness, thereby becoming aware of our subjective accomplishments and of the intentionality that is at play. In short, phenomenology must be understood as a philosophical analysis of the different types of world-disclosure (perceptual, imaginative, recollective, etc.), and in connection with this as a reflective investigation of those structures of experience and understanding that permit different types of beings to show themselves as what they are.

Consciousness and Self-Consciousness

When it comes to the relation between consciousness and self-consciousness, literally all of the major figures in phenomenology defend the view that a minimal form of self-consciousness is a constant structural feature of conscious experience. Experience happens for the experiencing subject in an immediate way and as part of this immediacy, it is implicitly marked as my experience. For the phenomenologists, this immediate and first-personal givenness of experiential phenomena must be accounted for in terms of a prereflective self-consciousness.

Self-consciousness on this account is not merely something that comes about the moment one scrutinizes one's experiences attentively (let alone something that only comes about the moment one recognizes one's own mirror image, refers to oneself using the first-person pronoun, or is in possession of a theory of mind, or an identifying knowledge of one's own life story). Rather, self-consciousness comes in many forms and degrees. It makes perfect sense to speak of self-consciousness whenever I consciously perceive an external object an oak, a table, or the moon because to consciously perceive something is not simply to be conscious of the perceptual object, but also to be acquainted with the perception of the object. The difference between a nonconscious perception of a sunset, and a conscious perception of the sunset, is that there is something it is like to consciously perceive a sunset. Although my attention is on the object, the experience itself remains conscious. Not in the sense that I am thematically aware of it, but in the sense that there is something it is like to be in that state, it has phenomenal

qualities. The notion of prereflective self-consciousness is consequently related to the idea that experiences have a subjective feel to them, a certain (phenomenal) quality of what it is like or what it feels like to have them. As it is usually expressed in analytical philosophy of mind, to undergo a conscious experience necessarily means that there is something it is like for the subject to have that experience. One reason experiences are said to be subjective is because they are characterized by a subjective mode of existence in the sense that they necessarily feel like something for somebody. Our experiential life can consequently be said to entail a primitive form of for-me-ness. This for-me-ness is not a quality like yellow, salty, or spongy. It does not refer to a specific content of experience, to a specific what, but to the unique mode of givenness or how of experience. It refers to the first-personal givenness of experience. In its most primitive and fundamental form, self-consciousness is a question of the ongoing first-personal manifestation of experiential life.

This claim should not be misunderstood. The phenomenologists are not advocating a strong thesis concerning total and infallible self-knowledge; rather they are calling attention to the constitutive link between experiential phenomena and the first-person perspective. When speaking of a first-person perspective it is important to be clear about the distinction between having or embodying such a perspective and being able to articulate it linguistically. Whereas the latter presupposes mastery of the first-person pronoun the former is simply a question of the subjective manifestation of one's own experiential life. Although both capabilities deserve to be investigated, phenomenologists have mainly been accentuating the significance of the former. To emphasize the importance of the first-person perspective is consequently simply to stress that there is a distinctive way experiential episodes present themselves to the subject whose episodes they are.

Phenomenologists explicitly deny that the self-consciousness that is present the moment I consciously experience something is to be understood in terms of some kind of reflection, or introspection, or higher-order monitoring. That is, the self-consciousness in question is not a new consciousness. It is not something added to the experience, an additional mental state, but rather an intrinsic feature of the primary experience. Thus, when phenomenologists speak of self-consciousness as a permanent feature of consciousness, they are not referring to what is usually called reflective self-consciousness. Reflection (or higher-order monitoring) is the process whereby consciousness directs its intentional aim at itself, thereby taking itself as its own object. But on the phenomenological account such a form of objectifying self-consciousness is derived. It presupposes the presence of the more fundamental prereflective self-consciousness. The latter form of self-consciousness is not thematic or attentive or voluntarily brought about; rather it is tacit, and very importantly, thoroughly nonobservational (i.e., it is not a kind of introspective observation of myself) and nonobjectifying (i.e., it does not turn my experience into a perceived or observed object). I can, of course, reflect on and attend to my experience, I can make it the theme or object of my attention, but prior to reflecting on it, I was not mind- or self-blind. The experience was already something for me, and in that sense it counts as being prereflectively self-conscious.

It is because of their conviction that consciousness is as such characterized by a primitive, tacit, self-consciousness, that many phenomenologists have also found it appropriate to ascribe a primitive sense of self to the experiential phenomena. The basic idea is that an understanding of what it means to be a self calls for an examination of the structure of experience, and vice versa. Thus, on this minimal definition the self is not something that stands opposed to the stream of consciousness, it is not an ineffable hypothesized precondition, nor is it a mere social construct that evolves through time; it is taken to be an integral part of the structure of our conscious life. More precisely, they argue that the self possesses experiential reality, and link it to the first-personal givenness of the experiential phenomena. In short, the self is conceived as the ubiquitous dimension of first-personal givenness in the multitude of changing experiences.

Incidentally, this view makes it clear that self-experience is not to be understood as an experience of an isolated, worldless self; nor is the self located and hidden in the head. To have a self-experience is not to interrupt the experiential

interaction with the world in order to turn one s gaze inward; on the contrary, at its most primitive self-experience is simply a question of being prereflectively aware of one s own intentional consciousness. It is as we shall see in a moment always the self-experience of an embodied and embedded mind. It would consequently be a mistake to interpret the phenomenological notion of a minimal self as a Cartesian-style mental residuum, that is, as some kind of self-enclosed and self-sufficient interiority. The phenomenological notion of self is fully compatible with a strong emphasis on the fundamental intentionality, or being-in-the-world, of consciousness.

Intentionality

The concept of intentionality has a long history that stretches back at least as far as Aristotle. It played a central role in medieval scholastic epistemology, but the modern revival of the term intentionality is due to Brentano. Intentionality has to do with the directedness or of-ness or about-ness of consciousness, that is, with the fact that when one perceives or judges or feels or thinks, one s experience is about or of something. The phenomenologists have primarily been interested in intentionality as a decisive feature of consciousness. Moreover, they have specifically focused on an account of intentionality from the first-person perspective, that is, from the subject s point of view. In fact, none of the phenomenologists have been engaged in a naturalization of intentionality, if that is understood as an attempt to explain intentionality reductively by an appeal to nonintentional mechanisms and processes. If one thinks that a theory of intentionality must issue in a reductive account one will be bound to find the phenomenological treatment of intentionality disappointing.

What is the aim of the phenomenological account of intentionality then? First and foremost, to provide a descriptive analysis of intentional the structures of consciousness. In doing so, however, the phenomenologists also seek to clarify the relation between mind and world (rather than the relation between mind and brain). This latter investigation, which basically intends to demonstrate the world-involving character of the mind and rejects the view that consciousness is a subjective sphere that exists independently from the world that is revealed through it, has some more overarching philosophical implications.

Husserl s Logical Investigations contains the first proper phenomenological investigation of intentionality. Like Brentano, Husserl argues that one does not merely love, fear, see, or judge, one loves a beloved, fears something fearful, sees an object, and judges a state of affairs. Regardless of whether we are talking of a perception, thought, judgment, fantasy, doubt, expectation, or recollection, all of these diverse forms of consciousness are characterized by intending objects, and cannot be analyzed properly without a look at their objective correlate, that is, the perceived, doubted, expected object. The converse is also true: The intentional object cannot be analyzed properly without a look at its subjective correlate, the intentional act. Neither the intentional object nor the experience that intends it can be understood apart from the other. Acts of consciousness and objects of consciousness are essentially interdependent: the relation between them is an internal rather than an external one. That is to say, from the perspective of phenomenology one cannot first identify the items related and then explore the relation between them. Rather one can identify each item in the relation only by reference to the other item to which it is related.

It is customary to speak of intentional relations as being aspectual or perspectival. One is not simply conscious of an object, one is always conscious of an object in a particular way. One always has a certain perspective or point of view on the object; the object is always presented in a certain way or under a certain aspect for the subject. More specifically, however, we need to distinguish the intentional object in the how of its determinations and in the how of its givenness. To take a simple example, let us consider a perception of an iPod. I can see the iPod from one perspective or another, I never see it in its totality all at once. Furthermore, the iPod always appears to me in a certain illumination and with a certain background. Moreover, it also appears in a certain context with a determinate meaning. Depending on my previous experiences and current interests, I might see it as a present I got from a good friend, as an efficient

way to storage my music collection, as a headache and source of irritation (because it is not working properly), etc. But apart from intending different properties of the object, apart from varying what the object I am intending is presented as, I might also vary the very form of presentation itself. Instead of perceiving the iPod, I can also imagine it, judge about it, remember it, etc.

Husserl, for instance, typically distinguishes between signitive, imaginative (pictorial), and perceptual ways of intending an object or state of affairs: I can talk about a withering oak which I have never seen, but which I have heard is standing in the backyard, I can see a detailed drawing of the oak; or I can perceive the oak myself. Similarly, I can talk about how terrible it must be for homeless people to sleep on the streets, I can see a television program about it; or I can experience it myself. For Husserl these different ways of intending are not unrelated. On the contrary, the modes can be ranked according to their ability to give us the object as directly, originally, and optimally as possible. The object can be experienced more or less directly, that is, it can be more or less present. The lowest and most empty way in which the object can be intended is in the signitive act. These (linguistic) acts certainly have a reference, but apart from that, the object is not given in any fleshed out manner. The imaginative (pictorial) acts have a certain intuitive content, but like the signitive acts, they intend the object indirectly. Whereas signitive acts intend the object via a contingent representation (a linguistic sign), pictorial acts intend the object via a representation (picture) which bears a certain resemblance to the object as seen from a certain perspective. It is only the actual perception, however, which gives us the object directly. This is the only type of intention which presents us with the object itself in its bodily presence. Thus, on the Husserlian account, perception does not confront us with pictures or images of objects — except, of course, in so far as we are perceiving paintings or photographs — but with the objects themselves. When we say that something appears perceptually, this should consequently not be understood in the sense that the perceptually given is a picture or sign of something else.

For Husserl, intentionality is not an ordinary relation to an extraordinary object, but a special kind of relation to the intended object; a special relation that can hold, even if the object does not exist; and that can persist even if the object ceases to exist. When it comes to intentions that are directed toward unreal objects, they are in his view just as much characterized by their directedness as are ordinary perceptions. In contrast to normal perceptions, however, the referent — that which would satisfy or fulfill the intention — does not exist, neither intramentally, nor extramentally. In the case of a hallucination, the pink elephant exists neither inside nor outside of consciousness, but the hallucination is still about a pink elephant. In short, although one of the peculiar features of the mind is its ability to think about objects that do not exist, we should not accept the reality of nonexistent objects. The intentional object is not a special kind of object, but rather the answer to the question of what a certain intentional state is about. If the answer refers to some nonexisting object, the intentional object does not exist. If the answer refers to some existent thing then the intentional object is that real thing. So if I look at my fountain pen, then it is this real pen which is my intentional object, and not some mental picture, copy, or representation of the pen.

Throughout this kind of analysis, phenomenology contends that consciousness is characterized by an intrinsic intentionality, and resists the attempt to provide a reductive account of intentionality, for example, by trying to explain it by appeal to nonintentional factors such as causality. But how exactly does intentionality work? How do we intend objects? This is where the notion of meaning becomes central. For the phenomenologists, intentionality is a question of meaning. We intend an object by meaning something about it.

Does this emphasis on meaning commit the phenomenological account of intentionality to some form of internalism? In this context, internalism refers to the view that a subject s beliefs and experiences are wholly constituted by what goes on inside the mind of that subject, so that matters in the subject s natural and cultural environment have no bearing on their content. By contrast, externalism argues that mental states are externally individuated. What we think about, what we refer to, depends upon what actually exists in the (physical, social, and cultural) environment; what we experience depends upon factors that

are external to the subject possessing the experience in question.

There has been a widespread tendency to argue that whereas Husserl was a classical Cartesian internalist, existential phenomenologists like Heidegger and Merleau-Ponty favored a form of externalism since they were fully committed to the view that the mind is essentially determined by its intentional relationship to the world. Although it is quite true that later phenomenologists to a somewhat larger extent than Husserl emphasized the importance of practical and bodily forms of intentionality this standard interpretation nevertheless remains too simplistic. It ignores plenty of evidence suggesting that Husserlian phenomenology differs from traditional internalism. At the same time, however, it is by no means obvious that Heidegger or Merleau-Ponty can be classified as straightforward externalists.

A frequent claim found in the phenomenological literature is that the relation between mind and world is an internal relation, a relation constitutive of its relata, and not an external one of causality. Mind and world are not distinct entities; rather they are bound constitutively together. Considering the way in which phenomenologists conceive of the mind world relationship, considering how they would reject the dualism between a self-contained mind and a mindless world, it is questionable whether it really makes much sense to classify their views as being committed to either internalism or externalism.

Ultimately, we should appreciate that the phenomenological investigation of the condition of possibility for manifestation is more fundamental than any division between inner and outer a division that the alternative between internalism and externalism remains bound to. Phenomenology might teach us that the forced choice between internalism and externalism is misguided, since there are other options available.

Embodiment

The phenomenological investigation of the body is not the analysis of one object among others. That is, it is not as if phenomenology in its investigation of a number of different ontological regions (the domain of logic, mathematical entities, utensils, works of art, etc.) also stumbles upon the body and then subjects it to a close scrutiny. On the contrary, the body is considered a constitutive or transcendental principle, precisely because it is involved in the very possibility of experience. It is deeply implicated in our relation to the world, in our relation to others, and in our self-relation.

Phenomenologists object to the metaphysical division between *res extensa* and *res cogitans*. If one accepted such a division the only place for the body would seem to be on the side of the *res extensa*. But phenomenologists deny that the body is a mere object in the world. The body is not merely an object of experience that we see, touch, smell, etc. Rather the body is also a principle of experience, it is that which permits us to see, touch, and smell, etc. Obviously, the body can also explore itself. It can take itself (or the body of another) as its object of exploration. This is what typically happens in physiology or neurology, etc. But such an investigation of the body as an object is not exhaustive.

The phenomenological emphasis on the body entails a rejection of Cartesian mind body dualism. But this does not entail an endorsement of some kind of Cartesian materialism. It is not as if the phenomenological way to overcome dualism is by retaining the distinction between mind and body, and then simply getting rid of the mind. Rather the notion of embodiment, the notion of an embodied mind or a minded body, is meant to replace the ordinary notions of mind and body, both of which are considered derivations and abstractions. Merleau-Ponty famously spoke of the ambiguous nature of the body, and argued that bodily existence is a third category beyond the merely physiological and the merely psychological.

The first and most basic phenomenological distinction to be made is between the objective body and the lived body (Husserl s distinction between Kørper and Leib, respectively). This is, a phenomenological distinction rather than an ontological one. It is not meant to imply that each of us has two bodies: one objective and one lived. Rather it is meant to explicate two different ways that we can experience and understand the body. Whereas the former notion focuses on the body as seen from an observer s point of view, where the observer may be a scientist, a physician, or even the embodied

subject herself, the latter notion captures the body understood as an embodied first-person perspective. A description of the lived body is a description of the body from the phenomenological perspective. On the one hand, it is the way the body appears in immediate experience. On the other hand, it is more than that it is the way the body structures our experience. The body is not a screen between me and the world; rather, it shapes our primary way of being-in-the-world. This is also why we cannot first explore the body by itself and then subsequently examine it in its relation to the world. On the contrary, the body is already in-the-world, and the world is given to us as bodily revealed. Indeed as Sartre points out, the body is operative in every perception and in every action. It constitutes our point of view and our point of departure.

As perceivers and agents we are embedded and embodied agents. Already Husserl stressed that perception rather than being a mere passive intake of information, involves activity, more specifically, bodily movement. We see with mobile eyes set in a head that can turn and is attached to a body that can move from place to place; a stationary point of view is only the limiting case of a mobile point of view. What we see and hear and touch and taste and smell is shaped by what we do, and what we are capable of doing. In ordinary experience perception and movement are always united. I touch something by moving the arm. I see something by moving the head and eyes. That which is perceived is perceived as nearby or further way, as something that can be approached and explored. Perceptual intentionality presupposes a moving and therefore embodied subject. To understand perception is to understand the intentionality of our own body.

When I perceive an object, say, a bookcase, the object is never given in its totality but always incompletely, in a certain restricted profile or adumbration. It is never the entire bookcase, including its front, backside, underside, and inside which is given intuitively, not even in the most perfect perception. Despite this, it seems right to say that the object of my perception is the bookcase, and not simply the perspectivally given surface of the front and side of the bookcase. We so to speak see more than we sense. How is this possible? Phenomenologists have suggested that the absent profiles of the object are given to us as something that is perceptually accessible. Whereas the actual given front of the bookcase is correlated with a particular bodily position, the horizon of the cointended but momentarily absent profiles of the bookcase (its backside, bottom, etc.) is correlated with my sensorimotor capacities. The absent profiles are linked to an intentional if then connection. If I move in this or that way, then this or that profile will become visually or tactually present.

All perception and action involves a component of bodily self-experience. I am sitting in a restaurant. I wish to begin to eat, and so I need to pick up my fork. But how can I do that? In order to pick up the fork, I need to know its position in relation to myself. That is, my perception of the fork must contain some information about me, otherwise I would not be able to act on it. On the dinner table, the perceived fork is to the left of me, the perceived knife is to the right of me, and the perceived plate and wineglass in front of me. Every perspectival appearance implies that the embodied perceiver is herself the experiential zero point, the indexical here in relation to which every appearing object is oriented. As an experiencing, embodied subject I am the point of reference in relation to which all of my perceptual objects are uniquely related. I am the center around which and in relation to which (egocentric) space unfolds itself, or as Merleau-Ponty would put it, when I perceive the world, the body is simultaneously revealed as the unperceived term in the center of the world toward which all objects turn their face. Whereas I can approach or move away from any object in the world, the body itself is always here as my very perspective on the world. That is, rather than being simply another perspectivally given object, the body itself is precisely that which allows me to perceive objects perspectivally. In a primary sense, I am not conscious of my body as an intentional object. I do not perceive it; I am it.

The body tends to efface itself on its way to its intentional goal. We do not normally monitor our movements in an explicitly conscious manner. Fortunately, for had we been aware of our bodily movements in the same way in which we are attentively aware of objects, our body would have made so high demands on our consciousness that it would have interfered with our daily life. When

I play tennis, my movements are not given as intentional objects. My limbs do not compete with the ball for my attention. Had that been the case, I would have been so inhibited that I would have been unable to play efficiently. However, this might change if something goes wrong. Imagine that you are playing tennis. Your attention is directed at the ball, which is heading toward you with high speed, as well as on the position of your opponent. Your body tightens in order to return the ball in a masterful smash, but suddenly you feel a sharp and intense pain in your arm. Your smashing opportunity is lost, and the pain is now demanding all your attention. It attracts your attention whether you want it to or not. There is nothing that reminds us of our embodiment (our vulnerability and mortality) as much as pain. Moreover, the painful body can occasionally be experienced as alien, as something that is beyond our control. As is frequently the case in life, it is the privation that teaches us to appreciate what we take for granted.

Nothing in this conception of embodiment should lead us to conceive of the body as something static, as if it has a fixed set of skills and abilities. Not only can the body expand its sensorimotor repertoire by acquiring new skills and habits, it can even extend its capacities by incorporating artificial organs and parts of its environment. The classical example is the blind man s cane. To put it differently and perhaps even more strikingly, the lived body extends beyond the limits of the biological body. It does not stop at the skin.

To take embodiment seriously is to contest a Cartesian view of the mind in more than one way. Embodiment entails birth and death. To be born is not to be one s own foundation, but to be situated in both nature and culture. It is to possess a physiology that one did not choose. It is to find oneself in a historical and sociological context that one did not establish. Ultimately, the issues of birth and death enlarge the scope of the investigation. They call attention to the role of historicity, generativity, and sexuality. Indeed, rather than being simply a biological given, embodiment is also a category of sociocultural analysis. To gain a more comprehensive understanding of the embodied mind, one cannot just focus narrowly on perception and action, one also has to consider sociality.

Intersubjectivity

The phenomenological tradition contains rich but also quite diverse and even occasionally competing accounts of intersubjectivity. They all share, however, a rather critical appraisal of the argument from analogy. The classical argument runs as follows: In my own case, I can observe that I have experiences when my body is causally influenced, and that these experiences frequently bring about certain actions. In the case of others bodies, I can observe that they are influenced and act in similar manners, and I therefore infer by analogy that the behavior of foreign bodies is associated with experiences similar to those I have myself. Thus, the argument from analogy can be interpreted as an inference to best explanation. An inference bringing us from observed public behavior to a hidden mental cause. Although this inference does not provide me with indubitable knowledge about others and although it does not allow me to actually experience other minds, at least it gives me more reason to believe in their existence, than in denying it.

As phenomenologists have pointed out, however, the argument presupposes that which it is meant to explain. If I am to see a similarity between, say, my laughing or screaming and the laughing or screaming of others, I need to understand their bodily gestures and behavior as expressive phenomena, as manifestations of joy or pain, and not simply as physical movements. If such an understanding is required for the argument of analogy to proceed, however, the argument presupposes what it is supposed to establish. In other words, we employ analogical lines of reasoning only when we are already convinced that we are observing minded creatures but are simply unsure precisely how we are to interpret the expressive phenomena in question.

The argument also operates with some questionable assumptions. First, it assumes that my point of departure is my own consciousness. This is what is at first given to me in a quite direct and unmediated fashion, and it is this purely mental self-experience that is then taken to make possible the recognition of others. One is at home in oneself and one then has to project into the other, who one does not know, what one already finds in oneself.

Second, the argument also assumes that we never have direct access to another person s mind. We can never experience her thoughts or feelings. We can only infer that they must exist based on what we perceive, namely her bodily and behavioral appearances. Although both of these assumptions might seem perfectly obvious, many phenomenologists have rejected both. They have argued that we should not fail to acknowledge the embodied and embedded nature of self-experience and that we should not ignore what can be directly perceived about others. In other words, they have denied that our initial self-acquaintance is of a purely mental nature and that it takes place in isolation from others, and they have also denied that our basic acquaintance with others is inferential in nature.

The phenomenological account differs from the accounts proposed by the dominant positions within the theory of mind debate, that is, the theory-theory of mind and the simulation theory of mind. Both of the latter positions deny that it is possible to experience other minded creatures. It is precisely because of the absence of an experiential access to others that we need to rely on and employ either theoretical inferences or internal simulations. Both accounts share the view that the minds of others are hidden, and they consider one of the main challenges facing a theory of social cognition to be the question of how and why we start ascribing such hidden mental entities or processes to certain publicly observable bodies.

But it is no coincidence that we use psychological terms to describe behavior; indeed we would be hard pressed to describe it in terms of bare movements. Affective and emotional states are not simply qualities of subjective experience, rather they are given in expressive phenomena, that is, they are expressed in bodily gestures and actions, and they thereby become visible to others. There is, in short, something highly problematic about claiming that intersubjective understanding is a two-stage process of which the first stage is the perception of meaningless behavior and the second an intellectually based attribution of psychological meaning. In the majority of cases, it is quite difficult (and artificial) to divide a phenomenon neatly into a psychological aspect and a behavioral aspect think merely of a groan of pain, a handshake, an embrace. In the face-to-face encounter we are confronted neither with a mere body, nor with a hidden psyche, but with a unified whole. Scheler spoke of an expressive unity. It is only subsequently, through a process of abstraction, that this unity can be divided and our interest then proceed inward or outward.

Phenomenologists have tended to take an embodied perceptual approach to the questions of understanding others and the problem of intersubjectivity. We begin from the recognition that our perception of the other s bodily presence is unlike our perception of physical things. The other is given in its bodily presence as a lived body, a body that is actively engaged in the world. As Sartre pointed out, it is a decisive mistake to think that my ordinary encounter with the body of another is an encounter with the kind of body described by physiology. The body of another is always given to me in a situation or meaningful context, which is codetermined by the action and expression of that very body.

Some phenomenologists suggest that our understanding of others involves a distinctive mode of consciousness which they call empathy. Empathy is defined as a form of intentionality in which one is directed toward the other s lived experiences. The phenomenological conception of empathy stands opposed to any theory that claims that our primary mode of understanding others is by perceiving their bodily behavior and then inferring or hypothesizing that their behavior is caused by inner experiences similar to those that apparently cause that kind of behavior in us. Rather, in empathy, we experience the other directly as a person, as an intentional being whose bodily gestures and actions are expressive of his or her experiences or states of mind.

Most phenomenologists have argued that it makes no sense to speak of an other unless the other is in some way given and accessible. That I have an actual experience of the other, and do not have to do with a mere inference or imaginative simulation, does not imply, however, that I can experience the other in the same way as she herself does, nor that the other s consciousness is accessible to me in the same way as my own is. The second-(and third-) person access to psychological states differs from the first-person access, but this difference is not an imperfection or a shortcoming. Rather, the difference is constitutional. It is what makes the experience in question, an experience of

an other, rather than a self-experience. In short, we should not make the mistake of restricting and equating experiential access with first-person access. It is possible to experience minds in more than one way. When I experience the facial expressions or meaningful actions of others, I am having an experiential access to the life of the mind of others.

To get a fuller picture of the phenomenological take on intersubjectivity, it must be emphasized, however, that many phenomenologists have denied that sociality is primarily a question of a thematic encounter between individuals, where one is trying to grasp the experiential states of the other. On the contrary, the very attempt thematically to grasp the experiences of others is the exception rather than the rule. Under normal circumstances, we understand each other well enough through our shared engagement in the common world. We encounter others in worldly situations, and our way of being together and understanding each other is codetermined in its meaning by the situation at hand. In fact, in contrast to mentalistic theory-of-mind approaches that define the problem of other mind as amounting to the question of how to gain access to the other person s (hidden) mind, phenomenological approaches suggest that a more productive focus is on the other person s world. In effect, to understand other persons I do not primarily have to get into their minds; rather I have to pay attention to the world that I already share with them. On the phenomenological account, self, world, and others belong together; they reciprocally illuminate one another, and can only be understood in their interconnection.

See also: Concepts and Definitions of Consciousness; Intentionality and Consciousness; Mental Representation and Consciousness; Meta-Awareness; The Mind-Body Problem; Perception: The Binding Problem and the Coherence of Perception; Philosophical Accounts of Self-Awareness and Introspection; Psychopathology and Consciousness; Self: Body Awareness and Self-Awareness; Self: Personal Identity; Self: The Unity of Self, Self-Consistency; Sensory and Immediate Memory; Subjectivity; First- and Third-Person Methodologies; Visual Experience and Immediate Memory.

Suggested Readings

Drummond JJ (1990) Husserlian Intentionality and Non-Foundational Realism. Dordrecht: Kluwer Academic Publishers.
Gallagher S (2005) How the Body Shapes the Mind. Oxford: Oxford University Press.
Gallagher S and Zahavi D (2008) The Phenomenological Mind. London: Routledge.
Gurwitsch G (1966) Studies in Phenomenology and Psychology. Evanston: Northwestern University Press.
Heidegger M (1996) Being and Time. Stambaugh J (trans.). Albany: SUNY Press.
Henry M (1975) Philosophy and Phenomenology of the Body. Etzkorn G (trans.). The Hague: Martinus Nijhoff.
Husserl E (1982) Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy. First Book. Kersten F (trans.). The Hague: Martinus Nijhoff.
Leder D (1990) The Absent Body. Chicago: Chicago University Press.
Merleau-Ponty M (1962) Phenomenology of Perception. Smith C (trans.). London: Routledge and Kegan Paul.
Moran D (2000) Introduction to Phenomenology. London: Routledge.
Sartre J-P (1956) Being and Nothingness. Barnes HE (trans.). New York: Philosophical Library.
Scheler M (1954) The Nature of Sympathy. Heath P (trans.). London: Routledge and Kegan Paul.
Sokolowski R (2000) Introduction to Phenomenology. Cambridge: Cambridge University Press.
Thompson E (2007) Mind in Life: Biology, Phenomenology, and the Sciences of Mind. Cambridge, MA: Harvard University Press.
Waldenfels B (2001) Das leibliche Selbst. Vorlesungen zur Phänomenologie des Leibes. Frankfurt am Main: Suhrkamp.
Zahavi D (2003) Husserl's Phenomenology. Stanford: Stanford University Press.
Zahavi D (2005) Subjectivity and Selfhood. Cambridge, MA: MIT Press.

Biographical Sketch

Dan Zahavi (1967) is a professor of philosophy and the director of the Center for Subjectivity Research at the University of Copenhagen. His research areas include phenomenology, philosophy of mind, and cognitive science. He is a past president of the Nordic Society of Phenomenology and a current coeditor in chief of the journal *Phenomenology and the Cognitive Sciences*. He has written and edited more than 15 books including *Self-Awareness and Alterity* (1999), *Exploring the Self* (2000), *Husserl und Transcendental Intersubjectivity* (2001), *One Hundred Years of Phenomenology* (2002), *Husserl's Phenomenology* (2003), *The Structure and Development of Self-Consciousness* (2004), *Subjectivity and Selfhood* (2005), and *The Phenomenological Mind* (with Shaun Gallagher) (2007).

Philosophical Accounts of Self-Awareness and Introspection

W Seager, University of Toronto Scarborough, Toronto, ON, Canada

Introduction: Self-Awareness

This article provides an overview of philosophical approaches to the topic of self-awareness and introspection. It does not attempt to survey the voluminous scientific work in this area (see 'Suggested Readings' for references). Some remarkable experimental work has been done, which by and large seems to support the hypothesis that we are quite poor at self-understanding. It is worth noting that the experimental work provides some support for what are called 'evidence theories' below which assert that self-knowledge involves self-interpretation (and hence is subject to a host of distortions and biases, often engendered by possession of background 'theory' of mind and self). However, the scientific work generally investigates relatively sophisticated and high-level self-knowledge rather than the core of basic introspectible mental states about which our accuracy remains unchallenged. That is, while it may well be that our ability to introspectively know such things about ourselves as that we are not jealous of a colleague is highly suspect, or if what we are feeling is really love, this does not call into question our introspective knowledge that we are, say, now in pain or seem to be seeing an apple. We seek here to investigate the philosophical roots of self-awareness and introspection.

It is important to distinguish the two ideas in the title of this article. Self-awareness is a broad concept that encompasses a much wider range of phenomena than does the concept of introspection. The remarks immediately following are intended primarily to highlight the domain and nature of introspection; for more on nonintrospective self-awareness see the entry for Self: Body Awareness and Consciousness. Some remarks here about the broader phenomenon are nonetheless in order.

It is legitimate to include within self-awareness our ongoing awareness of our own bodies, especially proprioceptive monitoring of such things as the positions of our limbs, muscle tone, joint extension, and so on. In the sudden absence of such information, our sense of ourselves and our place in the world would be radically altered. But 'body image' extends far beyond these basic physical parameters, to include culturally specified bodily norms, both of appearance and deportment.

One might also include various forms of what might be called 'situatedness' within the purview of self-awareness. At least two forms are notable. One is physical or spatial situatedness – our sense of orientation in local space, which includes implicit action patterns, sensitive to the spatial distribution of objects with the potential either to interfere with or aid our actions (an example might be the way we smoothly avoid obstacles as we move around a room). This sort of awareness is seldom the focus of attention; indeed and famously, focusing attention on the smooth interaction we normally enjoy with the environment interferes with, rather than enhances it. Nonetheless, it is a foundational part of our awareness of ourselves in the world. It seems appropriate to include our sense of spatial situatedness within the shadowy (and still insufficiently explored) realm of what William James called the 'fringe' of consciousness. Spatial situatedness is just one of the many low-level or background features of consciousness that underlie our sense of 'fitting into' our environment.

It is interesting and highly suggestive that our sense of, or at least certain aspects of, spatial situatedness can remain in the absence of other forms of conscious awareness of self and environment. An illustration of this is provided by certain forms of visual agnosia such as the well known case of DF, who lacks the ability to recognize objects (even at the level of a basic geometric form, such as the difference between a square and a triangle), but who nonetheless retains the ability to grasp objects appropriately and is able to walk through complex environments without

difficulty. Although DF's abilities are amazing in themselves, what is interesting from our perspective is the oft noted 'confidence' DF exhibits when interacting with objects via visual information. Thus it seems that her sense of (local) orientation in the world has been retained despite the complete absence of consciousness of the objects as such, around her.

Another important sort of situatedness is 'social situatedness.' Social situatedness is our sense of orientation toward others as they are reacting to their own perception and understanding of us. This encompasses both sophisticated culturally mediated aspects of the social environment but also much more basic ongoing 'assessments' of the emotional reactions of others to one's own behavior. There has recently been much debate about how we acquire knowledge of the mental states of others, especially focused on the issue of whether such knowledge is acquired via simulation – the empathic ability to get ourselves into a mental state that is hopefully similar to our target's mental state – or via an internal theory of the mind. Probably both are necessary to generate our sense of 'belonging' in our social environment. This certainly involves an ongoing 'comparison' of our own mental states with those of the people around us, which gauges the appropriateness of these states in relation to each other. Social situatedness and spatial situatedness come together in our sense of 'personal space,' the jealously guarded, but culturally relative need to preserve a certain distance from others.

As in the case of spatial situatedness, there can be breakdowns in social situatedness, many of which are very familiar (e.g., the paranoia exhibited in a number of syndromes). The very striking, rare, and bizarre Capgras syndrome illustrates the complexity and deep importance of social situatedness. Victims of this affliction form delusional beliefs that those closest to them (typically parents or spouses) have been replaced by 'duplicates' (twins or even – in modern times – robots). The delusion is of such irrepressible strength that tragic consequences sometimes result. The cause of Capgras syndrome is unknown, but the most favored hypothesis is that it results from a breakdown in the emotional response system, so that a kind of dissonance is created between the perceptual familiarity of someone and a lack of any appropriate feelings toward that person. The dissonance is resolved by the 'hypothesis' that a loved one has been replaced by a stranger toward whom no emotional attachment exists.

These forms of self-awareness are vitally important for normal functioning, and while they are in all probability a part of the necessary background enabling conditions for introspection, they do not themselves constitute introspection.

The Characteristics of Introspection

Introspection is a special kind of knowledge we have of our own mental states. One way it is special is that it is does not appear to be mediated by inference from other knowledge we may possess. For example, it is conceivable that one might come to know that one was in a certain mental state by inferring this from someone else's opinion (e.g., one's psychoanalyst). This might be a legitimate form of self-knowledge, but it would not be introspective knowledge.

Thus one important question about introspection is the 'mechanical' one of how it works to provide us with knowledge of our own mental states in this 'special' way. A naturally connected further question is whether the specialness of introspection means that introspective knowledge is somehow more secure or indubitable compared to other forms or sources of knowledge.

However, before investigating these issues a more basic characteristic of introspection must be noted. Introspection produces propositional knowledge about our own mental states. Thus it is crucial to emphasize that since all such knowledge is conceptual in nature, introspective knowledge must involve the application of concepts – mental concepts – to ourselves. This is in contrast to the fundamentally nonconceptual aspects of the sorts of self-awareness discussed above and also in contrast to the 'first-order' awareness provided by consciousness itself.

When we introspect we find any number of mental states of a quite staggering variety: sensory, bodily, cognitive, emotional, or more broadly affective, and so on. Introspection is a kind of knowledge, and in order to 'know' that we are in any of these mental states, we must possess the

concepts necessary to categorize them correctly. Luckily, we are all in possession of a pretty full range of concepts well tailored for this task. The totality of them and their 'dynamics' are frequently referred to as folk psychology, and the invention of this system of concepts should perhaps be ranked as the greatest and most indispensable innovation humankind has achieved. The core idea of folk psychology is based on two foundational mental states: belief and desire. Belief represents the world as we take it be, desire represents the world as it should be (at least from our own point of view). Action results from the rational evaluation of how our beliefs provide a transformation, taking the world as it is to the world as we wish it to be. Thus a thirsty man heads toward rather than away from the river. Who knows how this scheme was first articulated, although obviously it long postdates the existence of creatures who 'have' beliefs and desires, but know nothing of them as such. More recently, folk psychology has been explicitly characterized as a kind of theory, of which the biological basis, development, and significance in both normal and pathological cases is the subject of ongoing current debate. And, of course, its complexity, measured both in the kinds of mental states recognized and their interactions, has grown immensely.

Whether folk psychology ought to be considered a theory or not, it provides us with a hugely powerful way to understand, explain, and predict the thoughts and actions of our fellow human beings (not to mention helping us to successfully deal with animals). The subtlety and power of advanced folk psychology is most evident in literature. Here, drawn pretty much at random from Jane Austen's *Sense and Sensibility* is an illustrative passage:

> . . .Marianne awoke the next morning to the same consciousness of misery in which she had closed her eyes.
>
> Elinor encouraged her as much as possible to talk of what she felt; and before breakfast was ready, they had gone through the subject again and again; with the same steady conviction and affectionate counsel on Elinor's side, the same impetuous feelings and varying opinions on Marianne's, as before. Sometimes she could believe Willoughby to be as unfortunate and as innocent as herself, and at others, lost every consolation in the impossibility of acquitting him. At one moment she was absolutely indifferent to the observation of all the world, at another she would seclude herself from it for ever, and at a third could resist it with energy. In one thing, however, she was uniform, when it came to the point, in avoiding, where it was possible, the presence of Mrs. Jennings, and in a determined silence when obliged to endure it. Her heart was hardened against the belief of Mrs. Jennings's entering into her sorrows with any compassion.

Every sentence of this passage deploys mentalistic notions, exhibiting a supreme knowledge on Austen's part of how mental states interact with each other, to the end of giving us an excellent sense of the characters of the persons described.

Of course, Austen could not have written this without a deep introspective awareness of her own mental states, no less than her evident acuity in understanding others. Introspection is then the application of these mentalistic concepts to ourselves, the very same concepts we use to describe the inner lives of others.

While the range of mental ascription is the same for self-as well as other descriptions, introspection appears to possess some epistemic properties that decisively distinguish it from our knowledge of the mental states of others. We can mark out this distinctiveness in terms of a few (putative) core properties of introspective knowledge: privileged access, immediacy, infallibility, and transparency.

Privileged access is the idea that introspective knowledge is in some way more secure and has a basis distinct from that of others' knowledge of our mental states. While it is possible, and generally quite easy, for other people to know what we are thinking or feeling, our own access to these states is of a different nature.

One of the most famous defenders of introspection as a special and epistemically privileged mode of mental access was René Descartes. Descartes's dualism held that the mind was an independent substance, causally connected to the body by a divine act, which linked a particular mind with a particular body in a system which Descartes's called the 'substantial union.' All conscious experience is a purely mental phenomenon, hence any knowledge of the body is at best indirect. Of course, this general thesis led Descartes to skeptical problems about the nature and existence of the

external world. But skepticism does not extend into the realm of the mind. It is impossible to be wrong about one's own existence because any mental act guarantees the existence of the thinker. But how do we know that we are conscious? It would seem to be impossible to be wrong about this, even if we grant the possibility of error about the content of one's mental states. For try to imagine or conceive that you are not at this moment conscious. The act of conceiving this is by itself a conscious act. Evidently we have some kind of very special access to at least some of the features of consciousness.

Introspective access also seems to be unmediated by any other knowledge. When, for example, you are suffering from the pain of toothache there does not seem to be anything 'between' the pain and your introspective awareness of the pain. It is important to emphasize that these nonetheless are entirely different states. The pain is a sensation with its own features: sensory qualities of feeling and location plus a distinctive sense of aversion (we know these are distinct because they can become dissociated in certain conditions, notably opiate analgesia). It does not involve the concept of pain, as evidenced by the fact that animals and very young children or infants can feel pain even though they seem to completely lack any mentalistic concepts. The introspective awareness of pain however requires both the existence of the conscious sensation of pain and also possession of the concepts needed to encode knowledge of it.

This may suggest a worry about the thesis of immediacy: what guarantees the accuracy of the act of conceptualization? This worry is misplaced at this point and is really directed at the claim of infallibility. Introspective knowledge is infallible if it is impossible for the deliverance of introspection to be in error, or, more precisely, if it must be the case that any subject who believes on the basis of introspection that they are in mental state S is in S. An example where infallibility seems highly plausible is one's introspective awareness of intense pain. Imagine your reaction to a dentist who replied to your anguished complaint of excruciating toothache that you were mistaken – that you were not in pain at all. The seeming inconceivability of the dentist being right about this illustrates the infallibility of introspection.

Contrast infallibility with transparency, which is the thesis that nothing in the mind is hidden from introspection, or, if a subject is in mental state S, then S can always become introspectively aware of S. A stronger form of transparency would replace 'can become' with 'is.' Once again, intense pain presents a plausible illustration of transparency. Imagine that you are experiencing an excruciating toothache. Now try to imagine that you are not or even cannot become introspectively aware of the pain. This scenario seems almost as absurd as the idea that one could be mistaken about being in intense pain.

Despite the superficial plausibility of infallibility and transparency in these cases, and despite a historical tendency to embrace them (e.g., as in Descartes), upon inspection, neither infallibility or transparency seem likely to be true, or at least not true of all mental states in general. The idea that there could be unconscious mental states, a thesis anathema to Descartes, but embraced by the seventeenth century philosopher Leibniz and now utterly commonplace, puts pressure on the transparency thesis. However, given the weak formulation above, transparency fails only if there are mental states of which introspective access cannot be attained. Freudian-type examples thus seem to fail to undermine transparency since they are such as can – under the appropriate circumstances of psychoanalysis – become introspectively known. But a host of 'subliminal' perceptual states which have been shown to affect behavior in ways consistent with their mental status show quite clearly that transparency is not true in general. For example, in semantic priming it can be shown that the 'meaning' of a word which is presented visually for too short a time for subjects to be aware of it affects processing of consciously apprehended words. It also seems entirely possible that a conscious experience could be so overwhelming that their subjects would not be able to 'step back' to engage in introspection (perhaps the first time one attempts sky-diving would provide an example of such an experience). It is also possible that the character of certain experiences would be necessarily altered by any attempt to introspect them. It is also worth mentioning the pervasive phenomenon of self-deception, which strongly suggests that there must be vast areas of the mind hidden from

introspection. At least, it is hard to see how the self-deceived could persist in their delusion if they were introspectively aware of the underlying cognitive mechanisms (e.g., denial, self-interest, etc.)

Infallibility is not much more plausible than transparency when regarded as a general thesis. Humans are notoriously adept at self-deception and will with apparent sincerity avow mental states (beliefs, desires, etc.), which it is very hard to believe they actually possess. A typical example might be a man who steadfastly maintains that he believes his wife is faithful, but who persistently and carefully avoids all opportunities that might present evidence suggesting otherwise. The best account here might well be that he actually believes his wife is unfaithful but 'cannot admit it to himself' as we say. Closely akin to – perhaps a form of – self-deception is the phenomenon of cognitive dissonance in which belief and desire formation is distorted by various – usually emotionally charged – factors (the fable of the fox and the grapes provides a homely illustration: the fox realigns his beliefs about the grapes on the basis of their unavailability). It appears clear in all such cases that there is a large gap between how we think our minds are working and their actual operation, which is at odds with infallibility.

Despite the difficulties with these putative features of introspection, it remains clear that introspective knowledge of our own mental states is a special form of knowledge, possessing a clear form of privileged access, and at least a high degree of transparency and infallibility.

Theories of Introspection

Modern views of introspection broadly divide into four quite distinct forms, which I will label here 'perception' theories, 'evidence' theories, 'transcendental' theories, and 'displaced perception' theories. All of them, and certainly the first two and the fourth approaches, take one of their primary tasks to be the integration of introspection into a naturalistic view of the world. The main force of the term 'natural' is 'anti-Cartesian,' specifically anti-Cartesian dualism. These theories are also opposed to Descartes's endorsement of infallibility and transparency and aim to provide a theory of introspection that will smoothly integrate with the scientific world view and current psychological evidence.

Perceptual Theories

The most obvious approach to introspection is simply to equate it with perception. According to naturalism, mental states are states of the brain (more or less) and so introspection is perception of the brain. Of course, if a surgeon holds a mirror up so that I can see my brain as she operates on it or if I am watching the functional magnetic resonance imaging (fMRI) readout of my own brain operations, I am not, in virtue of this unusual perception of my own brain, also introspecting. Just as in the case of perception, introspection requires its own sense organ, which must itself be a part of the brain with its own specialized 'inner receptors.'

Thus is born the inner-scanner theory of introspection, which postulates a functional brain system – let's call it the 'I-scanner' – designed, by evolution presumably, to actively seek out information about those brain systems which realize mental states. David Armstrong explains the action of the I-scanner thus:

> In perception the brain scans the environment. In awareness of the perception another process in the brain scans that scanning. . . .

The I-scanner is by its nature rather more intimately connected to the objects of its perceptual capabilities than are our various sensory 'scanners,' and this is supposed to explain both the accuracy of introspection and the long-standing philosophical illusion of perfect introspective infallibility and transparency. According to I-scanner theorists, introspection is, as a matter of fact, 'practically' infallible, but no more so, and only in the way that perception of nearby objects directly before us in good light, and so on, is characteristically infallible. However, following the analogy of sense perception one can envisage at least the philosophical possibility of radical errors in introspection. For example, an adept science fictional brain surgeon could, in principle, activate a variety of I-scanner states that would yield introspective belief in all sorts of mental states, none of which

were actually present. This is the natural result of downgrading the privileged access supposedly enjoyed by introspection to a mere generally reliable causal link.

One could raise several objections against the I-scanner theory, but the most salient difficulty stems from a curiosity of the phenomenology of introspection. If introspection were properly understood as a kind of perception, one would expect there to be a distinctive phenomenology of introspection; mental states ought to have 'introspectible qualities' just as the objects of perception have perceptible qualities such as color and form. Unfortunately, it is a plain fact that there is no distinctive introspective phenomenology whatsoever. That you have current introspective access to your perceptual states of consciousness no one will deny, but all the phenomenology you can find is exhausted by the perceptible qualities of the objects of your current perception. Attending to your perceptions may well heighten or otherwise alter your perceptual state, but it does not introduce a new realm of introspectible qualities. It would be desperate folly to claim that the introspectible qualities of mental states are an exact 'copy' of the perceptible qualities of objects, and even worse to claim that we are never aware of anything but introspectible qualities of the mental. Lacking the nerve to make this latter assertion, we might ask the I-scanner theorist why, when we introspect our perceptual states of consciousness, the evident 'outerness' of perception is not replaced by an 'innerness' appropriate for introspective awareness of the brain (or, even, the mind)?

The introspection of intentional mental states is generally of greater significance to us than the introspection of perceptual states. The I-scanner theory seems to fare even less well in this domain. It is very difficult to convince oneself that beliefs and desires are known by a process analogous to perception. Intentional states do not have a set of quasi-perceptual qualities by which they are internally recognized. This is evident in common wisdom. If Mary wonders whether she really loves Tom, she 'thinks' about Tom and their relationship. The advice to forget about thinking and just look inside oneself gets one nowhere. Of course, there are various feelings Mary might have when she thinks about Tom and she can be conscious of these as feelings, but the question is, are they 'signs' of love. And there, the problem is that there just does not seem to be anything more to 'look at'; our intentional mental states do not form an inner world of objects with their various introspectible properties.

A defender of a perceptual account of introspection might try the following reply. Couldn't Mary imagine, for example, what it would be like to live with Tom? She might imagine various domestic scenes or possible situations and 'play out' how Tom might act in these situations. Of course she could and this is no doubt part of what thinking about our own intentional states involves, but the central point is that this is not introspection. Consciously imagining something is 'not' in and of itself introspection. I can vividly imagine the Canadian flag, with its bright red stripes and maple leaf, but I am not necessarily introspecting when I do this. Of course, it is not a great leap from imagining to introspection. Knowing that and what I am imagining, as well as how I am imagining it, is introspective knowledge, but I do not get this knowledge by somehow perceiving my imagining of the flag. My imagining does not look like anything, though my image of the flag does (at least there is a distinctive and obviously perceptual phenomenology of imagining). But 'looking' at my image is just the imagining itself, not an act of introspection, although of course I can, even in imagination, attend to particular aspects of the imagined sensory qualities.

Evidence Theories of Introspection

What I shall call 'evidence theories' provide an entirely different approach to introspection. The basic claim of evidence theories is that our knowledge of our own mental states is formed in precisely the same way as our knowledge of others' mental states; there is no asymmetry between the 'first-person' and 'third-person' knowledge of the mental. Gilbert Ryle was the most forceful exponent of such theories. Ryle put forth his view succinctly as follows: ''Our knowledge of other people and ourselves depends upon our noticing how they and we behave.''

There are some advantages to the evidence theory. It dissolves the surprisingly vexing problem of

other minds, or at least transforms it into no less a problem about other minds than about our own. However, this advantage depends upon the assumption that the problem of other minds does not stem from a genuine and fundamental asymmetry in the relation we bear to our own mind and the minds of others. Thus, an obvious objection to the evidence theories turns the advantage on its head. A direct refutation of the evidence theory follows from its denial that we stand in relation to our own states of mind, quite distinct from the relation we bear to the mental states of others.

The evidence theory also exhibits a nice theoretical economy. Since it is given that we do have methods for attributing mental states to others it is rather elegant to enlist these methods for self-attribution. This eliminates any theoretical need for a special and, at least in its traditional guise, epistemically suspicious introspective faculty or sense, or even any special mode of self-interpretation. It also retains an acceptable, albeit rather weak, form of 'privileged' access to our own mental states in its allowance that we all generally have access to more, if not a different kind of, evidence about ourselves than others, while possessing abundant means to account for the evident failures of self-knowledge.

Now, it is hardly surprising that in some cases people will attribute mental states to themselves on the same basis that they attribute them to others, for people are capable and fond of thinking about themselves and their motivations. As noted above, the complex system of concepts of mental states and their dynamic interconnection of folk psychology provides a very rich and hugely successful system for understanding ourselves and others. We quite naturally apply this knowledge to ourselves no less than to others. But this is a frail basis upon which to build an evidence theory of introspection for it simply ignores the vast range of self-knowledge to which the model applies very badly if at all.

It is, as philosophers have long remarked, extremely easy to gain introspective knowledge of one's own perceptual states. In fact, this ease of access provides the core intuitive support for the special features of introspective knowledge such as infallibility and transparency. But it is evident that this access does not require one to observe, imagine, or remember one's own behavior or utterances. I don't need to say to myself "I am seeing red" or hear myself saying this to someone else, to know that I am in a state of 'seeing red,' nor do I need to imagine red (in fact, when one is actually seeing red, it is next to impossible to imagine red at the same time, and obviously such imaginings would gain nothing over the perception). Of course, 'seeing red' is not by itself an introspective act, so perception is not itself an objection to the evidence theories, but introspective knowledge of perceptual states is freely available simply by virtue of the fact that these are conscious states. We simply do not have to watch what we do or say to know that or what we are seeing or hearing, whereas the only way to know what perceptual experiences others are having is by observing this sort of behavior.

The evidence theory no more plausibly explains introspective knowledge of the vast majority of intentional states such as beliefs and desires than it does our knowledge of perceptual states. Knowing that I believe that, for example, 2 + 2 = 4 is (a rather trivial) case of introspective knowledge. It is ludicrous to suppose that my knowledge of this belief depends upon or stems from an observation of my own behavior or from some act of imagination (what would it be: imagining myself writing down '2 + 2 = ' and then noting that my imaginary self completes the equation with '4'?). It is no different in the case of desire. I know that I like strawberry pie, but not because I find myself ordering it in a restaurant, or because I can remember that I have ordered it often in the past, or because it is easy to imagine myself ordering it in the future, or because I remember hearing myself say frequently enough "I like strawberry pie," or because I can easily imagine hearing myself say this, or yet because I frequently say to myself "I like strawberry pie" (which in fact I do not).

Nonetheless, the evidence theory is not entirely hopeless. Although it cannot be regarded as a serious candidate for a theory of introspection *per se*, not all my self-knowledge is introspective. The evidence theory is plausible only for cases of sophisticated or 'difficult' attribution of complex intentional states and only a tiny fraction of my self-knowledge, albeit a highly significant fraction, involves these. This zone of accuracy for the

evidence theory is important. It points to the very real intellectual and practical problems we face in knowing ourselves as complicated intentional beings within the complex social environment we have constructed for ourselves. A good part of our self-knowledge stems from self-interpretation as the evidence theory claims. In fact, it is arguable that the most important elements of our self-knowledge are based on self-interpretation of the kind the evidence theory appeals to – introspective knowledge is generally more easily obtained and the majority of it is relatively trivial. But despite the insights the evidence theory embodies, it seems to be entirely inadequate as a general account of introspection. If we insist on regarding it in this light the evidence theory is seriously misleading about the fundamental source and nature of introspective knowledge.

Transcendental Theories of Introspection

In philosophy, it is always at least thought to be possible that the appearance of a problem is mere appearance, which can be resolved by a more careful analysis of the concepts involved. And so it is with the problem of introspection. The 'transcendental' theories of introspection are really more of a denial that there is a problem of introspective knowledge than a substantive theory of introspection. I adopt the term 'transcendental' from Kant's usage in which a transcendental 'proof' of some subject matter depends on laying out the conditions of its possibility. For example, Kant gives a transcendental proof that our experience conforms to the dictates of causality (or reveals a world to which the category of causation applies) by trying to show that one of the conditions for the possibility of experience is that it conforms to causality. Thus, given that we do have experience, we can infer that it meets the condition of causality.

How is this supposed to work as applied to the case of introspection? Of course, any transcendental account of introspection will admit the existence of extensive self-knowledge and will go so far as to allow that this sort of knowledge has a very special nature and perhaps has a more secure status than most other kinds of knowledge. But the story of introspection as told by transcendentalists is quite different from those we have heard so far. The basic idea is that a subject S having accurate self-knowledge is a condition for the possibility of all attributions of mental states to S. Philosophers, notably Donald Davidson, have argued that a general failure of self-knowledge would render the interpretation of a subject as believing and desiring incoherent. Roughly speaking the argument goes like this. Suppose that in general some person acted as if they believed, say, that elephants have tusks, even going so far as affirming that elephants have tusks (i.e., uttering 'elephants have tusks'), but denied that they believed this or even claimed that they could not tell whether or not they believed that elephants have tusks. Such 'metabehavior' would render the initial interpretation of the subject's belief unsustainable and if such paradoxes were the norm rather than very rare exceptions, it would throw into doubt the idea that the person had beliefs and desires at all. Note that this sort of *a priori* defense of self-knowledge leaves room for errors of introspection, so long as these errors remain relatively infrequent and do not become, so to speak, flagrant.

A very elegant and compressed argument for a transcendental theory of introspective knowledge has been provided by Akeel Bilgrami who ingeniously links self-knowledge with the possibility of interpreting people as moral agents. Roughly speaking, the proposal is that only those who are aware of, or know, what they are doing can be held morally responsible for the consequences of their actions. Then, Bilgrami argues, since it is a condition for the possibility of taking others to be persons that they be thought of as morally responsible for their actions (at least most of the time), it is a condition for the possibility of taking others to be persons that they (generally) know what they are doing, and one cannot know what one is doing without some insight into the intentional states, which actually explain one's actions. Self-knowledge of one's own intentional states thus emerges as a condition of personhood.

An immediate problem with such accounts is that they threaten to implausibly reduce the kinds of creatures which can possess minds. Intuitively, many creatures enjoy consciousness and a fairly rich mental life, while lacking the conceptual equipment necessary for introspective knowledge.

I think a very large number of animals fall into this category (and it may well be that no animals save for human beings are capable of introspection). Probably, human infants and even young children perhaps up to age three or so also have not yet grasped the mentalistic concepts needed for introspection (there is a large literature on the way that children acquire mental concepts and in general it seems to support the idea that full acquaintance with folk psychology arrives rather late and may have a genetic component). Thus there is a serious problem lurking here with respect to the minds of animals and young children. If having a mind itself somehow entails that the subject possesses generally accurate self-knowledge then they will end up mindless. I take it that this is not a welcome result and thus requires some kind of hierarchy of minds, only some of which will be capable of introspection. It is not altogether clear that those of a transcendentalist bent can accommodate a scale of more or less primitive animal minds.

Leaving that worry aside, there is no doubt that there is something correct in transcendental analyses like these, but it is very far from clear that they either solve or dissolve the problem of introspective knowledge. This is because of two related failings that reveal the inability of the transcendental theories to evade the need to provide a positive account of introspection.

First, although the transcendentalist may well be right that generally accurate self-knowledge is a condition of the possibility of ascribing mental states to others and also ourselves, one must nonetheless ask how creatures for which this possibility arises are structured so as to realize the possibility. The original transcendentalist, Kant, consigned this structure to an unknowable noumenal realm into which human thought dare not venture on pain of unintelligibility. This is hardly a strategy modern, naturalistically inclined transcendentalists about self-knowledge should wish to emulate. Nor is it defensible to refuse the request for an explanation of how self-knowledge is attained simply by appealing to the transcendental proof that, after all, it 'must' be attained. Here, the transcendental methodology is closely akin to the use of the anthropic principle in cosmology, which well illustrates the general limitations of a transcendental or anthropic account of some fact. For example, we observe that the Earth is neither too near nor too far from the Sun to support life. This is no surprise, for the Earth being at such a distance is a condition of the possibility of observers. The anthropic, 'transcendental' account of the Earth–Sun distance is, however, no substitute for an explanatory tale of the genesis of the solar system in which each planet attains its appointed place via natural forces, with no help from the fact that the Earth would someday spawn astronomers who can measure the Earth–Sun distance. Every realized possibility must be realized via some mechanism; so too our introspective capacities must have some actual source which, when articulated, will amount to a positive theory of introspection.

Second, it is evident that self-knowledge is a cognitive achievement that takes some effort. Self-knowledge, like any other knowledge, does require evidential warrant. When we have such knowledge, we generally also know why we know what we know about ourselves. This point is intended to be uncontroversial and not metaphysically or epistemologically deep. Suppose I report that I am in pain and someone asks me how I know this. The obvious, and correct, answer is ''I feel it'' (given that, of course, I know what pain is). Self-knowledge cannot be freely generated simply by asserting something about oneself. Self-knowledge must answer to the same epistemic canons to which any potential domain of knowledge must answer, such as those of evidence, warrant, and the possibility of error. Further, my self-knowledge must be integrated with my knowledge of others' mental states and my knowledge of the world. But then the range and type of evidence required for self-knowledge is problematic and demands an account.

Displaced Perception Theories of Introspection

Recently, a novel theory of introspection has been advanced as an adjunct to an account of consciousness that identifies consciousness with certain forms of mental representation (the theory is usually known as the representational or intentional theory of consciousness and has been developed in most detail by Fred Dretske and Michael Tye).

It denies that there are peculiarly mental features of all experience (the so-called qualia), and instead asserts that consciousness is essentially a matter of presentation to the mind of the content of mental representations. For example, if one imagines the Canadian flag the experience is of a vivid red maple leaf shape on a white background flanked by two similarly bright red bars. The mechanism of consciousness is that the content of a representation of the flag is the content of the conscious experience. This has certain obvious advantages. It avoids postulating bizarre entities with little prospect for assimilation in the scientific picture of the world (e.g., the imaginary Canadian flag which cannot be found anywhere in the physical universe). The prospect of integration with a naturalistic outlook is also present, or at least reduces to the problem of naturalizing mental representation. It is worth mentioning here that the main rival to this account of consciousness is the so-called higher order thought (HOT) theory (see Seager article on consciousness). The mechanisms of introspection of the two theories are quite similar insofar as both require conceptualization of the mental state, as mental as a condition for conscious introspection. (For HOT theories, unconscious introspection exists and in fact is what accounts for the target of the introspective state being a conscious state.)

The theory of introspection that naturally accords with this view of consciousness asserts that introspection depends on an interplay of consciousness and concepts via a mechanism called displaced perception (which extends beyond the realm of introspection). An example of Dretske's is, learning that the postman has arrived by perception of the dog's barking. To get such knowledge one must hear the dog (consciousness) and one must also know what the dog's barking signifies (concepts). Introspective knowledge of our own perceptual states similarly requires that we consciously perceive, but also that we know what perceiving is. As we have noted, knowledge is conceptual and so requires an appropriate field of concepts for its formulation. Introspective knowledge requires the field of concepts that together form our notion of the mind, or folk psychology, a set of concepts ready to hand for all reasonably mature human beings.

This theory does not require that there be any perception of our mental states as such, but instead simply requires that we possess certain mentalistic concepts that we can successfully apply to ourselves as we perceive the world (including our own bodies). Thus, when I am consciously perceiving red, I have introspective knowledge of my experience because I can and do apply the concept of perceiving red to this instance of my conscious perceptual experience. I don't need to perceive my perceiving (as the I-scanner theory asserts at the bottom) to make this application any more than I need to perceive my perceiving of a barking dog to apply the concept of 'barking dog' to that object. Of course, I do need to be perceiving red to make the introspective application of the concept 'perceiving red,' but that is simply a matter of being perceptually conscious.

The displaced perception model of introspection shares with the evidence theories the idea that introspection requires an input of information from which mentalistic conceptualization will follow, but it denies that introspection depends on exactly the same kind of evidence we use in attributing mental states to others. Instead, it exploits the fact that we are conscious beings and we can develop or acquire the ability to describe our consciousness at, so to speak, one step removed, as a mental feature of ourselves. The ability to comprehend the epistemic distance between the world and the experience of the world is not some kind of benighted proto-Cartesianism; it is a vital step toward self-consciousness and an awareness of one's own identity.

One difficulty with the displaced perception model is that it appears to transform introspective knowledge into inferential knowledge, thus disparaging even the acceptable level of infallibility and transparency which introspection evidently possesses. That is, it does not account for what is truly special about introspective knowledge. It does indeed seem odd to say that we infer that we are experiencing, say, the taste of an apple based on our consciousness of the apple's taste. Although many proponents of the displaced perception model of introspection liken introspective knowledge to inferential knowledge, the idea can be resisted. Instead, it may be possible to liken introspective knowledge to nothing more than

conceptual categorization. It does not seem quite right to say that one 'infers' that a dog is in the yard (a dog which is perfectly visible and in fact is seen) from some underlying purely sensory material, even if such material plays a role in conceptualization. Similarly, we don't need to infer that we are 'seeing' a dog from our visual experience. We need only be in possession of the concept of seeing and have the ability to apply that concept in the appropriate circumstances.

An interesting question about this model of introspection is how it is to be extended beyond the realm of introspective knowledge of conscious perceptual states, to include intentional and emotional mental states. For example, consider the problem of how we attain introspective knowledge of our own beliefs or desires. In all probability you, the reader, believe that there are no giraffes on Mars and you have, at least you have now, introspective knowledge that you possess this belief. But it does not seem that one can discover any special internal mental feature by which one recognizes, first, which belief is in question (e.g., giraffes live on Mars vs. grass is green) and, second, whether that content is actually one which is believed (as opposed, say, to being entertained, wondered about, doubted, etc.). Instead, it seems that one knows that one believes that no giraffes live on Mars simply in virtue of figuring out (or recalling, if for some reason you had worked this out in the past) that no giraffes live on Mars. It is as if one says to oneself: ''giraffes on Mars? No way (what would they breathe, how did they get there, etc.),'' and from this, one both verifies that giraffes do not live on Mars and one believes that giraffes do not live on Mars.

The most straightforward way to assimilate this insight about believing is to posit some feature of conscious experience, which we could label 'taking for true.' There is, on this view, a basic experiential difference, which differentiates the thought ''I am six feet tall'' when thought by me and when thought by someone who knows their height to be less than six feet. I take this for true, it seems correct to me, I happily absorb or maintain it within my cognitive economy; the other person rejects it, finds it – in the literal sense of the term – incredible, and will not take it up into his belief system.

If the postulate that there is an act of the mind of 'taking for true,' then the displaced perception model of introspection can be easily extended to the case of belief. One knows that one believes that giraffes do not live on Mars because one 'sees' the truth of this proposition and knows that this licenses application of the concept of belief to it.

The model can also be extended to other intentional states. Besides belief, the second and equally fundamental intentional state is desire. To accommodate desire one must posit something like an experiential feature of 'taking to be good.' Perhaps this is intuitively even more acceptable than is 'taking for true.' We have many expressions that seem almost directly to refer to such a feature of experience, for example, when we say of something that it 'looks good to eat.'

With an account of introspective knowledge of belief and desire in place it is not difficult to further extend the theory to other intentional states and the highly significant subclass of states – involving both cognitive and sensory elements – that comprise emotional consciousness.

Apart from the worry that the displaced perception account makes introspection implausibly similar to inferential knowledge, another serious worry is that it mistakenly equates introspection with a kind of self-generation of our mental states rather than a kind of access to preexisting mental states. Roughly, but hopefully vividly, the problem is that this account cannot distinguish between introspective knowledge of a belief and the (presumably also introspective) knowledge that one has just acquired a belief. In fact, there is a danger that displaced perception could yield a false introspective claim if it is taken to generate knowledge about what one's beliefs were, prior to setting the displaced perception mechanism into action. It is not entirely clear that this is a bug rather than a feature. After all, if a child comes to disbelieve in Santa Claus simply by reflecting on the implausibilities inherent in the usual stories, then it seems quite correct for introspection to generate the claim that the child does not believe in Santa Claus.

But there is an interesting issue here about whether introspection should be regarded more as a kind of 'measuring instrument' revealing the current state of one's mind or more as an engaged

faculty, which is at least in part actively forming the mind according to principles of rationality as well as, perhaps, other normative principles. The notion that introspection plays a role in self-construction is attractive (and has obvious affinities with the transcendental approach), but does not seem to be at odds with the positive accounts of introspective mechanisms discussed above.

Conclusion

The topic of introspection, self-awareness, and self-knowledge is complex and far from settled. The nature and status of introspective knowledge remains hotly debated and the issues raised in discussions of introspection are closely linked to the wider philosophical topics of mind, naturalization, and consciousness. In addition, the recent and accelerating growth in our understanding of the neural processes underlying introspection is likely to have a huge impact on the philosophical accounts and general attitudes toward our knowledge of ourselves.

See also: Phenomenology of Consciousness; Self: Body Awareness and Self-Awareness; Self: Personal Identity; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Abilgrami A (1995) Belief and Meaning. Oxford: Blackwell.
Aydede M (2003) Is introspection inferential? In: Gertler B (ed.) Privileged Access: Philosophical Accounts of Self-Knowledge. Burlington: Ashgate.
Byrne A (forthcoming) Introspection. Philosophical Topics.
Carruthers P (1996) Simulation and self-knowledge: A defense of theory. In: Carruthers P and Smith P (eds.) Theories of Theories of Mind. Cambridge: Cambridge University Press.
Davidson D (1987) Knowing one s own mind. Proceedings and Addresses of the American Philosophical Association 60: 441 458.
Dretske F (1995) Naturalizing the Mind. Cambridge, MA: MIT Press.
Dunning D (2005) Self-Insight: Roadblocks and Detours on the Path to Knowing Thyself. New York: Psychology Press.
Gertler B (ed.) Privileged Access: Philosophical Accounts of Self-Knowledge. Burlington: Ashgate.
Gertler B (2003) Self-knowledge. In: Zalta EN (ed.) The Stanford Encyclopedia of Philosophy. Stanford, CA: Metaphysics Research Lab. http://plato.stanford.edu/archives/spr2003/entries/self-knowledge/.
Gertler B (2008) Do we look outward to determine what we believe? In: Hatzimoysis A (ed.) Self-Knowledge. Oxford: Oxford University Press.
Goldman A (2006) Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading. Oxford: Oxford University Press.
Gopnik A (2001) Theory of mind. In: Wilson R and Keil F (eds.) The MIT Encyclopedia of the Cognitive Sciences. Cambridge, MA: MIT Press.
Kind A (2006) Introspection. The Internet Encyclopedia of Philosophy. http://www.iep.utm.edu/i/introspe.htm.
Lycan W (1997) Consciousness as internal monitoring. In: Block N, Flanagan O, and Güzeldere G (eds.) The Nature of Consciousness. Philosophical Debates, pp. 755 771. Cambridge, MA: MIT Press.
Lyons W (1986) The Disappearance of Introspection. Cambridge, MA: MIT Press.
McLaughlin B and Rorty A (1988) Perspectives on Self-Deception. Berkeley: University of California Press.
Moran R (2001) Authority and Estrangement: An Essay on Self-Knowledge. Princeton, NJ: Princeton University Press.
Nichols S and Stich S (2003) Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds. Oxford: Oxford University Press.
Ryle G (1949) The Concept of Mind. London: Hutchinson & Co.
Schwitzgebel E (2008) The unreliability of naïve introspection. Philosophical Review 117: 245 273.
Seager W (2002) Emotional introspection. Consciousness and Cognition 11(4): 666 687.
Shoemaker S (1996) The First-Person Perspective and Other Essays. Cambridge: Cambridge University Press.
Tye M (1995) Ten Problems of Consciousness. Cambridge, MA: MIT Press.
Wegner D (2002) The Illusion of Conscious Will. Cambridge, MA: MIT Press.
Wilson T (2004) Strangers to Ourselves: Discovering the Adaptive Unconscious. Cambridge, MA: Belknap.

Biographical Sketch

William Seager is a professor of philosophy at the University of Toronto. He was born and raised in Edmonton, Alberta. He obtained his BA and MA from the University of Alberta and then moved to Toronto for his PhD work, where he has remained ever since. His main research interests are in the philosophy of mind, especially the problem of consciousness. He has published *Theories of Consciousness* (Routledge, 1999) and many articles on mind and consciousness, the most recent being 'The intrinsic nature argument for panpsychism' (in A. Freeman (ed.) *Consciousness and Its Place in Nature*, Imprint 2007). William Seager would especially like to thank Matt Habermehl and Adrienne Prettyman, as well as a number of assiduous anonymous referees, for very helpful comments on this article.

The Placebo Response

L Y Atlas and T D Wager, Columbia University, New York, NY, USA

Glossary

Analgesia – Significant lessening or complete termination of pain sensation.
Nociception – Pain sensation.
Placebo – An administered treatment that on its own has no known beneficial effect on a given condition.
Placebo effects – Observed improvements in outcome measures attributed to the influence of the placebo treatment.
Placebo response – The set of endogenous processes whereby conscious expectancies and conditioning recruit brain and body mechanisms to elicit beneficial effects on symptomatology and processing.

Introduction

The history of placebo effects over the last century has been a turbulent one. They have been variously dismissed as artifacts and embraced as compelling evidence of the healing power of the mind. In either case, the rigorous scientific study of placebo provides a unique window into the relationship between the mind, brain, and body. Such research is beginning to reveal the potential power of, and the limits on, how the brain affects the physiological and psychological manifestations of health and disease. In some cases, positive outcomes have been demonstrated to result from conscious expectations of therapeutic benefit. But some of the most potent placebo responses have been elicited by learning that may be outside of conscious awareness and at least partially impenetrable to volition and conscious thought. These studies, and more broadly the kinds of evidence that have and have not demonstrated the power of placebos, are the subject of this article.

A placebo treatment is one that is expected to have no direct physical or pharmacological benefit – for example, a starch capsule given for anxiety or pain, or a surgery where the critical surgical procedure is not performed. For this reason, placebos are routinely used as comparison conditions in clinical studies, against which to evaluate the effects of investigational treatments. However, placebo treatments have also frequently been used to actually treat a variety of ailments; they have had a place in the healer's repertoire for thousands of years, and are used as clinical treatments by physicians in industrialized countries today with surprising frequency.

To the degree that placebo treatments are healing agents, their power lies in the psychobiological context surrounding treatment, resulting in an active response in the brain and body of the patient. This endogenous response is referred to as a placebo response, or meaning response. In many cases, the psychological meaning of the treatment induces shifts in cognition, emotion, and corresponding brain and nervous system activity that produce a palliative effect. In other cases, placebo responses result from associations learned in the brain through a process known as conditioning. In this process, associations between the elements of the treatment context (the shape of a hypodermic needle, or the color of a pill) and helpful neurobiological responses are formed in the brain. These associations may be learned and activated outside of the patient's consciousness. Thus, although a placebo treatment is itself inert, the placebo response on the part of a patient may have real healing benefits. What is typically measured in a study, however, are placebo effects – observed improvements in signs or symptoms attributed to the influence of the placebo treatment. Although the two terms are often used interchangeably, for the purpose of this article we distinguish between the placebo effect, an observed effect on specific outcome measures (e.g., the difference between a placebo-treated group and an untreated group in a study), and the placebo

response, which is a set of endogenous processes (brain and body mechanisms) recruited by placebo-induced expectations and conditioning. By this definition, placebo effects are directly observed in studies, and placebo responses are psychophysiological processes whose nature must be inferred by observing placebo effects. A placebo effect may thus provide evidence for an underlying placebo response, though the strength of that evidence depends on the nature of the comparison performed in the study and range of alternative explanations. (We note that many scholars use the term 'placebo response' to refer to any improvement on a placebo treatment, which is not the sense in which we use the term here.)

The potential clinical significance of the placebo response has led to the standard use of placebo groups in clinical trials for therapeutic drugs and procedures. Clinical trials are experimental studies performed to test the therapeutic efficacy of medicines, surgical procedures, and other interventional procedures. The use of placebos has proven invaluable in providing a baseline against which to assess interventions; however, clinical trials are not typically designed to test the therapeutic effects of placebo treatments themselves. Thus, for reasons described below, they provide little evidence either for or against the existence of placebo responses as defined above.

In an influential paper published in 1955 entitled 'The powerful placebo,' Henry K. Beecher analyzed a series of placebo-controlled clinical trials in order to examine the efficacy of the placebo itself. Stemming from this first effort, another class of experimental studies is designed to specifically assess the therapeutic effects of placebo treatment and patients' expectancies, and these studies provide the most direct evidence for active psychobiological placebo responses. These studies offer scientists a unique opportunity to investigate interactions between the brain and body. Behavioral, pharmacological, physiological, and most recently neuroimaging methodologies have allowed us to begin to understand the mechanisms by which placebos exert their effects. This, in turn, offers scientists a window into understanding how the brain exerts control over behavior, emotional experience, and physiological processes in the body. This overview is devoted to a critical review of these studies.

False Placebo Effects

Placebo effects have been the center of heated debates among researchers that have persisted in different forms throughout the past century. The central debate revolves around the issue of whether active, psychobiological placebo responses exist, and whether they affect health and disease processes in meaningful ways. Two alternative explanations for observed placebo effects have been offered: First, placebo effects in some studies may be statistical artifacts; and second, placebo effects in many studies may reflect changes in subjective reporting processes only, and not in meaningful disease processes.

The argument that the so-called 'placebo effects' are statistical artifacts applies primarily to clinical trials that are not designed to test the efficacy of placebo treatments, and are thus not carefully designed to avoid such artifacts. The argument that observed placebo effects are subjective reporting biases applies to clinical trials and experimental studies of placebo alike. In this section, we outline each argument, and in the following section, we evaluate the experimental evidence on reporting biases in various disease processes.

Placebo Effects as Statistical Artifacts

In a typical clinical trial, patients are randomly assigned to treatment either with a therapeutic intervention (e.g., a study drug) or a placebo, and outcomes are assessed under 'double-blind' conditions: neither the patient nor the assessor knows which treatment a person is taking. There are many variants on this procedure, but the vast majority of trials share this element in common. Comparisons between the treatment and placebo groups are performed in order to estimate the active effects attributable to the treatment (treatment efficacy). Two assumptions are made in this comparison: First, that all nonspecific effects of being in the study (natural outcome improvement or worsening over time, effects of patient expectation and motivation, health care setting effects, etc.) are common to both treatment and placebo groups, and second, that nonspecific effects and drug effects combine additively, so that a simple subtraction between the two will yield the active

treatment effects. The critical question for placebo research is what those 'nonspecific' effects are, and how much of them can be attributed to causal effects of the placebo treatment itself (i.e., the placebo response). Put in concrete terms, placebo groups in studies of depression typically improve about 7–8 points on the Hamilton depression inventory (a clinical measure with a range from 0 to 54 points, with 24 or above indicating severe depression); but it would be an error to attribute all of this effect to the placebo response, because other factors also contribute to 'nonspecific' improvement in placebo groups. Some of these are (1) spontaneous remission, (2) natural symptom fluctuation, (3) regression to the mean, and (4) participant sampling bias.

Natural history: spontaneous remission and natural symptom fluctuation

Every disease has a natural history, a time-course that the disease state would follow without any intervention. Many disease states are time-limited, and patients experience spontaneous remission – most depressed patients, for example, eventually recover, as do those with anxiety disorders, sleep disorders, and many other conditions. Improvement in a placebo group in a clinical trial could therefore be due to spontaneous remission. Other diseases – such as hypertension, pain syndromes, Parkinson's disease, irritable bowel syndrome, and many others – do not have high spontaneous remission rates, but the signs and symptoms fluctuate over time. A study may end while a patient is relatively symptom-free, which would look like healing attributable to placebo, but may in fact be a transient improvement after which symptoms may reappear.

Regression to the mean

Regression to the mean refers to the observation that when measurements are repeated, subsequent measurements are likely to be closer to the mean than the initial ones. This phenomenon is especially applicable to clinical improvements in placebo groups. If a patient seeks treatment and is enrolled in the study when his or her disease is worse than average, the patient's state is likely to improve by the time the disease is next measured due to the natural course of the condition, rather than anything having to do with placebo administration. It is tempting in many contexts to look at subgroups of individuals on the basis of their initial disease state – that is, the most depressed patients – and follow them over time. But this is problematic due to regression to the mean: If one were to follow the most severe patients in the placebo group in almost any disease state, they would be seen to improve over time.

These first three factors demonstrate the importance of including a natural history (non-treatment) control group in clinical trials for assessing placebo effects. If a no-treatment group is included, then a comparison between the placebo-treated and the no-treatment group can be used to assess the effects of the placebo *per se*. Unfortunately, very few clinical trials include such groups, because the placebo effect itself is not of primary interest in these studies.

Participant sampling bias

Participant sampling bias comes about as a result of the fact that participants who experience beneficial results from a treatment are more likely to continue in the study and adhere to treatment regimens than those who experience either no effect or adverse effects. Thus, the participants who complete clinical trials in placebo groups (as well as active treatment groups) are more likely to be those who improve over the course of the study. The result is that participants may appear to improve over the course of the study, but this apparent improvement actually reflects changes in the sample over time.

Placebo Effects as Reporting Biases

Another important source of potential error has to do largely with the important subjective aspects of many conditions in which the placebo effect is studied. In pain, depression, and nearly every illness, an essential component of the illness is the subjective experience of the condition.

In some domains, there are clear objective measures of current state, such as motor performance or outwardly observable physical symptoms, but in many important clinical states objective measures are not available. For example, because pain is a subjective experience, clinical and experimental measures of pain are based on patients' reports.

Physiological measures based on skin conductance, heart function, and pupillary response can be measured, but they are indirectly related to pain; thus, while they may be more precisely measured in some cases, they cannot be assumed to accurately reflect the pain experience.

The problem with self-report-based measures is that they might be influenced by a number of factors that are not central to the disease process being studied. Thus, demonstrating placebo effects in subjective outcomes may say little about the power of placebo treatments to effect meaningful changes in disease progression. For example, a treatment that improves subjective well-being in cancer patients may be beneficial for this reason alone, but its viability as a specific treatment for cancer must certainly rest on its effects on the growth of the cancer. A few of the factors that create biases in subjective outcome assessment are described below. Ruling out reporting biases as causes of observed placebo effects is difficult in many cases; we address relevant evidence in the following sections.

Demand characteristics and Hawthorne effects

Demand characteristics refer to changes in patients' reports and behavior on the basis of their perception of how they are expected to behave. Often, these expectations are communicated by inadvertent cues from the investigator: Even a very subtle nod or widening of the eyes from a physician may communicate agreement or surprise on the physician's part. Such influences can be avoided by keeping experimenters and physicians blind to condition, when possible. In one interesting study, for example, physicians' expectations about a drug's effectiveness were shown to affect how effective patients thought it was in relieving pain. Other types of cues about expectations might result from the study procedures themselves or the way questions are worded. The question ''How much did your pain decrease?'' implies that it should have decreased to some degree.

Being observed often causes changes in behavior. Such changes are classically referred to as Hawthorne effects, after a landmark study that showed workers' performance improved dramatically once they were under observation in the study. Hawthorne effects generally result from social influences that affect participants' reports and behavior. Male experimental participants, for example, will tolerate higher levels of painful stimulation when an experimenter is watching them.

Demand characteristics and reporting biases cover a range of different psychological effects, including Hawthorne effects in some cases, social compliance effects (in which patients say what they feel should be said), self-presentation biases (individuals often say what makes them look better in the eyes of others), and self-consistency biases (consistency with past behavior, to avoid admission that past behavior was in some way incorrect).

Response bias

Many experiments have demonstrated that decisions about the presence or absence and intensity of a stimulus (such as a disease symptom) are not a function of a stimulus alone, but of prior beliefs and the relative benefits and costs of the decision. This approach is the basis of signal detection theory (SDT). Consider the following simplified scenario. A patient is given a medication and is asked to judge whether the drug relieved pain. In simplified form, the patient truly feels different or she does not, and she may choose to respond ''Yes, the drug helped'' or ''No, it did not help''. A 'yes' response might be true pain relief or a false affirmative (an error). A 'no' response might reflect a true lack of pain relief or a false negative (another kind of error). The answer that the patient reports depends partly on the how well the patient can discriminate true from null effects, and partly on the relative costs and benefits of making the two kinds of error. The SDT framework acknowledges that a cost–benefit analysis influences patients' decisions on what to report, and provides a way of estimating both the discriminability of true versus null effects and the patient's response bias towards responding either 'yes' or 'no.'

SDT analysis has been employed since the late 1960s in order to investigate whether placebo administration alters sensitivity to pain (measured by the ability to discriminate between different levels of noxious stimulation), response bias, or both. These studies suggest that the placebo effect is actually a change in response bias alone, without any change in sensitivity. A complicating factor is

that a change in 'response bias' after placebo treatment amounts to subjects reporting that noxious stimulation simply feels less painful after the treatment, even if they are no less accurate at discriminating different levels of intensity. These studies cannot provide evidence on whether those reports of feeling less pain occur because of changes in pain processing in the brain and spinal cord, or simply because of a cost–benefit analysis involved in the reporting decision. One way of disentangling these alternatives lies in the direct measurement of brain responses to painful stimuli, and we return to this alternative below.

Active Placebo Outcomes

It is clear that placebo treatments may influence subjective reports and observable conditions without affecting any underlying disease process. That is not to say that placebo effects on some outcomes, such as reported relief from pain or improvement in reported quality of life, are not valuable in their own right: In many cases, these outcomes are central to the comfort and happiness of patients, and to their ability to work productively and maintain positive social relationships. Nonetheless, it is valuable to review evidence for placebo effects on measurable biological processes in the brain and body. These are most often likely to be essential elements of a clinically meaningful placebo response in disease progression.

In this section, we review some of the evidence for placebo effects on physiological outcomes, focusing on physical pain – one of the best-studied outcomes experimentally – as a model system. In brief, placebo effects have been reported on physiological indices of nociceptive (pain-related) processing, including measures of brain activity and neurochemical responses to painful stimuli, and placebo analgesic treatments have been shown to interact with active pain-relieving drugs. In clinical studies, there is a growing body of literature demonstrating placebo effects on reported clinical pain (a clinically meaningful outcome) in pain disorders such as irritable bowel syndrome.

In other domains, particularly when conditioning is involved and sensory or internal cues become associated with active treatment, experimental studies have found placebo effects in immune function, cortisol levels, and growth hormone levels. In Parkinson's disease, placebo effects have been reported in disease-relevant neural and neurochemical activity and in clinical signs of disease severity. In major depressive disorder, treatment with placebo induces changes in brain activity that mimic treatment with an active drug. Finally, in asthma, placebo effects have been reported in clinical measures of airway responsiveness. It is possible that placebo effects exist in other physiological processes; more research remains to be done to demonstrate the causal effects of placebo treatment in these and other areas.

Placebo Effects in Pain Physiology

While placebo effects have been demonstrated in a wide range of domains, the majority of laboratory studies demonstrating causal effects of placebo treatment have been conducted in the context of pain. Pain is a common and debilitating condition with great biological significance for all organisms. Though it typically has a peripheral, physiological origin and physiological consequences in the body, pain is a subjective experience thought to be determined by an interplay between sensory, affective, and emotional brain systems. Because pain can be manipulated experimentally, and because much is known about the neural and physiological correlates of human and animal nociceptive processing, pain is uniquely suited for experimental investigations of the placebo response at multiple levels, from the involvement of specific endorphins and neurotransmitters, to the contributions of different brain regions, to the motivational significance of placebo effects for an individual.

Because of the clinical significance of pain experience, the vast majority of placebo studies to date have demonstrated effects on subjective experiences of analgesia or pain relief. The classic example of placebo analgesia is when a person is (1) experiencing pain, which may be ongoing or caused by noxious (normally painful) stimulation in a laboratory study; (2) the person is given a placebo treatment, often with information indicating that the treatment will diminish pain; and (3) pain under the placebo treatment is compared with pain in an otherwise comparable no-treatment

condition, either in the same or different individuals, and pain reports decrease reliably with the placebo. However, in spite of the predominant (and clinically appropriate) use of reported pain as a primary outcome, several lines of research demonstrate placebo effects in objective physiological outcomes.

Early studies of placebo analgesia, in the late 1960s, suggested that placebo treatment affected pain reports, but not the ability to discriminate stimuli of different temperatures from one another. This research was done in the tradition of then-recently developed SDT, which provided a method to separately measure an observer's ability to discriminate stimuli of two or more categories and the so-called 'response bias' in reporting on the basis of costs and benefits of making different types of errors. Because placebo effects were found in 'response bias,' but not discrimination ability, it was concluded that placebo treatments may amount to nothing more than an increase in the decision criterion for reporting a given stimulus as painful.

However, another set of studies published in the late 1970s provided partial refutation of this conclusion by providing evidence that placebo treatments resulted in the release of brain opioids – endogenous (produced by the brain) neuropeptides known to play a key role in clinical pain relief. In these studies, the researchers gave participants a placebo treatment, which produced analgesia; but when they gave the same placebo treatment along with naloxone, a drug that blocks opioid receptors, the placebo analgesic effect was eliminated. These findings suggested that endogenous opioid release was a necessary component of at least one type of placebo analgesia (that elicited by verbal instructions and consequent expectancies of pain relief). Because opiate drugs are among the most widely used and best-known treatments for clinical pain and because opioids are involved in inhibition of pain-related neural signals at the earliest stages of processing in the spinal cord, the broader implication was that placebo treatment must have affected the brain physiology of pain.

These studies were replicated and extended in the late 1990s and early 2000s. Although naloxone can itself block pain or enhance pain depending on the dose and perhaps on the psychological conditions of the study, several studies demonstrated that naloxone could reverse placebo analgesia without otherwise affecting reported pain, and that placebo analgesic effects were specific to the particular body site to which they were applied. This latter finding implied that placebo analgesia in these paradigms was not a general, 'global' response to the placebo treatment, but rather was mediated specifically by expectancies that pain would be reduced at a particular body site. Another important set of findings was that placebo effects could be driven by expectations or conditioning with an opiate drug, in which case they tended to be naloxone-reversible and thus implicate endogenous opioid release, or they could be created by conditioning with nonopiate analgesics, in which case they do not appear to be naloxone reversible, implicating nonopioid mechanisms.

While the inferences from these naloxone studies suggest that active endogenous opioid processes play an important role in the placebo response, these indirect inferences fail to fully resolve the question of whether pain processing itself is affected under placebo. Opioids could affect pain reports in nonspecific ways or offset the aversiveness of pain without affecting specific nociceptive processing. In order to resolve whether nociceptive processing is affected, changes in the brain processes underlying pain must be examined.

Neuroimaging methodologies, including positron emission tomography (PET) and functional magnetic resonance imaging (fMRI), have allowed researchers to examine the brain's activity in response to pain and to identify a network of regions involved in the pain response, frequently referred to as the 'pain matrix.' The first study to use neuroimaging to examine the placebo response used PET to identify a common network between placebo analgesia and opioid analgesia. When compared to a condition in which participants received noxious stimulation without any analgesic administration, both placebo and opioid conditions revealed increased activity in rostral anterior cingulate (rACC) and the brainstem, two regions that are known to play important roles in pain modulation.

A second approach to the study of the placebo response assumes that if the placebo response truly decreases pain processing, this ought to be evident through decreased pain-related activity in regions of the brain known to be responsive to pain. Thus,

researchers can contrast the response to stimulation under placebo with stimulation in a nonplacebo control condition and look for differences in pain-related neural activity. The first study to use such an approach used fMRI, and did indeed find evidence of placebo-induced decreases in regions of the pain matrix that have been shown to be associated with the affective components of pain, including rACC and insula, as well as thalamus. In addition to these decreases, placebo administration induced increases during pain anticipation in orbitofrontal cortex (OFC), lateral prefrontal cortex (PFC), and, importantly, an area of midbrain surrounding the periacqueductal gray (PAG), a structure rich in endogenous opioids that is known to play an important role in pain modulation. Decreases in thalamus and insula, however, occurred late in the pain response, near the time when participants were asked to make ratings of the pain they had experienced; this therefore did not fully resolve whether the pain process itself had changed or whether the observed changes had been mediated by evaluation processes.

Researchers followed this initial investigation with studies employing methodologies with improved temporal resolution, including electroencephalography (EEG) and magnetoencephalography (MEG). Just as research using PET and fMRI has identified specific regions involved in the pain experience, EEG research has identified specific temporal components that are known to be associated with pain. Thus, a similar approach can be employed in EEG and MEG research to examine whether the placebo response changes pain processing; researchers can look for placebo-induced decreases in the amplitude of evoked nociceptive event-related potential components. Importantly, the temporal resolution of these methodologies allows researchers to specifically examine pain-related components that follow nociceptive stimulation quickly enough to be unaffected by subsequent modulatory cognitive processes (such as poststimulus decision making). This approach is particularly suited to examine a long-standing hypothesis known as the gate control theory, which hypothesizes that descending mechanisms inhibit nociceptive input at the level of the spinal cord, thereby preventing nociceptive information from ascending to the cortex for processing.

Several studies have employed this approach, and all have offered support for at least some decrease of pain processing as part of the placebo response. One study examined the effect of cues on perceived pain, and found that invalid cues (those inducing expectations for lower or higher pain than applied stimulation) modulated activity in components associated with activity in secondary somatosensory cortex (SII) that occurred just 160 ms after pain stimulation. However, later components reflecting activity in posterior ACC varied with stimulus intensity, but were not affected by induced expectations and changes in perceived pain. A study that examined two of the major components of laser-evoked pain potentials, the P2 and N2, found that the amplitude of the P2 response was decreased with placebo treatment. Importantly, the P2 differences habituated over the course of the experiment, although placebo differences in reported pain persisted, suggesting that effects on early nociceptive processing and reported pain are dissociable. Furthermore, the extent of P2 decreases was less than that which would be expected if the gate control theory entirely explained changes in pain processing under placebo. These results suggest that while the placebo response does directly affect nociceptive processing, cognitive components are also important in the placebo response, as indicated by the large and persistent placebo effects on reported pain, relative to the smaller effects on P2 amplitude, which habituated over time.

Since these initial findings, recent neuroimaging studies have added to our understanding of how the placebo response affects neural processing. Similar to the rationale behind the high temporal resolution studies reviewed above, one study sought to resolve whether placebo-induced decreases occurred early in the pain period. Rather than using a methodology with high temporal resolution, the authors used fMRI but acquired images only during nociceptive stimulation, and stopped the scanner during the pain rating periods. Decreases were observed in the right mid-insula, medial thalamus, and ACC during painful stimulation, again suggesting that the placebo response can indeed decrease nociceptive processing in the brain.

Together, these studies provide evidence that the placebo response modulates pain processing,

and begin to provide evidence of how this modulation may occur. Consideration of placebo responses in other domains allows us to consider candidate central and proximal mechanisms for these observed effects.

Physiological placebo effects in other domains

A controversial meta-analysis of clinical trials that compared placebo groups with no-treatment control groups found evidence of beneficial placebo effects only in subjective outcomes, particularly in studies of pain. The authors asserted that this was presumably a result of reporting biases and artifacts. The series of neuroimaging studies of placebo analgesia reviewed above clearly demonstrate that the placebo response does indeed influence physiological processing of nociceptive information. However, the subjective component of pain may play a critical factor in the etiology of the placebo response.

In order to determine the kind of effect the placebo response may have on outcomes without any subjective component, a meta-analysis was conducted that examined placebo effects in clinical trials of peripheral disease processes that used outcome measures assessing the state of peripheral organs, tissues, and body fluids was conducted. There was a significant positive effect of placebo administration on physical processes, such as blood pressure and forced expiratory volume (a clinical sign of asthma severity), but no consistent placebo effect was demonstrated in biochemical process measures, such as cortisol and cholesterol levels.

Thus, placebo effects appear to be strongest in physiological processes that can be directly controlled by the brain, and their effects on biochemical processes in the body are likely to be substantially weaker. However, these findings do not conclusively show that there are no effects on biochemical processes. First, some of the individual clinical trials included in the meta-analysis may indeed have shown support for placebo effects on biochemical processes, but there may have been significant heterogeneity among studies, causing the overall results of the analysis to be null. Second, a limitation of analyses of clinical trials is that they are not designed to study placebo effects, and expectancy effects in these studies may in some cases be very weak. Laboratory studies use instructions and conditioning to elicit positive expectations (i.e., 'This cream is known to be effective in relieving pain' for the placebo, 'This is a control cream with no known effect' for the control), whereas in double-blind clinical trials, participants are told that they may receive either an active agent or an inert substance. Consistent with this notion, another meta-analysis has shown that laboratory studies of placebo mechanisms elicit stronger placebo effects than clinical studies in which placebo groups are used as a control.

Experimental investigations have demonstrated that placebo treatments can alter peripheral physiology across a variety of conditions. In asthma, placebo bronchodilator administration increases forced expiratory volume, a clinical measure of lung function. Preconditioning with active agents has been shown to elicit powerful placebo effects in the domains of growth hormone and cortisol secretion, as well as immunosuppression. Placebo analgesia not only affects pain-related physiology in the brain but has also been shown to reduce b-adrenergic activity in the heart. Experiments designed to investigate nocebo effects – whereby expectations for increased pain or negative outcomes lead to hypersensitivity or increased symptomatology – reveal that nocebo responses in pain involve hyperactivity of the hypothalamic–pituitary–adrenal axis, as evidenced by increased cortisol release and adrenocorticotropic hormone.

Mechanisms of Placebo Effects

The evidence reviewed above powerfully supports the existence of a placebo response, albeit one that varies in magnitude across outcome measures and disease processes, and with the strength of expectancy manipulations and use of conditioning procedures. However, a mechanistic understanding of the placebo response is critical to understanding how, in what outcomes, and under what conditions placebo responses occur. In summary, the placebo response can be divided to two mechanistic levels: central mechanisms and proximal mechanisms. Central mechanisms are those that may operate across domains (i.e., in pain, Parkinson's disease, and depression), whereas proximal mechanisms

tend to be domain-specific. We will again employ pain as a model system, since much is known about the central and peripheral pathways of pain processing.

Central Mechanisms

Central mechanisms are those involved in translating verbal instructions into positive expectancies and maintaining the activity corresponding to those expectancies in the brain. They are the mechanisms of generating and maintaining positive beliefs, which are likely to operate in similar ways across disorders. Conditioned placebo effects may involve additional, separate mechanisms. Studying placebo effects can thus provide a window into how these basic processes work, and how they shape mind–body interactions across a range of conditions.

Expectancy versus conditioning: two routes

An important theoretical debate about the central mechanisms subserving the placebo effect seeks to understand the psychological processes that give rise to the placebo response. Do conscious expectations mediate the placebo effect, or are changes in placebo due to classical conditioning? We will review the arguments of each perspective, noting that they are not mutually exclusive; both general expectancies and specific learning (conditioning) are likely to play important roles in placebo effects.

Conditioning and the placebo effect

Some researchers have argued that placebo effects are a result of conditioning. In the most common understanding of classical conditioning, an unconditioned stimulus (US) that normally elicits a certain response (the unconditioned response, or UCR) is paired with a neutral stimulus that elicits no response on its own. Over repeated pairings of the two stimuli, the previously neutral stimulus comes to elicit the same response as the US; the previously neutral stimulus is then referred to as a conditioned stimulus (CS), and the evoked response is a conditioned response (CR). Conditioning can occur in aversive contexts (in fear conditioning, a light may be paired with a shock, to elicit freezing in response to the light) or appetitive contexts (as in Pavlov's classic experiments, food can be paired with a tone and animals eventually salivate in response to the tone).

In the case of the placebo effect, there are several routes by which conditioning may result in placebo effects. For example, a pharmacological agent (US) might be administered in the form of a pill (CS) to elicit the specific effects of the drug (UCR); subsequent administration of the pill without the active pharmacological ingredients might still elicit the same effects. Researchers have even suggested that the medical context itself can serve as the conditioned stimulus; thus, people may associate intrinsically neutral items such as syringes or even doctors with the changes associated with treatment, and placebo effects may result from conditioning to these contextual stimuli. Proponents of the classical conditioning model of placebo have argued that a lifetime of medical treatments serve as conditioning trials to pair the medical context (CS) with therapeutic effects (CR). Support for the classical conditioning model of placebo comes from research demonstrating that placebo effects are stronger after exposure to active drugs, that placebo analgesics are more effective when labeled with well-known brands, and from research that shows that some side effects of drugs that are unlikely to be consciously perceived or mimicked – such as respiratory depression following opiate treatment – may be reproduced by placebo treatments after conditioning. In addition, some placebo effects, such as increased cortisol and growth hormone after conditioning with an active drug, are not reversible by telling subjects that the medication is a placebo, suggesting that some learning has occurred that is not modifiable by conscious expectancies.

Brain mechanisms of conditioning

While conditioning has been studied for more than half a century, and we know much about the neural circuitry involved in fear conditioning, mechanistic research that would directly support a conditioning model of placebo is quite scarce. Primary support for a model in which conditioning recruits endogenous mechanisms comes from studies of conditioned immunosuppression. In a series of studies in rats, saccharin was paired with an

immunosuppressive agent, and the rats that had been preconditioned exhibited decreased antibodies when given saccharin in a later test phase. A similar approach has been used to show evidence of conditioned immunosuppression in humans, measured by immune factor expression in mRNA and lymphocytes (white blood cells). These studies offer support for conditioning-based placebo effects on immune responses.

However, while this offers some insight into the role of conditioning in endogenous processes, little is known about brain mechanisms specific to conditioning-based placebo responses in humans, and mechanisms supporting conditioned immunosuppression are unlikely to generalize to other domains, such as pain. In one PET study mentioned earlier, researchers compared brain responses to painful stimulation under opioid-based analgesia with responses to placebo analgesia. Opioid administration always preceded the placebo analgesia condition, which may have induced a conditioning-based placebo effect. Brain responses to each were compared with a pain control condition, and both were associated with increased activity in rACC and increased rACC-brainstem connectivity. While promising, this and other studies have not directly compared conditioning processes with nonconditioning expectancy manipulations (verbal instructions only), and the nature of the conditioning-specific placebo response remains yet to be elucidated.

Expectancy and the placebo effect

An alternative view proposes that conscious expectancies mediate the changes associated with placebo effects. In this view, the internal beliefs and expectations associated with the inert treatment are responsible for the endogenous regulation of processes in order to produce the requisite changes associated with placebo response; put simply, one experiences changes associated with placebo administration because one expects to.

Expectancies involve appraisals of the significance of a stimulus or event in the context of its anticipated outcome. For the most part, expectancies are conscious at the time when decisions are made, or, if they are not conscious (as may happen during rapid decision making), expectancies can be brought into consciousness when attention is drawn to them. This is an important distinction from conditioning theories, which assume that organisms need not have conscious awareness of contingencies between stimuli in order for conditioning to occur. Thus, one way to define the distinction between expectancy effects and conditioned responses is that expectancy effects depend on the participant's state of mind, whereas conditioned responses do not. By this definition, expectancy-based effects can be altered by instructions to subjects, whereas conditioned effects cannot.

The power of expectancies has been illustrated in many areas of research, from basic perception to complex physiological processes. One way that researchers have learned about the influence of expectations on physiological processes is by comparing the effects of hidden and open drug administration. Positive expectancies (expectations for relief) are active when patients are aware that they are receiving a given drug, as is the case when drugs are administered in full view of the patient. When this is contrasted with conditions in which drugs are administered surreptitiously (e.g., when a drug is administered intravenously under the guise of saline administration), researchers generally find that the open administration has a greater beneficial effect. When the contextual cues surrounding treatment are the same in both open and hidden cases, a conditioning-based explanation for the effect is unlikely. A practical application of this research is evident in hospitals, where patients are allowed to self-administer analgesic agents such as morphine; it takes far less morphine to produce the same pain-relieving effect when patients control drug delivery and expect relief than when doctors administer the drug without patient expectations.

The same positive expectancies may therefore be the driving force behind the power of placebo. The expectancy model of placebo is supported by research demonstrating that placebo effects can occur with verbal instruction alone (i.e., without prior experience with a drug or active treatment). In addition, in some studies, placebo effects on pain that have been induced through a conditioning procedure have been reversed completely by revealing to the participants that the placebo treatment was a sham. Thus, the placebo effects in these studies do not meet the criteria for conditioned responses (involving specific learning not modifiable by beliefs).

It is also worth noting that for conditioning to occur, the brain must be capable of learning an association (i.e., forming a pathway) between a CS and either the UCS or the UCR, and reactivation of that pathway must be able to elicit the UCR. Placebo responses do not always fit these criteria. For example, during the extinction phase of conditioning, a conditioned stimulus that is not reinforced will cease to elicit a conditioned response; however, placebo effects can remain far longer than extinction would permit. In other cases, an association formed over repeated experiences can be reversed immediately by a change in instructions, which is not consistent with models of conditioning.

Brain mechanisms of expectancy

Placebo instructions change the cognitive context in which pain stimulation is perceived, and these altered appraisals of the situation give rise to changes in expectations about pain, harm, and pain relief. The brain mechanisms of such expectations are likely to be similar to those involved in executive functions – basic cognitive processes coordinating the maintenance and manipulation of information. Information about context is known to require dorsolateral prefrontal cortex (DLPFC), which interacts with working memory – systems for maintaining information in an active state in the brain – in order to maintain expectancies induced by placebo manipulations. Expectations about the value of upcoming stimuli are also critical to the pain process and are potentially highly involved in the placebo response across domains. Effective placebo administration induces expectations for reduced symptomatology or diminished pain, which affects how the brain processes the condition or stimulus. The processes most likely to be altered are those that assign value and meaning (for the self, or survival) to the stimulus. Orbitofrontal cortices and rACC have been shown to be highly involved in the process of valuation. Brain representations of active contextual information and stimulus value may directly or indirectly influence more basic perceptual, behavioral, and somatic (peripheral) processes through connections with other parts of the brain that represent percepts, motor actions, and somatic states.

By drawing on knowledge from brain mechanisms of executive function, researchers can define reasonable hypotheses about brain mechanisms supporting an expectancy-based placebo response. These can be tested by contrasting anticipatory activity in a placebo condition to anticipatory activity during a control condition, so that one can identify processes related to pain expectancy that are shaped by placebo treatment. This approach was used in an fMRI study of placebo analgesia, which revealed increases in DLPFC, OFC, and rACC activity during anticipation of pain with placebo. These anticipatory increases correlated with placebo effects on reported pain, and anticipatory increases in DLPFC and OFC correlated with subsequent placebo-induced reductions in brain activity during thermal stimulation. Other studies have replicated and extended this result, showing that placebo treatments for negative emotion activate the same brain regions, and that endogenous opioids – neurochemicals linked to relaxation, euphoria, and pain relief – are released in these regions following placebo treatment.

The expectancy versus conditioning debate

It is difficult to resolve the relative contributions of expectancy and conditioning to placebo effects, because the two are not always mutually exclusive; in some cases, conditioning procedures are likely to shape both learning and expectations. There are two ways to distinguish between learning and expectancy mechanisms: One relies on behavioral observations and the other on measurement of the brain. Earlier, we suggested that conditioning results in learning that persists over time, in spite of expectancies; when a CS is presented without the UCS, extinction of the CR is relatively slow. Thus, effects that can be reversed in a single trial or affected by verbal instructions are not likely to be the result of conditioning, but rather expectancies. In a classic placebo study, expectancies of pain relief were manipulated by surreptitiously turning down the stimulus intensity during testing of a topical placebo solution. Placebo effects on pain developed over the course of this manipulation. However, some participants were informed that the intensity was being reduced during the placebo administration, while others were told that it was

not; this latter group presumably attributed the reduction in pain to the placebo. Although the physical stimuli were identical for the two groups, including the putative CS (placebo application) and UCS (reduction in pain), placebo effects were about 7 times as large when the verbal instructions led participants to expect large reductions in pain attributable to the placebo.

Other studies have demonstrated that 'conditioned' placebo effects in pain and even basic fear conditioning in humans are reversible by changing the instructions to subjects, suggesting that expectations are mediating the effects rather than automatic, learned associations. However, some kinds of placebo effects are not reversible by changing the instructions. In another classic study, researchers repeatedly injected sumatriptan, a drug that induces cortisol and growth hormone release, thereby forming an association between the injection and the drug response. After this conditioning procedure, injecting saline alone elicited cortisol and hormonal increases, and these increases were not blocked by changing the verbal instructions.

A second way to discriminate between conditioning and expectancy is by measuring brain activity itself. The patterns of placebo-induced activity increases in OFC and rACC, and increases in DLPFC, suggest that general mechanisms of appraisal and expectancy are at work. Such effects have been found in pain and, though less well studied, depression; in pain, these same expectancy- and appraisal-related regions have been shown to exhibit placebo-induced opioid release. A difficulty, however, is that there is no way to ensure by looking at the brain that these responses are not the result of some conditioned association being activated. Another difficulty is that it is currently difficult or impossible to measure learned associations directly in the human brain; whereas synapse strength, gene expression, and other molecular markers of learning can be investigated in animal models, the techniques for probing them are invasive and cannot be used in humans – and, in addition, the area in the human brain where cellular learning underlying placebo effects may be taking place is still unknown.

Other central mechanisms

Opioid release in anticipation of pain under placebo correlates with dopamine release in the nucleus accumbens – a region known to be highly involved in reward processing. Similarly, the placebo response in Parkinson's disease has been linked to increased dopamine release in this region. A promising new theory of placebo has thus suggested that placebo administration may actually be rewarding, and that it is this positive affective shift that results in observed placebo effects on pain and motor performance in Parkinson's disease. Work demonstrating correlations between dopamine release under placebo with subsequent activity in a reward task shows that individual differences in reward response correlate with individual differences in placebo. Researchers have postulated that two activation systems that are mutually exclusive exist – a positive, approach system (behavioral activation system), and a negative, withdrawal system (behavioral inhibition system). Placebos may induce a shift from a withdrawal system to an approach system, and this may induce concomitant effects on pain, depression, Parkinson's disease, and other conditions, which would be observed as placebo responses.

Proximal Mechanisms

Proximal mechanisms refer to the pathways whereby central mechanisms interact with the actual processing of a stimulus or condition in order to elicit changes in domain-specific activity. Proximal mechanisms are therefore likely to be different for different disorders. We will specifically review proximal mechanisms for pain, depression, and Parkinson's disease, as these placebo responses are the most well understood.

Proximal placebo mechanisms in pain

A central question in the study of placebo analgesia has concerned the level (or levels) of the pain pathway at which the placebo response has its effect. The naloxone studies reviewed earlier illustrate the important role of endogenous opioids in the expectancy-based and opioid-conditioned placebo responses. Endogenous opioids facilitate modulation of both descending pain-control circuits and central processing of nociceptive information, illustrating their significance as one of the key mediators of the placebo response in pain.

The gate control theory and its successors postulate that descending modulatory signals under placebo can cause inhibition of nociceptive processes in the spinal cord, before signals reach the brain. In animal work, a specific type of opioid receptor, the μ-opioid receptor (MOR), has been shown to inhibit nociceptive transmission at the level of the spinal cord's dorsal horn. The PAG, a region that has been repeatedly demonstrated to be involved in placebo analgesia, is rich in MORs and plays an important role in this descending modulation in animals. As mentioned earlier, an fMRI study that directly examined the placebo response found placebo increases in an area of the midbrain surrounding the PAG during anticipation of pain. This activity could be consistent with the gate control hypothesis, in that the placebo context would increase opioid release by the PAG, and descending opioids would inhibit subsequent pain at the level of the spinal cord's dorsal horn. Further support for the role of the PAG in the placebo response comes from PET studies showing that placebo treatment causes increased opioid release in PAG during pain. Spinal inhibitory mechanisms in humans are also supported by a study showing that placebo manipulations can reduce the area of skin that develops hyperalgesia with repeated thermal stimulation. Animal studies have shown that this hyperalgesia is due to spinal neuron sensitization; hence, placebo effects that decrease the area of hyperalgesia are considered as evidence of descending modulation at the level of the spinal cord.

While these studies and others suggest that early modulation of ascending nociceptive signals may play a key role in placebo analgesia, it is important to recognize that this may not be the only factor responsible for observed placebo effects in pain. The EEG results reviewed earlier suggest that while early inhibition does indeed occur, it cannot entirely account for the magnitude of decreases in reported pain. Recent studies have taken a mechanistic approach to investigating the role of endogenous opioids in the placebo response, elaborating on the knowledge available from naloxone studies. These studies have shown where in the brain placebo changes opioid release, strengthening the argument for placebo changes in central nervous system processing. In PET studies of placebo analgesia that measured MOR activity directly, PAG opioid activity both decreased in anticipation of pain under placebo and increased during painful heat, suggesting that placebo might diminish the threat associated with upcoming nociceptive stimulation and enhance opioid release induced by pain. These studies also demonstrated that placebo treatment increases μ-opioid system activation in the OFC, perigenual ACC, rACC, right anterior insula, left dorsolateral PFC, thalamus, amygdala, and the nucleus accumbens, suggesting that placebo treatments affect opioid responses in many regions critical for affective valuation. The rACC and insula regions corresponded to those that had been shown to decrease during pain in the aforementioned fMRI experiments, suggesting that these changes in pain responses may have been opioid-mediated. In one of these studies, placebo increased connectivity between PAG and rACC, as well as other distinct subsystems of correlated regions, suggesting that placebo treatment involves central opioid release that increases functional integration across regions. These effects could influence the central processing of pain above and beyond effects mediated by descending spinal control systems.

Proximal placebo mechanisms in major depressive disorder

In depression, placebo effects on the brain have been examined by using PET imaging to measure baseline metabolic activity before, during, and after treatment with either placebo or an active medication. Many changes that were observed as part of successful treatment with the active drug were also observed in placebo responders, including metabolic decreases in subgenual anterior cingulate. This region has been shown to be consistently affected in depression and is a target of deep-brain stimulation in patients who do not otherwise respond to treatment. Other common sites of activity included increases in prefrontal, parietal, and posterior cingulate cortex. Importantly, these common results differ from patterns of brain activation over the course other types of treatment, such as cognitive behavioral therapy. This suggests that both active drug and placebo treatments work in part by changing central systems involved in affective valuation and motivation. Much more work remains to be done

to unpack the brain mechanisms involved in both verum and placebo treatment for depression.

Proximal placebo mechanisms in Parkinson's disease

In Parkinson's disease, active placebo responses are likely to involve dopaminergic pathways. PET studies of dopamine D2 receptor activity have provided evidence that placebo treatments lead to dopamine release in the striatum. Complimentary evidence has been obtained from neurosurgical studies, in which placebo (sham) stimulation of the subthalamic nucleus – a stimulation site used in the treatment of Parkinson's – has been shown to affect both subthalamic nucleus activity (decreased bursting and neuronal frequency discharge) and muscle rigidity, which is a clinical sign of the disease. Recent evidence indicates that expectations of effective treatment affect dopamine levels and muscle activity in a dose-dependent fashion.

Whether dopaminergic mechanisms play a role in placebo effects in pain, depression, and other domains remains to be tested more thoroughly. It has been proposed that dopamine release underlies positive affective and motivational shifts that lead to improved outcomes across disorders. Whether outcomes can be influenced by changes in affective and motivational states may determine whether placebo treatments are effective; ongoing research is now being conducted to test this hypothesis.

Summary

While many factors can potentially lead to observed placebo effects without changing underlying processing, careful experimental manipulations provide evidence of active placebo responses in the domains of pain, Parkinson's disease, depression, asthma, and conditioned immunosuppression. There is clinical evidence for effects on other conditions, such as hypertension, anxiety, and heart function, but more experimental research is needed to assess whether placebo treatments can truly affect outcomes in these domains. Neuroimaging methodologies allow researchers to identify and examine active central and proximal mechanisms supporting the placebo response. This, in turn, provides a powerful window into mind–body interactions.

See also: Hypnosis and Suggestion; Psychoactive Drugs and Alterations to Consciousness.

Suggested Readings

Benedetti F, Colloca L, Torre E, et al. (2004) Placebo-responsive Parkinson patients show decreased activity in single neurons of subthalamic nucleus. Nature Neuroscience 7(6): 587–588.

Benedetti F, Pollo A, Lopiano L, Lanotte M, Vighetti S, and Rainero I (2003) Conscious expectation and unconscious conditioning in analgesic, motor, and hormonal placebo/nocebo responses. Journal of Neuroscience 23(10): 4315–4323.

de la Fuente-Fernandez R, Ruth TJ, Sossi V, Schulzer M, Calne DB, and Stoessl AJ (2001) Expectation and dopamine release: mechanism of the placebo effect in Parkinson's disease. Science 293(5532): 1164–1166.

Hrobjartsson A and Gotzsche PC (2001) Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. The New England Journal of Medicine 344(21): 1594–1602.

Kong J, Gollub RL, Rosman IS, et al. (2006) Brain activity associated with expectancy-enhanced placebo analgesia as measured by functional magnetic resonance imaging. Journal of Neuroscience 26(2): 381–388.

Meissner K, Distel H, and Mitzdorf U (2007) Evidence for placebo effects on physical but not on biochemical outcome parameters: a review of clinical trials. BMC Medicine 5: 3.

Montgomery GH and Kirsch I (1997) Classical conditioning and the placebo effect. Pain 72(1–2): 107–113.

Petrovic P, Dietrich T, Fransson P, Andersson J, Carlsson K, and Ingvar M (2005) Placebo in emotional processing – induced expectations of anxiety relief activate a generalized modulatory network. Neuron 46(6): 957–969.

Petrovic P, Kalso E, Petersson KM, and Ingvar M (2002) Placebo and opioid analgesia – imaging a shared neuronal network. Science 295(5560): 1737–1740.

Ploghaus A, Becerra L, Borras C, and Borsook D (2003) Neural circuitry underlying pain modulation: Expectation, hypnosis, placebo. Trends in Cognitive Science 7(5): 197–200.

Pollo A, Vighetti S, Rainero I, and Benedetti F (2003) Placebo analgesia and the heart. Pain 102(1–2): 125–133.

Scott DK, Egnatuck CM, Wang H, Koeppe RA, Stohler CS, and Zubieta JK (2008) Opposite dopamine and opioid responses define placebo and nocebo effects. Archives of General Psychiatry 65: 220–231.

Scott DJ, Stohler CS, Egnatuk CM, Wang H, Koeppe RA, and Zubieta JK (2007) Individual differences in reward responding explain placebo-induced expectations and effects. Neuron 55(2): 325–336.

Stewart-Williams S and Podd J (2004) The placebo effect: Dissolving the expectancy versus conditioning debate. Psychological Bulletin 130(2): 324–340.

Wager TD, Matre D, and Casey KL (2006) Placebo effects in laser-evoked pain potentials. Brain, Behavior, and Immunity 20(3): 219–230.

Wager TD, Rilling JK, Smith EE, et al. (2004) Placebo-induced changes in fMRI in the anticipation and experience of pain. Science 303(5661): 1162–1167.

Wager TD, Scott DJ, and Zubieta JK (2007) Placebo effects on human m-opioid activity during pain. Proceedings of the National Academy of Sciences of the Unite States of America 104(26): 11056–11061.

Biographical Sketch

Lauren Y Atlas began her doctoral work at Columbia University in September 2006. She completed her undergraduate education in 2003 at the University of Chicago, where she worked with John Cacioppo in the Social Neuroscience Laboratory. Her undergraduate thesis research examined cardiovascular differences between lonely and nonlonely participants during active and passive coping. After graduating, she worked as fMRI project coordinator in Stanford's Mood and Anxiety Disorders Laboratory under the direction of Ian Gotlib, where she was involved in projects investigating the neural bases of cognitive and affective processing in major depressive disorder, social anxiety disorder, and bipolar disorder. Her graduate work with Dr. Tor Wager at Columbia takes a mechanistic approach to the study of how expectancies modulate affective experience. Current projects use fMRI and psychophysiology methodologies to examine brain pathways mediating the relationship between applied nociceptive stimulation and subjective pain.

Dr. Tor D Wager received his PhD from the University of Michigan in cognitive psychology, with a focus in cognitive neuroscience, in 2003. He joined the faculty of Columbia University as an assistant professor of psychology in 2004. His primary research interest is in the neural and psychological bases of cognitive and affective control. His research quantifies behavioral performance and brain activity to investigate the neural mechanisms by which humans have flexible control over their behavior. This approach emphasizes the mutual constraints on interpretation afforded by studying behavior and functional anatomy at the same time. His main research interests along those lines are brain and psychological mechanisms underlying the cognitive control of pain and affect; individual differences in selective attention, inhibition, task switching, and other executive processes; and the relationship between affective regulation and cognitive control. He is also interested in developing image analysis and statistical

modeling methods that will improve our ability to use fMRI as a research tool in cognitive and affective neuroscience. Current projects along these lines include optimization of experimental design for fMRI experiments, meta-analysis of functional imaging data, nonlinear alternatives to hemodynamic response fitting, robust regression techniques in massively univariate linear models, and application of multivariate techniques.

Psychoactive Drugs and Alterations to Consciousness

A Dietrich, American University of Beirut, Lebanon

Glossary

Hallucination – Abnormal perceptual experiences that occur in the absence of a physical stimulus and while the person is awake and conscious.
Opiates – Drugs that act on opiate receptors and mediate relief from pain and produce feelings of euphoria.
Psychedelics – Drugs that induce hallucinations or have 'mind manifesting' or 'expanding' properties.
Psychoactive drugs – Drugs that act on the central nervous system and alter behavior and mental processes.
Sedatives and hypnotics – Drugs that depress or inhibit brain activity and produce drowsiness, sedation, or sleep; relieve anxiety; and lower inhibition.
Stimulants – Drugs that produces behavioral arousal via a variety of neural mechanisms.
Therapeutic index (TI) – A measure of the safety of a drug that takes into account the difference in dose between an effective dose and a lethal dose.
Tolerance – A compensatory process in which the effect of a particular drug diminishes with repeated exposure to the drug. There are several types of tolerance, such as cross, metabolic, physiological, behavioral, and learned tolerance.
Transient hypofrontality theory – A theory that proposes a common neural mechanism for all altered states of consciousness, the transient downregulation of brain activity in the prefrontal cortex.

Introduction

Drugs have been used and abused to change consciousness since recorded history. Indeed their use goes back probably much longer, as animals are known to seek out psychoactive substances. Alcohol, perhaps the oldest such drug, was discovered independently by most cultures and the psychological effects of many naturally occurring compounds were known to the Chinese, Indians, Egyptians, and Greeks. The pursuit of intoxicated happiness is so pervasive – even among other animals – that it has been called the fourth drive, after hunger, thirst, and sex. The history of drug use and abuse from antiquity to the present day drug pharmacology – especially their mechanisms of action – and the legal and health implications are all relentlessly fascinating topics, but this article is not the place to pursue these issues. This article focuses instead on the changes to phenomenology, with particular emphasis on the mind-altering effects induced by hallucinogens and psychedelics. Apart from these chemical altered states, this article also examines the problem of defining altered states in general and how they fit, with nondrug states, into an overall framework for altered states of consciousness (ASC).

Problems with Definitions

Consciousness is notoriously hard to define. This makes ASC even harder to capture in a single neat definition. Several attempts have been made to crystallize the concept into a usable definition but, to put it bluntly, these efforts have not been widely adopted. In consequence, we have a poor understanding of the subject matter under investigation and a lack of agreement about the type of phenomena or states that should be counted as examples of ASC. From an empirical perspective, the matter gets worse due to the ephemeral nature of these altered states. Altered mind states are difficult to induce reliably in the laboratory and any effect an induction procedure might have on consciousness is even harder to observe, let alone quantify. In short, the independent variable cannot

be readily manipulated and the outcome measures rely on introspective verbal report – a set of circumstances that has led to little research being done on the subject.

In addition, ASC is a topic that most neuroscientists do not see as a suitable subject for building a scientific career. There are, of course, exceptions – sleep research or experimental hypnosis, most prominently – but by and large research on altered states is not supported by funding agencies and, as a consequence, few psychologists or neuroscientists can afford to make them their primary area of interest. This polite neglect is also due to the bad reputation of ASC. Many view them with suspicion – some sort of abstruse psychopathology at the lunatic fringe frequented mostly by potheads and meditating yogis. This bias is made worse by the excessive use of esoteric language used to describe these mind states. Thus, an undeserved aura of lawlessness and unscientific hogwash sadly surrounds them. However, studying altered states is probably one of the best ways to study consciousness itself because most phenomena, mental or otherwise, are usually studied by manipulating them. As is the case for consciousness as a whole, there has been a resurgence of sound research on altered states in recent years. This is motivated also by evidence pointing at bona fide medicinal benefits for some altered states, such as hypnosis, meditation, or cannabis use. The advantage of this progress, apart from restoring some scientific legitimacy to the topic, is much exciting new data that informs our understanding of the phenomenology and brain mechanisms of ASC. The prospect of understanding altered states in terms of their neural substrates holds the great promise that we might clarify at a deeper level questions regarding the nature of altered states, such as, for instance, how ASC differ from normal consciousness and how they are all related to one another.

Trying to fit states of consciousness into some kind of system – a continuum or hierarchy, perhaps – is not novel but the recent advances in cognitive neuroscience have made this prospect more possible than ever. Two earlier models, one by Charles Tart and a more recent, neurobiologically informed attempt by Allan Hobson arrange conscious states by positioning them in a multidimensional space. Hobson's AIM model maps states of consciousness in a brain–mind space along three parameters: Activation, an energy dimension that depends on the activity of the reticulo-thalamocortical system; Input–output gating, an information source dimension that estimates the extent to which external and internal information is being processed; and Mode, a chemical modulation dimension that estimates the mix of aminergic (norepinephrine and serotonin) and cholinergic (acetylcholine) influences, which runs from low aminergic to high aminergic. This is particularly useful for drug states as the mechanisms of action of most psychoactive substances that induce profound alterations to consciousness are well known.

Tart's phenomenal space consists of two dimensions: irrationality and ability to hallucination. Tart, who also coined the term and popularized the concept of ASC, sees states of consciousness as discrete entities. Some positions in this multidimensional space are stable and can be occupied for a long time, while others are unstable and cannot be occupied except for very brief moments. The twilight between sleep and waking is an example of such an unstable position. We cannot operate in this zone for long and consciousness quickly gives way to the stable state of either waking or sleep. This led Tart to propose the model of what he called state-specific sciences. The concept is that each state of consciousness (SoC) has its own reality, logic, and physical laws. Any one state of consciousness cannot, therefore, be understood by a person in a different state, as it is observer dependent. It follows that all knowledge, scientific or otherwise, is only relevant to the SoC in which it is produced and can only be understood by someone residing in that same state. This places consciousness in the center of science – a bold move indeed. If you find it challenging to wrap your mind around the idea of state-specific sciences and its implications consider that the reason for your doubts may have to do with you not being in the right states of consciousness! This position also argues against the commonsense notion that there is one normal state and all other states are altered states. There is, then, no qualitative difference among SoC, echoing the thinking of William James who wrote many years earlier: "Our normal waking consciousness is but one special type of consciousness." This is in tune with others who

have also emphasized that ASC should not be considered higher, or lower, as the case may be, states of consciousness, because this introduces a value judgment without proper evidence.

Another upshot arising from the new evidence on the neural underpinnings of ASC is the old hope that altered states could be defined, in the future, by objective criteria, perhaps through better neuroimaging tools or an entirely new technology not yet invented, rather than by introspective self-report. In the meantime, ASC can only be defined in subjective terms. Two such definitions are commonly used in this context. The first comes from Tart, who defined ASC as "a qualitative alteration in the overall pattern of mental functioning, such that the experiencer feels his consciousness is radically different from the way it functions ordinarily." A nearly identical definition is proposed by William Farthing, who identifies ASC as "a temporary change in the overall pattern of subjective experience, such that the individual believes that his or her mental functioning is distinctly different from certain general norms for his or her normal waking state of consciousness."

Although definitions such as these do seem to capture the essence of ASC, several obstacles arise that make it impossible to use them in scientific research. Because subjective definitions rely, by their very nature, exclusively on introspection, we can only determine for ourselves whether or not we have entered an altered state. This makes altered states unverifiable from an objective, third-person perspective. Consider a sample of troublesome questions that expose the problem clearly. To start, take two people who have the exact same level of alcohol intoxication (matched for variables such as tolerance, weight, etc.), but one claims to be in an altered state while the other vehemently denies it? Or what about antidepressant drugs? Is a Prozac state an ASC? It is easy to see that this is an unsatisfactory state of affairs. The same problem arises when trying to compare some altered experience to our default mode of consciousness. What is the default here, we might want to reasonably ask? Normal consciousness itself is highly variable and the baseline, if there is one for each person, differs not only among people but also with each person. Consider more examples highlighting this predicament. If someone is madly in love, should we consider this feeling an ASC? In light of the above, subjective definitions, it is certainly not easy, let alone clear, to decide on this. We can draw out the same predicament for other situations, such as PMS? Apart from bodily changes, women often describe their symptoms in terms of alteration to cognition and emotion. Does that count as an ASC?

As for drug states, scientists and users typically agree that drug states constitute ASC. However, even here this troublesome mess cannot be entirely avoided. There is an agreement – by consensus, not via objective criteria though – that psychedelics and opiates alter conscious experience, but what about some less spectacular substance, say, amphetamines, Valium, or ecstasy And even for the clear, putative cases – LSD, mescaline, PCP, etc. – we must consider factors such as dosage or tolerance before we can be sure that a person has entered an ASC.

One final aspect of defining ASC is that all definitions, those above included, emphasize that altered states are temporary phenomena. Permanent changes to mental status, such as those that occur in neurological and psychiatric disorders, are usually not considered ASC, although it is typically assumed that consciousness is altered in these conditions. Drug abuse can, however, result in permanent brain damage. How to classify these changes to consciousness is not clear either.

Transient Hypofrontality Theory

Given the problem with subjective definitions, one solution has been to sidestep defining ASC altogether. This is also typically done in consciousness research in general. By simply considering, without further examination, certain mental states altered states – dreaming, daydreaming, meditation, hypnosis, runner's high, drug states, most prominently – researchers can make progress on the neural basis of these states in the hope of finding a better angle from which to consider ASC. One such approach has been the transient hypofrontality theory. The transient hypofrontality theory proposes a common neural mechanism for altered states that can be subjected to empirical study. The theory is explicitly based on functional neuroanatomy. It views consciousness as composed of various attributes, such as self-reflection, attention, memory,

perception, and arousal, which are ordered in a functional hierarchy with the frontal lobe necessary for the top attributes. Although this implies a holistic view in which the entire brain contributes to consciousness, it is evident that not all neural structures contribute equally to consciousness. This layering concept localizes the most sophisticated levels of consciousness in the zenithal higher-order structure: the prefrontal cortex. From such consideration, the transient hypofrontality theory of ASC can be formulated, which attempts to unify all altered states into a single theoretical framework.

Because the prefrontal cortex is the neural substrate of the topmost layers, any change to conscious experience should affect, first and foremost, this structure followed by a progressive shutdown of brain areas that contribute more basic cognitive functions. Put another way, the highest layers of consciousness are most susceptible to change when brain activity changes. It follows from this 'onion-peeling' principle of sorts that higher cognitive processes such as working memory, sustained and directed attention, and temporal integration are compromised first when consciousness changes. Anecdotal evidence, particularly for drug-induced altered states, supports this idea. All altered states share phenomenological characteristics whose proper functions are regulated by the prefrontal cortex, such as time distortions, disinhibition from social norms, or a change in focused attention. The evidence is particularly strong from psychopharmacology. For animal and human subjects, psychoactive drugs have been regularly shown to negatively affect exactly these mental processes – working memory, executive attention, and the ability to estimate the passage of time. This suggests that the neural mechanism common to all altered states, those induced by psychoactive drugs included, is the transient downregulation of modules in the prefrontal cortex.

The reduction of specific contents to consciousness is known as phenomenological subtraction. The deeper an altered state becomes, induced by the progressive downregulation of prefrontal regions, the more of those subtractions occur and people experience an ever greater departure from their normal phenomenology. In altered states that are characterized by severe prefrontal hypoactivity – various drug states such as those induced by LSD or PCP, for instance – this change results in an extraordinarily bizarre phenomenology such as hallucinations and delusions, most prominently. In altered states that are characterized by less prefrontal hypoactivity, such as long-distance running, the modification to consciousness is much more subtle. However, drug-induced states also differ from other altered states, such as meditation or hypnosis, in one important aspect. They add something to experience. So, on top of phenomenological subtraction, which is a hallmark of all ASC, drug states are also marked by phenomenological addition. The specific addition depends highly on the drug itself and the kind of neurotransmitter and mechanism of action it works on. In any event, the individual simply functions on the highest layer of phenomenological consciousness that remains fully functional.

A consequence of the transient hypofrontality theory is that full-fledged consciousness is the result of a fully operational brain. Thus, and despite popular belief to the contrary, default consciousness is the highest possible manifestation of consciousness, and all altered states represent, by virtue of being an alteration to a fully functional

Figure 1 The arachnologist Peter Witt wondered what would happen to the 'mind' of spiders on psychoactive drugs. The results? See for yourself. It makes you reconsider coffee consumption. Witt and Rovner, 1982, reproduced with permission from Palgrave Macmillan Limited.

brain, a reduction in consciousness. ASC that are often presumed to be as higher forms of consciousness, such as, for instance, transcendental meditation or the experiences reported after taking 'mind-expanding' drugs are therefore really lower states of consciousness, as they all reduce cognitive processes – attention, working memory, etc. – that are associated with the highest forms of consciousness. This view is also in contrast to the theories of James and Tart, who maintained that normal consciousness is not qualitatively different from any other state of consciousness. It is difficult to imagine how higher consciousness might look like in terms of brain activity or feel like in terms of phenomenology, but should not it entail an enhancement of mental abilities ascribed to the prefrontal cortex rather than, as is the case in the above examples, their subtraction?

If all drug states share this common neural mechanism, why, then, does each feel different? Specific drugs cause unique phenomenological changes that can be readily and reliably distinguished by drug users – or rats, for that matter – from those caused by some other drug. How can we reconcile this with the proposal that prefrontal downregulation is the underlying cause for all drug states, and indeed for all altered states? One clue is the specific phenomenological additions that come with each drug. Another might be that different drugs target slightly different sets of prefrontal modules, so removing quite specific computation from the conscious experience. Irrespective of these considerations, explanations of the neural basis of drug-induced altered states have focused, in the past, almost exclusively on neurochemical mechanisms. The transient hypofrontality theory makes clear though that a broader framework from cognitive neuroscience can provide a much better understanding of these phenomena of consciousness.

Drug Classification

Psychoactive drugs are compounds that act on the central nervous system and alter behavior and cognition. They are highly lipophilic and thus readily cross the blood–brain barrier that keeps neurons protected from circulating blood. Psychoactive drugs alter synaptic transmission by modulating neurotransmitter amounts and availability or by affecting receptor activity. In addition to their primary effects on mental processes – arousal, perception, mood, cognition, consciousness – these drugs produce a variety of nonbehavioral effects that are often far more dangerous to health, and, in some instances, can be lethal.

There are several different classification schemes for psychoactive substances – according to legal, medical, or pharmacological criteria – but the most common organization is based on their effect on behavior. This scheme classifies these drugs into four broad classes.

The first are the sedatives and hypnotics. Drugs in this class depress or inhibit brain activity and produce drowsiness, sedation, or sleep; relieve anxiety; and lower inhibition. Although depressants do not share a common neural mechanism, most of them either decrease the metabolic activity in the brain or increase the transmission of the principal inhibitory neurotransmitter, gamma-aminobutyric acid (GABA). All sedative compounds have the potential for addiction and dependency. Common examples include barbiturates, such as Seconal; benzodiazepines, also known as minor tranquilizers, such as Xanax or Valium; nonbarbiturate sedatives, such as methaqualone; nonbenzodiazepines, such as buspirone; antihistamines and anesthetics; and alcohol. In low doses, alcohol can act as a stimulant, but with increased dosage its main effects are almost always depressive. Marijuana, which is derived from the hemp plant *Cannabis sativa*, is often misleadingly classified in the category of psychedelic. Given that its most pronounced behavioral symptom is sedation and the fact that it rarely produces – and only in very high doses – sensory distortions of hallucinatory quality, it is also best categorized here as a sedative.

The second class are stimulants. These drugs produce behavioral and mental arousal. This class of psychoactive compounds also includes a variety of different substances each of which may also have a different neural mechanism. Common examples are amphetamines; cocaine; the methylxanthines, such as caffeine which is the most widely used psychoactive drug in the world, theophylline, which is present in tea, and theobromine, the psychoactive ingredient in chocolate; nicotine or

tobacco; appetite suppressants; and a variety of exotic plants, such as the betel nut, khat, yohimbe, and ephedra. Stimulants vary in strength, legal status, and the manner in which they are taken; however, all stimulants have addictive potential. None of these preparations, however, induces radical changes to mental status, such as, for instance, hallucinations or delusions, except perhaps cocaine and amphetamines when taken in high doses.

The third class consists of the opiates. Drugs in this class act on opiate receptors and their two main effects are that they mediate relief from pain and produce feelings of euphoria, which is experienced in a dreamlike state of consciousness. It is also these two effects, of course, that make them drugs of abuse. In pharmacology, the term narcotics, derived from the Greek word for stupor, is used interchangeably with opiates but the legal systems in most countries use the term narcotics to refer to all illicit drugs, regardless of their mechanism of action. Opiates are highly addictive and can either be natural, semisynthetic, or synthetic. Natural opiates such as opium are derived from the opium poppy. The active ingredients of opium are morphine and codeine. The most famous semisynthetic opiate is heroin, which is two hundred times more potent than morphine. If taken intravenously, heroin produces what is known as a rush, an intense feeling of pleasure of euphoria that, according to first-hand reports, is not comparable in strength to any other experience. This makes heroin the most psychologically addictive drug known. Examples of synthetic opiates, also known as designer drugs, include methadone, naloxone, and the prescription pain medication Demerol, which is the drug most abused by physicians.

The fourth and last class consists of the hallucinogens and psychedelics. Drugs in this category share the common feature that they induce hallucinations or have other 'mind manifesting' or 'expanding' properties. But this is also where the commonality ends. Pychedelics can either occur naturally in a plant, such as mescaline, which is derived from the peyote cactus, or be synthesized in the laboratory, such as LSD. In the latter case, such drugs are also known as designer drugs. In terms of consciousness – or behavior, for that matter – no overarching framework seems to be emerging that can organize their wide-ranging, phantasmagoric effects. Pharmacologically, they are also all over the place. For every classical neurotransmitter system, there is a drug in this class. In consequence, psychedelics are probably best classified according to the neurotransmitter system they primarily affect. Cholinergic psychedelics include, most prominently, physiostigmine, scopolamine, and atropine, which are well known to cause bad trips. Drugs that alter norepinephrine transmission include mescaline and the popular ecstasy (MDMA), which cannot be said to cause bone fide hallucinations. Drugs that alter serotonin transmission include lyseric acid diethylamide or LSD, psilocybin, and morning glory. These drugs are perhaps most commonly associated with the psychedelic experience. There are also drugs with psychedelic properties that alter dopamine (cocaine), glutamate (ketamine), GABA (muscimol), opioid (morphine), or cannabinoid (hashish) transmission; however, these preparations are not predominantly hallucinogenic and they are best classified elsewhere according to their respective main effect. A peculiar subclass of the hallucinogens and psychedelics category are the psychedelic anesthetics, such as phencyclidine or PCP (angel dust), ketamine, and several gases and solvents, such as ether or nitrous oxide. With the exception of the cholinergic psychedelics, all drugs in this class have a high margin of safety, that is, a high TI and are generally nonlethal – even when taken in large quantities.

In addition to these four broad categories, there a number of other drugs that affect the mind. These compounds are used to treat a variety psychological and neurological disorders and include antidepressants; antipsychotic medication, such as clozapine, chlorpromanzine, and haloperidol; and drugs for epilepsy, Parkinson's disease, the dementias, such as Alzheimer's disease, and spasticity.

In most countries, the legal system takes a different approach to classifying psychoactive drugs. In the United States, for instance, the Comprehensive Drug Abuse and Prevention and Control Act of 1970 regulates these substances. According to this scheme, drugs are classified into five schedules according to the perceived risk of developing dependency to that drug. Schedule I drugs such as heroin, marijuana, and most psychedelics have a high risk of dependency and no accepted medical

use. These drugs are forbidden and cannot be obtained by prescription or, often, even for research, for that matter. Schedule II drugs, such as morphine, codeine, amphetamines, and certain barbiturates have a high risk of dependency but are accepted by the medical community for some treatment, that is, they have a medicinal purpose. Schedule III drugs have a risk of moderate physical dependency or high risk of psychological dependency and include preparations with limited opiates (e.g., morphine) and barbiturates not included in Schedule II. Schedule IV drugs, such as the benzodiazepines, have a slight risk of mild physical or psychological dependency. Schedule V drugs have less risk of mild physical or psychological dependency. Finally, alcohol and tobacco are not classified under this law. They fall under the jurisdiction of the Alcohol, Firearms, and Tobacco agency, which is a division of the US Department of Treasury.

This is as good a place as any for a note of caution. Most explanations on the neural basis of altered states rely solely on neurochemical modulation. The reason for this is that drugs are the most common way to induce altered states and their known mechanism of action for changing brain function was for a long time the only hope to find a sound, mechanistic explanation for the mercurial nature of ASC. But there is a danger of leaning too hard on neurochemistry. An increase in serotonin does not cause happiness any more than an excess of dopamine causes paranoid schizophrenia. Neurotransmitters *per se* do not carry content in their messages. Drugs that work on the same neurotransmitter system, even via the same synaptic mechanism, can have very different effects on consciousness (Table 1). Prozac, for instance, increases serotonergic transmission and is a well-known antidepressant. LSD, on the other hand, is also a serotonergic agonist. LSD, however, sends you on a trip of some kind, but it certainly is not an antidepressant. Similarly, cocaine is a dopaminergic reuptake blocker and produces alertness and euphoria. Ritalin, the treatment of choice for children with attention deficit hyperactivity disorder (ADHD), works on exactly the same neural mechanism, blocking the reuptake of dopamine. Cocaine gets you high, though, to say nothing of prison, while Ritalin has a calming effect and the medical community has approved it to be administered to kids. The moral here is this. Behavioral pharmacologists have long given up on the naive idea that complex psychological phenomena can be attributed to changes in the concentration of a simple little molecule. Every transmitter system can be modulated by drugs that do and do not alter consciousness. As such, we must also consider the function of the neural structure in which the changes in chemical neurotransmission occurs.

Phenomenology

An obvious place to examine drug-induced alterations to phenomenology is the psychoactive substances classified under psychedelics and hallucinogens. Drugs in the other categories may produce changes to consciousness and its content as well but these are not anywhere as profound or interesting, at least for the study of consciousness, than drugs that induce perceptual changes, especially hallucinations, and changes to thought patterns, such as delusions. Note however that although not all psychedelics produce hallucinations or delusions, hallucinations have been studied most often by psychologists in this context. In consequence, this part of the article focuses mostly on drug-induced sensory distortions, particularly in visual perception – color, shape, depth, etc. These events occur, however, in all sensory systems and can have, in all of them, hallucinatory strength. Delusions are less frequent and occur mostly with the ingestion of a few selective psychedelics.

Sensory distortions, particularly bona fide hallucinations, do not typically occur spontaneously and are therefore usually not considered under normal experiences. In the Western world, such experiences are mostly associated with madness, as a symptom of a disease in the brain, such as psychosis, or malfunctioning of a bodily system, such as in sleep deprivation or starvation. When they occur in response to drug use, hallucinations are also typically seen as neurological junk or signs of temporary insanity. Not all cultures, however, view hallucinations in this negative light. Some cultures highly value hallucination, including drug-induced hallucinations, and consider them sources of insight and truth. Great spiritual significance is

Table 1 A sampler of psychoactive drugs detailing some of the more interesting effects on consciousness

Category	Drug and origin	Mechanism of action	Phenomenology
Sedatives and hypnotics	Valium	Binds to benzodiazepine site of the GABAa receptor facilitating the effects of GABA	Produces all degrees of behavioral depression from sedation, to relaxation, to coma; reduces anxiety, disinhibition, sleepiness, tiredness, sense of well-being
	Alcohol	Binds to barbiturate site of the GABAa receptor facilitating the effects of GABA	Relief of anxiety, sedation, disinhibition. Stimulant in low doses. Impairments in vision, speech, motor control, and judgment
	Marijuana (Cannabis sativa)	Receptor agonist on the CB1 receptor; also inhibits 5-HT3 receptors	Main effects are sedation and analgesia; others include heightened sensations, uncontrollable laughing, sense of well-being, intensified introspection, mental slowing
Stimulants	Amphetamine	Stimulates the release of DA and blocks its reuptake; also activates adrenergic receptors	Alertness, arousal, insomnia, loss of appetite, and combats fatigue; known as 'speed' when taken IV or as 'speedball' in combo with opioids (to reduce side effects)
	Cocaine (Coca plant)	DA reuptake blocker	Same as for amphetamines; more euphoria and paranoia; freebasing and crack result from purification procedures and intensify the experience, especially the euphoria
Opiates	Morphine, heroin (opium poppy)	Activates opioid receptors	Analgesia, peacefulness, feelings of warmth and well-being, boundless energy, dreamy imagery; when taken IV, heroin induces a 'rush' of intense pleasure that is without equal
Hallucinogens and psychedelics	Scopolamine, atropine (Nightshade)	Cholinergic antagonist	Restlessness, delirium, vivid hallucinations, disconnection with reality, no memory of the experience; OBEs are common; bad trips can easily be filled with ugly monsters
	Mescaline (Peyote cactus)	In a class of its own; structurally related to NE and alters NE transmission	Vivid hallucinations consisting of bright colors, geometric designs, and animals; synesthesia; no OBE, insight is retained; more color and less form distortion than LSD
	MDMA (Ecstasy)	Pharmacologically promiscuous but primarily stimulates the release and blocks the reuptake of 5-HT2	Induces feelings of empathy and sympathy; creates desire for intimacy and need for personal contact, heightens self-awareness; no hallucinations or loss of reality
	LSD	Activates 5-HT2a receptors	Vivid, kaleidoscopic sensory changes (e.g. mosaic patterns on all surfaces), powerful feeling of love, mystical oneness, synesthesia; bad trips include anxiety, panic, violence
	Psilocybin, psilocin (Mushrooms)	Structure resembles 5-HT and activates 5-HT receptors	Pleasant feelings of relaxation, distortions of space, visual hallucinations with more intense color than LSD; effects resemble more mescaline than LSD; fewer bad trips
	DMT (Virola tree)	5-HT agonist	Short-acting and intense psychedelic rush (businessmen's LSD); loss of all reality, overwhelming visual hallucinations; communication with Gods, spirits, or the deceased
	PCP (Angel dust), ketamine	Primarily NMDA receptor agonists but also alter opioid and monoamine transmission	Anesthesia with depersonalization, loss of ego boundaries, changes in body image, floating, OBE, NDE, distortion of time and space, full amnesia, euphoria; frequent bad trips

often attached to them and people interpret them as visions that allow them to see into other realms of reality, realms inhabited by Gods and their ancestors. Because of this sacred meaning, hallucinations are sought out in elaborate rituals and only selected members of the group, such as shamans, are entrusted to have them. These rituals, in addition to drugs, often involve repetitive and rhythmic motion – spinning, whirling, dancing, and jumping – which is coordinated with some sort of cadenced drumming, but hallucinations can also be induced through such behavioral techniques alone – without the aid of a psychedelic substance.

Drug states are not so much an altered state as they are a whole family of altered states. Depending on the nature of the compound, people can experience many different kinds of hallucinations, delusions, and emotional changes. On the positive end, hallucinations may include intense colors, kaleidoscopic visions, fantastic images of animals and landscapes, mosaic patterns on all surfaces, as well as vibrating, rotating, or exploding designs that retreat into infinity. This is especially the case for the three major psychedelics: LSD, psilocybin, and mescaline. The delusions may include merging with one's surroundings, such as a mystical feeling of oneness with God or the forces of nature; however, they may also be less sacrosanct experiences, such as becoming one with the ashtray on the table. The emotional changes may include pleasant relaxation, mild amusement, or even sweeping euphoria. On the negative end, changes may range from some tenseness or thoughts of paranoia to nail-baiting fear and terror. Drug users may encounter menacing beasts that chase them mercilessly with such a demented roar that would make the average nightmare look quaint in comparison. Other grotesque delusions may contain distortions to the sense of self, such as macabre changes in body image – the nose grows to double its size, for instance – bizarre depersonalizations where the psychonaut, a term Siegel introduced, floats in a void, and out-of-body or near-death experiences (OBE and NDE). As with most drug effects, whether you go to heaven or hell depends heavily on set and setting, that is, where and when you take the drug and in what state of consciousness or mood you happen to be in. Given the noncholinergic psychedelics' high margin of safety, the psychological scars that may result from the so-called bad trips may represent the biggest danger to health while visiting these alternate realities.

Some psychedelic drugs, even in low doses, cut all ties with reality and take the user into an altogether different realm of experiences. PCP and scopolamine delusions are good examples. The danger of losing all understanding of what is real and what is imagined is that PCP-powered phantasms become ontologically fully real and when psychonauts, in their desperation, run from them and jump out of a real window, they are really dead, in this world. Other psychedelic drugs keep the psychonaut on this side of reality and maintain some type of lucidity as into what is real and what is imagined. Mescaline, of peyote fame, ecstasy, and several other norepinephrine psychedelics are drugs that tend to preserve such a tentative hold on reality. This often allows users to retain the insight and thoughts they had during the trip. In addition, for some drugs the changes to mental status are so subtle that they allow users, within limits and at low doses, to temporarily snap out of it.

The anecdotal and esoteric descriptions by drug users, even the above broad overview, often give the mistaken impression that psychopharmacological karma comes in an infinite variety. Different drugs certainly evoke different cognitive and emotional effects, to the extent that they can, as said, be readily distinguished by drug users, even if they happen to be laboratory rats, but all still share common characteristics. The passage of time, for instance, loses all its meaning and people show no evidence of cognitive flexibility, that is, they simply lose any ability to extract themselves from their immediate surroundings – the freeze-frame of the here and now. Psychoactive substances, then, can be said to induce a sort of mental singularity, which, again, strongly questions the idea that these drugs are somehow mind-expanding, which is the original meaning of the term psychedelic, or induce a higher state of consciousness.

Another common feature is the loss of ego boundaries, that is, people feel that they dissolve into the Universal Ocean, merge with the Almighty, or become one with Void. Such experiences are clear indications for the disengagement of cognitive functions supported by the prefrontal

cortex, as it is there that we form a sense of the self and delineate it from other selves. But taken together, drug states are also quite different from nondrug states – meditation, hypnosis, or rhythmic movement – because they are, in addition to phenomenological subtractions, also marked by phenomenological additions, depending on the systems the drug stimulates.

Siegel has investigated the psychedelic experience in detail, collecting and categorizing the phantasmagoric images of the hallucinating mind in an effort to find patterns in what seems to be a phenomenon that can seemingly take on an infinite number of forms. He trained the participants of his experiments on a standardized hallucination code that classified their perceptions in precise language. They would report, for instance, that they see, say, 550 nm rather than a dirty red. His studies revealed that there is a hidden order in the chaos. The brain, it seems, hallucinates in only four basic geometric forms. All hallucinations have some uniqueness to them, some superimposed fiddly bits that provide some creativity, but all are variations on only four recurrent themes. These form constants, as they are called, were first discovered by Heinrich Klüver, who experimented with mescaline in the 1920s. Klüver called them the spiral, the cobweb, the tunnel or cone, and the lattice or honeycomb. Siegel, however, has found the same form constants in hallucinations induced by nondrug altered states, such as migraines, fever, auras, temporal lobe epilepsy, or sensory deprivation. They can be clearly seen in psychedelic-inspired art all around the globe, be they from Huichol Indian paintings, or mandelas, or tie-dye T-shirts. The reason for these form constants can be found in the functional architecture of the visual cortex. Computer models using a columnar organization similar to that of the cortex show that the same form constants also crystallize in computer simulations (Figure 1).

When the dosage is increased, the distorted sensory experiences become more involved and might contain people, storylines, and composite scenes. Siegel has also cartographed and indexed this more complex phase in the hallucinatory process and discovered that here, too, are underlying rules that govern how, exactly, consciousness is altered. There are, for instance, the effects of megalopsy and micropsy, phenomena in which people or objects grow huge or shrink infinitesimally in size. There is also duplication and multiplication. A tunnel always leads to more tunnels and one sunflower becomes an entire field of sunflowers. The way in which motion and metamorphosis occur also has order to it. Birds become bats (never the other way around) and things first pulsate before they revolve. The hallucinating brain is not an unconstrained image generator.

The best way perhaps to demonstrate this mental order is in visual perception. Siegel and Jarvik showed that specific drugs produce precise alterations in the color spectrum and color intensity and these changes that can be described in terms of dose–response curves. Figure 2 shows that the three major psychedelics LSD, psilocybin, and mescaline tend to be very colorful and bring out red, orange, and yellow hues, while marijuana is more on the bluish end of the spectrum. When psychonauts closed their eyes and used nothing but their imagination, blacks, whites, and violets dominated.

A few psychedelic drugs are particularly strange, even for psychedelics, in the way they affect phenomenal consciousness. One is PCP, also known as angel dust. This substance is not popular anymore but in the early 1970s it was for a brief amount of time the drug of choice. The mechanism of action of PCP is not fully understood. It is a promiscuous compound that binds to different types of postsynaptic receptors and thus alters several different neurotransmitter systems at once. PCP is a so-called dissociate anesthetic because it produces, in addition to the usual delusions and hallucinations, two prominent effects: analgesia, the suppression of pain sensations, and an eerie feeling of detachment or dissociation from the world. The phenomenal experience is often extraordinarily bizarre with sensations of depersonalization, complete loss of ego boundaries and self-identity, OBE or NDE, as well as peculiar distortions in body image and self-concept. Abuse of this drug frequently led to symptoms of paranoid schizophrenia requiring treatment and sometimes even hospitalization.

Another peculiar compound is the designer drug ecstasy. Like PCP, it is pharmacologically promiscuous and, although classified as a norepinephrine psychedelic, it acts mostly in serotonergic synapses. This, however, as so often, does not explain its weird

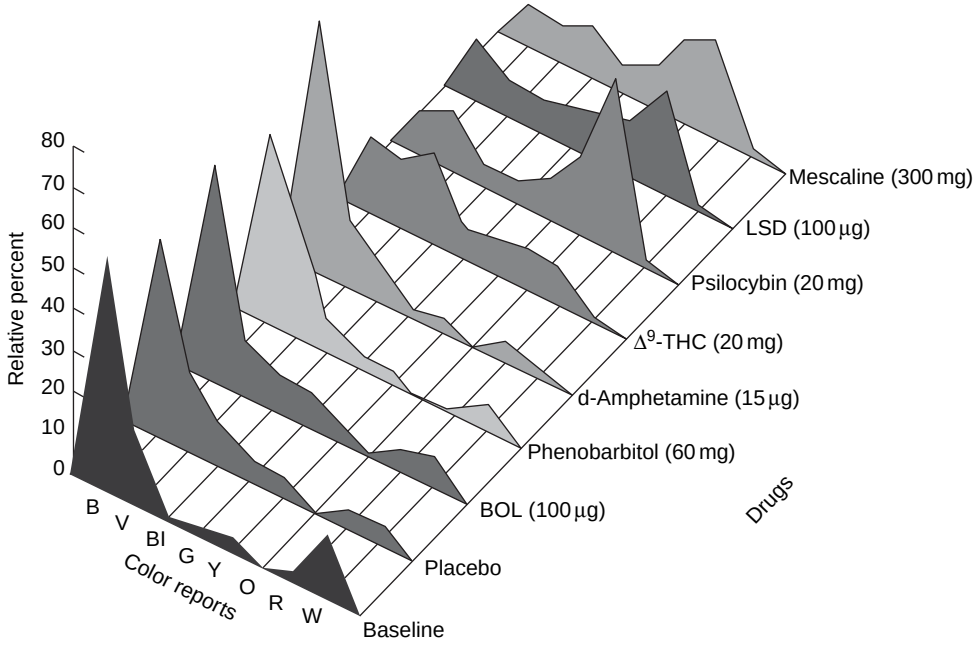


Figure 2 Color in hallucinations is a function of the drug and its dosage. Participants saw warm reds, oranges, and yellows on LSD, psilocybin, and mescaline; somber blues on marijuana and uninspiring blacks, whites, and shades of gray on amphetamine and sedatives. Reproduced from Siegel RK and Jarvik ME (1975) Drug induced hallucinations in animals and man. In: Siegel RK and West LJ (eds.) Hallucinations: Behavior, Experience, and Theory, pp. 81–161. New York: Wiley, with permission from Wiley (New York).

effects on phenomenal consciousness. Much disagreement surrounds ecstasy among psychologists and neuropharmcologists. For one, a controversy in the late 1990s about the neurotoxic properties of ecstasy, a controversy that included, most unfortunately, forged data, has not helped in our understanding whether or not this drug causes permanent brain damage. As for changes to the content of consciousness, ecstasy does not induce true, frank hallucinations. Rather, it heightens the sense of introspection in general and in particular seems to inspire intense feelings of empathy and sympathy. The user feels an increased need and even longing for personal contact and intimate tactile stimulation. This set of properties has led to the popular label of ecstasy as a love drug. In psychopharmacology, the proper term for this is entactogen. This led to two more controversies surrounding ecstasy. The first is the use of ecstasy as a date rape drug. The second is the proposal by some clinical psychologists to use ecstasy's feeling enhancing effects in psychotherapy. Because the user maintains a high level of cognitive functioning and retains any insights gained during the state of heightened introspection, ecstasy, goes the argument, might facilitate the self-discovery process. Needless to say, this proposal has been met by fierce opposition among psychologists and physicians, to say nothing of law enforcement.

Marijuana is a different matter still. Marijuana's mechanism of action was not understood until the mid-1990s when the endocannabinoid system was fully identified. This system is a transmitter system consisting of two receptors, CB_1 and CB_2, and their endogenous ligands, anandamide and 2-arachidonylglycerol (2-AG), which are the endocannabinoids, that is, endogenous cannabinoids. The fact that marijuana does not alter any of the classical neurotransmitter systems explains perhaps the distinct ASC it engenders. Delta-(9)-tetrahydrocannabinol, or THC for short, is the psychoactive constituent of marijuana. It exhibits high affinity to the CB_1 receptor and activation of this receptor complex, by endocannabinoids or by the ingestion of exogenous cannabinoids such as THC, induces emotional and cognitive changes such as analgesia, sedation, anxiolysis, and a sense of well-being. Other prominent effects include a shortened attention span, impaired working

memory, distortions of time estimation, all mediated by hypometabolism in prefrontal cortex regions, as well as enhanced sensory perception and a state of silent introspection. As said, unless high doses are taken, marijuana induces neither hallucinations nor delusions.

Summary

Empirical progress on understanding the nature of drug-induced altered states has been impeded by the lack of an operational definition for these experiences as well as the ephemeral nature of drug effects. Because no single definition is widely accepted, other approaches to the problem, such as the transient hypofrontality theory, have been suggested. Such frameworks, which are based on cognitive neuroscience, raise the hope that we someday might better understand these alterations to consciousness by, perhaps, linking the phenomenology of drug states, or any other altered states, to specific neural markers. These models have already cleared the ground of common misconceptions such as that these states are mind expanding or that they constitute some higher form of consciousness that allows people to connect to the Great Beyond.

We humans are curious about what it would be like to feel different. This is surely one of the main reasons that attract so many people to experiment with altered states, particularly drug states. It is also perhaps the main reason why drugs have featured so prominently in most societies and cultures throughout history. Drug states come in such a great variety and form that they seem to defy classification. There exist several different categories of psychoactive substances and each contains a host of structurally very different compounds that work via separate mechanisms of action involving all known neurotransmitter systems. What's more, there is the bewilderingly complex phenomenology that appears to be highly idiosyncratic. There is a pattern underneath this diversity, though. The brain hallucinates in only a very limited number of stable configurations – the four form constants – that give way to more intricate delusions and hallucinations also according to a number of seemingly simple rules. Colors, for instance, vary in an orderly fashion as a function of the type of drug and its dosage. All this suggests an underlying blueprint, a kind of universal hallucination code. Moreover, drug states share common themes, such as the loss of time perception, the lack of cognitive flexibility, and the disintegration of the self concept, which suggests that the one common neural mechanism underlying these altered states is the downregulation of the prefrontal cortex.

See also: Altered and Exceptional States of Consciousness; General Anesthesia; The Neurochemistry of Consciousness.

Suggested Readings

Bressloff PC, Cowan JD, Golubitsky M, Thomas JP, and Wiener MC (2002) What geometric hallucinations tell us about the visual cortex. Neural Computation 14: 473–491.

Diaz J (2007) How Drugs Influence Behavior: A Neurobehavioral Approach. Upper Saddle River, NJ: Prentice Hall.

Dietrich A (2003) Functional neuroanatomy of altered states of consciousness. The transient hypofrontality hypothesis. Consciousness and Cognition 12: 231–256.

Dietrich A (2007) Introduction to Consciousness. London: Macmillan.

Grinspoon L and Bakalar BB (1997) Marijuana, the Forbidden Medicine, 2nd edn. New York: Yale University press.

Grinspoon L and Bakalar BB (1997) Psychedelic Drugs Reconsidered, 3rd edn. New York: The Lindesmith Center.

Hardman JG and Limbird LE (eds.) (2000) Goodman & Gilman's the Pharmacological Basis of Therapeutics, 9th edn. Elmsford, NY: McGraw Hill.

Hobson JA (2001) The Dream Drugstore. Cambridge: MIT Press.

James W (1890) Principles of Psychology. New York: Holt.

Siegel RK (1989) Intoxication: Life in the Pursuit of Artificial Paradise. New York: Penguin.

Siegel RK and Jarvik ME (1975) Drug induced hallucinations in animals and man. In: Siegel RK and West LJ (eds.) Hallucinations: Behavior, Experience, and Theory, pp. 81–161. New York: Wiley.

Tart CT (1972) States of consciousness and state-specific sciences. Science 176: 1203–1210.

Tart CT (1975) States of Consciousness. New York: Dutton & Co.

Biographical Sketch

Arne Dietrich is an associate professor of psychology and the chair of the Department of Social and Behavioral Sciences at the American University of Beirut, Lebanon. He holds a PhD in cognitive neuroscience from the University of Georgia. Professor Dietrich has done research on the higher cognitive functions supported by the prefrontal cortex, focusing mostly on the neural mechanisms of (1) altered states of consciousness (ASC), (2) the psychological effects of physical exercise, and (3) creativity. His major publications include a new theory of ASC, the transient hypofrontality theory, published in *Consciousness and Cognition* in 2003 and 2004, the proposal of two new mechanistic explanations for the effects of exercise on brain function, the endocannabinoid theory and the transient hypofrontality theory, and a new overall framework for the neural basis of creativity, published in *Psychonomic Bulletin & Review* in 2004. Professor Dietrich is the author of an introductory textbook on consciousness, published by Palgrave Macmillan in 2007, and his research has been featured prominently in the international press.

Psychodynamic Theories of the Unconscious

M Macmillan, University of Melbourne, Melbourne, VIC, Australia

Glossary

Coconscious, subliminal conscious, subconscious, and unconscious – Alternative terms used by the pioneers of depth psychology for describing unconscious mental processes.

Instinctual drive – Freud's concept of a drive having a physiological source that is represented in the mind by an idea which has psychological consequences.

Libido – For Freud it is the psychological energy of the sexual instinctual drive; for Jung it is a general type of psychological energy that takes different forms.

Psychodynamic theory – A theory that casts mental functioning as the result of conflicting mental forces.

Repression – For Freud a mental force that prevented entry of unacceptable ideas to consciousness or the ego.

Structure – One of the three agencies of id, ego, and superego making up Freud's final theory, known as the structural theory, and which supplanted the three systems of the topographic theory.

System – One of the three components, Cons., Precons., and Uncons., that comprised Freud's early topographic theory of the mental apparatus.

> The hypothesis [of unconscious cerebration] rests entirely upon the testimony of consciousness, and this testimony should be considered very suspicious. (Alfred Binet, *Alterations of Personality*, 1892/1896, p. 355).

Introduction

Although the impression is sometimes given that Sigmund Freud *discovered* unconscious mental processes, this view is incorrect. Freud made no such discovery. What he did was to invent a *psychodynamic* concept of unconscious mental processes and *locate* them in a particular repository, the unconscious mind. There is now more than one such conceptualization, but each derives from Freud's.

It is now difficult to recall that in the late 1800s, when Freud invented his concept, a whole new kind of psychology called *depth psychology* had come into being. Depth psychology was not a monolithic psychology, but rather a loose movement, grouped around the central proposition that aspects of mental life relevant to understanding unusual mental states were unconscious. The theoretical and practical contributions to this view made by Jean-Marin Charcot, Pierre Janet, Alfred Binet, and Hippolyte Bernheim in France, Joseph Delboeuf in Belgium, Charles Samuel Myers and Frederic William Henry Myers in England, and William James, Morton Prince, and Boris Sidis in the United States, to name only a few of its pioneers, were regarded as being at least as important as Freud's.

The depth psychologies differed in three fundamental ways. First, they conceptualized nonconscious mentation very differently. For some, these nonconscious mental processes were like their conscious counterparts, or even superior to them, but for others they were inferior. Second, and related to the first difference, there was profound disagreement over what was hidden in the depths. Was it another kind of self, an essentially normal but subconscious personality, or was it material disavowed by normal consciousness organized according to principles very different from those of waking life? Lastly, what caused the mentation to be so lost? Had a passive process caused mental fragmentation or did it result from something active? To these questions we must add that of asking how each theorist explained how we become aware of its workings.

Consciousness and Unconsciousness

Although precise definition may elude us, most of us agree about the everyday use of the terms *conscious* and *unconscious*. Consciousness refers to what is currently in our minds or, more correctly, that of which we are presently aware. Our awareness of what occurred 20 min ago is different, but not markedly so, and can be retrieved with little effort. It is the mark of consciousness that it is there, and immediately so. Unconsciousness is quite different. Not being able to recognize it directly in ourselves, we do so only by being told by another about something that occurred when we were asleep or unconscious in some other way. As a pendant to our not being able to know about our unconsciousness, we usually have no difficulty in recognizing that state in others.

Lancelot Whyte showed in the early 1950s that there had been an extensive debate about unconscious mental processes in the European philosophical literature from the second quarter of the 1700s to 1880. One of the main pre-1890s philosophical issues was whether any meaning at all was attached to the term *unconscious idea*. Some argued that because an idea was not immediately present in consciousness it could not be an idea at all. To explain the recall of an idea experienced 20 min ago, it was assumed that its physiological basis had been revived strongly enough for it to cross the threshold of consciousness. What seemed to be an unconscious idea was merely a physiological process with a potential to become an idea – as a potential it was simply not an idea at all.

Sensory-Motor Physiology and Psychology

Throughout the second half of the 1800s, the dominant model underlying the workings of the nervous system and the psychological processes they supported was that of the *sensory-motor reflex*. Based on a generalization from the simple reflex arc, impulses were pictured as entering the nervous system through the sense organs, being transmitted to the brain and emerging from it as commands to move particular muscles in particular ways. Incoming stimuli and motor responses were associated with one another because they had affective or emotional consequences. In some mid-century writings the important consequence was the gaining of pleasure, but in Alexander Bain's physiological psychology of the 1850s it was the avoidance of pain or unpleasure.

Thinking itself was modeled on associationism. One idea gave rise to another because both had previously been experienced at the same time or because of some similarity between them. What was true for a single association was also true for complex chains: one moved from idea A to idea Z because experience or similarity provided the intermediate pathways $A \rightarrow B \rightarrow C \rightarrow D \ldots W \rightarrow X \rightarrow Y \rightarrow Z$. A pathway like this was usually conscious, but the notion of a threshold allowed for some degree of unconsciousness, such as when links $C \rightarrow D \ldots W \rightarrow X$ failed to be conscious. These links were present physiologically, but at an intensity that could not carry them across the threshold of psychological awareness or consciousness.

Freud and Associationism

Of the versions of this kind of thinking that influenced Freud, the most important were those formulated in the period 1860–90 by Theodor Meynert and Eduard von Hartmann. Meynert's was a physiological thesis in which associations were made by fibers in the white matter that connected the areas in the cortex where sensations were registered. In his famous example, fibers connected the auditory image of the sound of a lamb's bleating with the visual image of its appearance such that the visual image was revived when the bleat was heard. The causal presence of the real lamb was an inductive or logical inference from its sound. Associations, singly or in trains, were thus identical with logical and causal connections. Once the starting image or idea was revived, nothing could derail the train from reaching the image or idea at its destination. Von Hartmann proposed that trains of thought were ordinarily evoked by a consciously willed purpose or motive. Were the conscious purpose dropped or abandoned, an unconscious one took over and directed the thinking to its goal.

We know that Freud was quite familiar with Meynert's ideas. Although he came to disagree with his mentor about the localization of brain function, he consciously adopted Meynert's view that trains of association were physiologically and internally determined, and even used the example of the sound and sight of the lamb. Von Hartmann's books on unconscious mentation circulated very widely among the German reading intelligentsia. Freud had read them as a student although he claimed it was not until about 1914 that his attention was drawn to von Hartmann's thesis about the unconscious guidance of thinking.

Unconscious Ideas and Symptoms

None of the debate Whyte documented was about the unconscious production of symptoms or about their removal. Some fifteen years later, from about the mid-1960s, Henri Ellenberger located clinical concepts of unconscious mental processes in a late 1700s European development in what became known as psychiatry. The trends identified by Whyte and Ellenberger came to fruition in the work of many people by 1890, Freud among them.

The notion that some symptoms, which we would now class as psychiatric conditions resulting from mental processes, is an old one, going back in Europe to at least 300 BCE, to those Greek temples devoted to the healing cult of Aesculapius and to at least 600 CE in healing practices like those of the Islamic Sufis. Central to practices like these was a well-defined sleep ritual from which the patient woke cured. Where there are reasonably detailed descriptions of the symptoms, or realistic representations of them in the votive offerings of grateful patients, it is almost certain that most of the disorders so treated were produced by unconscious mental processes and were hysterical.

Here we need to consider very carefully what is meant by the term hysteria, especially as the concept is now often regarded as 'incorrect' in some politically irrelevant sense. Briefly, hysterical symptoms are a class of organic-like symptoms in which there is a loss of function, but which lack the organic basis that explains them. They include paralyses, contractures of the limbs, anesthesia, inability to speak or understand language, loss of sight or hearing, and sometimes convulsions. Hysteria in this sense has nothing to do with the extravagant displays of emotion of popular language and is not related to gender. It is equally important to recognize that its symptoms are not feigned and that patients are not malingering.

Charcot

It was through the experimental work of Jean-Martin Charcot, the great nineteenth century neurologist and student of hysteria, and the astute analyses of hysterical symptoms by Pierre Janet, his psychologist colleague, that ideas were positively identified as determining the symptoms of hysteria. Charcot observed that the symptoms of hysteria corresponded in every detail with those that could be produced by direct verbal suggestion under hypnosis. The characteristics of muscular paralyses and associated loss of feeling generated by hypnotic suggestion, 'You are unable to move your arm,' for example, exactly matched those of hysteria. There was a similar correspondence between hysterical symptoms that followed real physical trauma, such as falling on one's shoulder, and those caused by indirect suggestion when a hypnotized subject was suddenly hit on the shoulder with moderate force.

Charcot proposed that both physical trauma and verbal suggestion in hypnosis caused the ego to lose control over the *realization* of the ideas called up by the sensations. Here he applied and extended the concept of unconscious mental processing first proposed as part of sensory-motor physiology and psychology by Laycock in the 1840s and by Bain and Carpenter in the 1850s. Sensory-motor physiology had it that a mover could make a movement only if the idea of it was in the mind beforehand. Realization was the largely unconscious process that enabled its execution.

Charcot incorporated the idea of a *loss* of movement and a *loss* of sensation into this schema. The patient's real fall and the blow to the hypnotized subject called up sensations of paralysis and anesthesia in exactly the same way as a blow causing a 'corked' or 'Charley horse' arm. The ego would ordinarily have prevented that idea being realized, but after trauma and in hypnosis its control was

weak or absent so that the unconscious idea of loss was realized as a symptom. The basis of the symptom was thus an idea, or as it was sometimes called, 'the imagination.' Several times Charcot made the point that this did not mean the patient was malingering: although paralysis might result from an idea or the imagination, the paralysis was not imaginary.

Janet

Janet's contribution was to notice that in hysteria and hypnosis it was the idea of the function or organ that was lost rather than those functions dependent on the innervation of particular muscles and receptors. Hysterical paralysis or anesthesia of the arm affected the everyday concept of the arm, that is, everything from the shoulder girdle down to the wrist. Janet explained this observation – the validity of which is quite independent of any of Charcot's observations – by assuming that the idea of the arm had become dissociated from normal consciousness. The sensations making up the idea of the organ or function had not been synthesized into a unitary perception or assimilated into the subject's idea of his or her self. Only with this act of what Janet called personal perception could the subject say 'I feel.' Dissociation followed a restriction of the field of consciousness during a traumatic experience, or during hypnosis, or when there had been prolonged inattention to the sensations.

Breuer and Freud

Freud adopted Janet's concept of the role of everyday ideas in determining the details of symptoms (initially giving him full credit for it and later claiming it as his own). He and Breuer based their explanation of symptom formation on the principle that the nervous system acted to reduce excitation impinging upon it by directing it into movement, including the movements of speech. An idea was always associated with a charge or quota of excitatory energy, later to be called cathexis. When this quantity was especially large and ordinary channels of discharge unavailable, it was converted into muscular or sensory symptoms. Normal discharge could not take place in hypnotic-like states (Breuer) or might not do so if intentionally prevented (Freud). The idea or function was then lost because its excessive affect prevented it from forming associations with other ideas.

Freud

By 1900 Freud had abandoned the theories of Breuer and the French and set out a different theory of how mental systems and the mental forces housed in them caused symptoms. The systems comprised Conscious (Cons.), Preconscious (Precons.), and Unconscious (Uncons.) minds. A mental force located (vaguely) in Cons. or Precons. opposed unacceptable ideas from entering Cons. and banished them to and kept them in Uncons. The ideas were motivated by drives and once in Uncons. were governed by a primary process: drives brooked no delay in their gratification, drive energy could readily be displaced from one idea to another or combined with it, and there was no sense of time or order and therefore no logic. On the other hand the systems Cons. and Precons. operated according to a rational and logical secondary process, oriented toward reality.

Freud derived the notion of a force in Cons. and Precons. from the failure of his patients to follow the trains of their associations to the memories of the events he believed had caused their symptoms. The failure occurred at precisely the point where Freud thought the associations were closest to the causal memory, and it was there that patients displayed the most resistance to further recollection. It was also the point where Freud had to exert the most pressure or force to overcome the patients' failure to recall. Repression was, Freud concluded, a force located within the patient's self or ego that corresponded to the resistance.

From the three systems and their mode of operation, Freud derived explanations for the formation of symptoms, dreams, and faulty actions (parapraxes), like slips of the tongue and pen, and forgetting proper names. All these mental products were compromises between the efforts of repressive forces seeking to keep unacceptable ideas from Cons. and forces in Uncons. just as determined to push their way in.

Freud's systems and the forces within them make up what is usually referred to as the topographic theory. Many of the functions carried out by Cons. and Precons. are similar to those in the theories of his

rivals. What makes Freud's distinctly different is his picture of mental life as resulting from conflict between mental forces. It is that which constitutes his theory as the first psychodynamic theory.

Clinical Unconsciousness

All five, Charcot, Janet, Delboeuf, Breuer, and Freud, agreed that an unconscious memory was responsible for symptom formation, but they obviously differed about the mechanism. Charcot and Janet both hypothesized that lost memory was in a dissociated or *subconscious* part of the mind, and Delboeuf would probably have located it in a sleep-like state that *alternated* with normal consciousness. Breuer suggested that symptoms formed in a hypnotic-like or *hypnoid* state, an explanation he formulated only after Janet's and Delboeuf 's cases were published. It is almost certain that his thinking was influenced by Freud's knowledge of the role Charcot and Janet attributed to ideas and to their notion of subconscious processes.

Freud himself initially agreed with Charcot, Janet, and Breuer about the roles of hypnoid states and subconscious ideas. Gradually he came to think that it was always a psychological force that banished ideas, which conflicted with the individual's standards, into the unconscious mind. Nevertheless, exactly like his contemporaries, Freud also had to postulate a constitutional tendency to form a secondary personality as a necessary basis for symptom formation. It is also a postulate common to the concepts of *co-consciousness* or *subliminal consciousness*, formulated by Morton Prince and F.W.H. Myers, which are not considered here. All share the same logical difficulties. For a set of ideas to be cut off from, but able to intrude into normal consciousness, disconnection and connection had to be present simultaneously. More, as Binet pointed out, if the secondary state was a rudimentary secondary personality, its characteristics could not be explained by the associationist psychology on which all drew.

There was a particular problem with Freud's view of the unconscious as a repository of unacceptable ideas. Some patients appeared to recall and experience as real his reconstruction of what he said had happened to them; others had no memory at all, agreeing, he said, only because "the context calls for them inexorably." Others again said they knew and did not know at the same time. These different reactions make the nature of the unconsciousness of Freud's psychodynamic theory difficult to understand.

Unconscious Ideas and Therapy

From about the mid-1850s unconscious mental processes were sometimes seen as being responsible for the removal of symptoms. Ellenberger described a whole class of what he called *magnetic diseases* in which this occurred. These very rare conditions were characterized by marked alterations of consciousness and included lethargy, catalepsy, somnambulism, and multiple personality. In those affected, mental processes were very different from normal, and resembled those produced by hypnotism, still then mainly known as *animal magnetism*. Although rare, magnetic diseases were known well enough through the popularization of authors like Eugène Scribe, whose play was utilized by Bellini in his opera *La sonnambula*, and E.T.A. Hoffmann in his stories – although not in those utilized by Offenbach in his opera. They were frequently cured by the patient leading the person conducting the treatment to a particular form of therapy in which the patient set the conditions for cure, often on a date they specified. Direct suggestion under hypnosis was often added to the therapeutic mix.

Janet's and Delboeuf's Therapies

Once the role of ideas in determining hysterical symptoms was clarified further, more rational therapies based on modifying the ideas underlying the symptoms were worked out. Charcot made no direct contribution to this development because he was decidedly skeptical about the therapeutic value of hypnosis, using it only to demonstrate experimentally how ideas could produce and remove symptoms.

The first of the more rational therapies appears to have been that used by Janet in late 1888 or early 1889. He treated the 19-year-old Marie for recurrent hysterical crises accompanied by deliria, hallucinations, and violent bodily contortions. The attacks began two days after the onset of each of

her menstrual periods and menstruation was suppressed during this time. Janet hypnotized her and found her first menstruation to have been an entirely unexpected event to which she reacted with shame. She made an attempt to stop the menstrual flow by immersing herself in cold water. Menstruation ceased, but she then had a severe attack of shivering followed by several days of delirium. Menstruation did not recur until five years later and, when it did, the symptoms came with it. Marie had minor hallucinatory attacks of terror, which were repetitions of feelings she had experienced after seeing an old woman fall down some stairs and kill herself. She also had a left-sided facial anesthesia and was blind in the left eye.

Janet then used what he termed 'the singular means' of using hypnosis to revive her memory of 'the initial conditions' and convince her that the idea that her menstruation had been arrested was absurd. The symptoms disappeared completely. Similarly her hallucinations were removed once Janet modified the idea she had formed at the age of 16 years that the old lady had died. Her blindness and anesthesia were removed by modifying her memory of a time when she was 5 years old and had been forced to sleep next to a child with a left-sided impetigo.

Josef Delboeuf, the Belgian philosopher and psychologist, made a similar application of the same therapeutic principle apparently independently of Janet. In early 1889 he published his successful cure of a woman whose repetitive hallucinatory visions reenacted her inability to come to the aid of her dying son. Delboeuf had his patient revive the memory of the traumatic event and then used direct hypnotic suggestion to 'combat it' while it was in 'its state of rebirth.' Both Delboeuf and Janet used hypnosis to return their patients to the state in which the trouble had first manifested itself; both claimed to have cured their patients by removing or modifying the causal memories and images.

Breuer's Therapy

Breuer's patient, the 22-year-old Bertha Pappenheim – pseudonymously Anna O. – helped Breuer develop a similar method. When Breuer first examined Bertha in 1880 for a nervous cough he immediately diagnosed her as mentally ill. She frequently stopped in the middle of a sentence to pause before continuing, seemingly unaware of these *absences* (as Breuer came to call them). It also soon became apparent that hers was a case of dual personality in which the transition from one self to the other took place in a well-marked late afternoon state of auto-hypnosis. She had a number of specific and variable hysterical symptoms including frightening hallucinations, partial aphasia, visual and auditory disturbances, paralyses, and contractures.

In the best tradition of magnetic illness, Bertha gradually guided Breuer into what she termed the *talking cure*. During the afternoon auto-hypnosis she would tell stories like those of Hans Christian Andersen, but with sad or tragic themes, which nevertheless lightened her mood when she regained normal consciousness. After some months of so talking, Bertha extended her guidance by telling Breuer about the circumstances under which she had acquired her symptoms. Again, she gradually led him to identify what he called the root of her illness in an event that had occurred while nursing her seriously ill father. Dozing, she had had a hallucination of a snake about to strike at him, but her arm had 'gone to sleep' and she had been unable to raise it to defend him. Nor could she call a warning in German; she could utter only an English prayer. Her hallucinations, her inability to speak in German while retaining her ability to speak and write in English, and her paralyses and contractures repeated her inability to act and speak at that time, and were elaborated into symptoms during her later absences. Breuer's published account had it that each of her separate symptoms was removed after she told the stories of the events during which they had formed, and that she was finally relieved of them by reenacting, on a day she nominated as the day of her cure, the root cause of the snake hallucination. That, Breuer said, ended the whole of her hysteria.

Nevertheless there were significant limitations to Bertha's recovery: Some were reported by Carl Jung in 1923, others by Ernest Jones in his 1953 biography of Freud. Fuller details emerged in 1972 after Ellenberger found a copy of Breuer's notes on Bertha. Albrecht Hirschmüller's 1976 complete

publication of the notes and Breuer's related correspondence reveal that Bertha's talking calmed her mood temporarily, but had little or no effect on her symptoms. In fact, at one stage while regularly using the method, her symptoms became so much worse that she was 'forcibly' hospitalized. Unlike the published version, Breuer's notes state only that she had been greatly relieved by the treatment.

Only 5 weeks after the treatment ended Bertha was admitted to a convalescent hospital, partly or mainly for the treatment of a morphine addiction caused by Breuer's prescription of the drug for her facial pain, but some of the same hysterical symptoms, including the inability to speak and understand German, were still present. She remained there for four and a half months. Over the next 5 years she was hospitalized three more times for a total of 10 months, each time usually showing remnants of her hysterical aphasia and possibly other speech disturbances, absences, hallucinations, and convulsions, as well as evening facial pains, which because of their variability may have been part of the hysteria, even though storytelling now made them worse. She did recover, but what brought that about is a total mystery.

Freud's Therapy

Despite the limitations of Breuer's treatment, Freud began by basing his own therapy on a variant of it that became known as *cathartic* therapy. Whereas Breuer's published and unpublished descriptions of Bertha's treatment say she merely talked about her symptoms in much the same way as she told her stories, Freud's variant had it that there was an accompanying discharge of the excess affect attached to those ideas. The recovery of the lost idea, and hence the effectiveness of the treatment, was because of that discharge or *abreaction*. Freud began to use this technique between 1885 and 1895 at a time when, in the light of what he had learned from Charcot and Janet, he seems to have pondered over what Josef Breuer had told him about Bertha.

Freud's first use of a purely cathartic method in the waking state was probably toward the end of 1892 with Lucy R. She was governess to the Viennese family of a widower with two children, and troubled by a recurrent smell of burnt pudding – an olfactory hallucination. She easily recalled a recent event after which the symptom formed. Her thinking about leaving her employment had been interrupted by a game with the children during which a pudding had burned. Lucy R.'s presenting symptom was its smell. On Freud's further inquiry she fairly easily recalled a much earlier event during which she felt she was falling in love with her employer.

Freud found it difficult to accept these as the memories responsible. By this time he conceptualized sets of memories being linked in a *pathogenic memory structure* in which at least one memory had to have been deliberately repressed. Neither of Lucy R.'s had been, and at Freud's insistence, she recalled an event in which her employer had criticized her severely for allowing a female guest to kiss the children. Realizing that this meant there was no possibility of marriage, she had therefore intentionally pushed the thought out of consciousness. With a little further prompting, Lucy R. then recalled a fourth event in which a male guest had been criticized for kissing the children. Cigar-smoke had been present at the time and even though she had not banished the memory, a transient minor olfactory symptom of the smell of cigar smoke had troubled her.

At the center of the structure, Freud placed Lucy R.'s deliberately repressed memory of being criticized over the female guest. It formed an unconscious core or nucleus to which the memories of the feeling of love for her employer and the criticism of the male guest were drawn, even though neither of them had been repressed. During her thinking about leaving, which was also not repressed, the affect associated with all three memories was pooled, and reached sufficient intensity to be converted into maintaining the simultaneously present olfactory innervation caused by the burned pudding, the way for which had been paved by the cigar-smoke.

Sexuality and Repression

Freud had originally formulated repression for explaining hysteria, but soon generalized its scope to what he called *psychoneuroses*. In obsessional

neurosis, for example, the affect was displaced from the unacceptable idea to a neutral one. He made this and similar generalizations in a context partly determined by the sexual theory of nasal neurosis, formulated by his Berlin friend and colleague, Wilhelm Fliess. Freud set out to establish an exclusively sexual etiology for the *actual neuroses* of neurasthenia and anxiety neurosis. In doing so, he applied Koch's postulates, used for identifying the causes of bacterial diseases, in such an incomplete and methodologically indefensible way that his conclusion that they had sexual causes was unwarranted. All it did was to generate a false expectation that he would find sexual causation in the psychoneuroses when he later investigated them.

Perverse Sexuality, Determining Quality, and Free Association

In Charcot's patients and experimental subjects, the sensations experienced in the trauma were repeated in the symptoms. Freud called this property *determining quality* and adopted it as the criterion for identifying the traumatic memories at the core of the pathogenic memory structure. Symptoms containing oral, anal, and genital sensations, therefore, had to be based on memories of frankly sexual experiences in which those zones of the body were stimulated. Given that Freud wanted to establish sexual causes for the psychoneuroses, it is not surprising to find him placing enormous pressure on his patients to recall events possessing those kinds of sexual content.

The main outcome of Freud's pressure was the seduction theory, or more correctly, the seduction hypothesis. He came to believe that brutal perverse sexual assaults involving those zones had been perpetrated on his patients during childhood, usually by an adult, and sometimes by their fathers. He therefore expected his patients to have memories of them. Some did so, but for the most part the causal 'memories' were reconstructions he made from the very fragmentary recollections of his resisting patients. For those patients unable to recall sexual traumas of any kind, only the inexorable demands of the context forced them to agree with Freud.

In arriving at the seduction hypothesis Freud discounted his influencing the patients' recollections. To understand this we need to consider the assumptions underlying the method by which he gathered his data, that of free association. When patients thought in a nonpurposive way about their symptoms Freud required them to report with absolute honesty everything that came to mind, including thoughts that intruded, however irrelevant. Freud's application of what he termed, 'the fundamental rule of psychoanalysis' depended on trains of thought being psychologically continuous – gaps in them being filled by psychological and not physiological processes – and of them being guided by unconscious ideas. During free association, all the ideas that were produced, including those that intruded, were under the guidance of unconscious ideas related to the symptom. In one of his earliest implicit references to the rule, Freud said it was demonstrably untrue that trains of thought were ever purposeless; when conscious purposive ideas were abandoned, unconscious ones simply took their place. This was the thesis proposed explicitly by Meynert and von Hartmann and it gave no place to unconscious suggestion.

Not long after leaving the Salpêtrière, Freud had vigorously defended Charcot against the charge that many of the phenomena of hysteria and hypnosis produced there were products of unconscious suggestions transmitted to the patients and subjects. In believing that his patients' trains of thought were similarly uninfluenced, Freud was wrong. External factors like unconscious suggestion could not be disregarded: the determinants of what was produced were not completely inside the patient.

In addition to many of the seduction 'memories' being reconstructions, it is sometimes clear from Freud's reports that he imposed them on the patients. This is already evident in the pre-seduction hypothesis case of Elizabeth von R., on whom he practically forced the idea that her painful leg pain was the result of her repressing the idea that she loved her brother-in-law. The later post-seduction hypothesis patient, known as the Wolf Man, resisted until the end of his life, Freud's reconstruction of the root cause of his problems as his witnessing his parents' coitus *a tergo* when he was only 18 months old. What actually happened in these and other cases is difficult to establish. In the case of the Rat Man, for example, on whom some of Freud's case notes have survived, the discrepancies between the

notes and what Freud published make it impossible to know anything reliable about his reconstruction of the supposed threat made by the Rat Man's father.

A Sexual Instinctual Drive

The seduction hypothesis required that the traumatic childhood experiences at the nucleus of the neuroses be of perverse sexuality. At the beginning of 1898, when Freud repudiated the reality of these 'memories' because he no longer believed in them, he was left without a source for the sexual sensations in the symptoms. His solution was to propose a childhood sexual instinctual drive with *component drives* that generated the perverse sensations in the oral, anal, and genital zones of the patient's own body, that is, *autoerotically*, and which gave rise to fantasies of seduction. The only direct observational evidence Freud had for any component was for an oral drive and it was an unwarranted sexual interpretation of the 'sensual sucking,' described by the Hungarian pediatrician Samuel Lindner. In his study of mainly thumb and finger sucking in 500 children, Lindner described four 'exultant' suckers who went into a 'rapture.' Freud claimed that Lindner recognized pleasure-sucking in general and the rapture specifically as sexual, even though Lindner made no such claim and his descriptions do not warrant that interpretation.

Once proposed, Freud's hypothetical childhood sexual drive helped explain what he saw as the singularity of repression in being directed only at the memory of sexual events. He supposed that in remote prehistory the stimulation of all three zones produced sexual pleasure. During evolution, when humans adopted the upright stance, mental forces of disgust, shame, and morality emerged and reversed the affect. Stimulation of the zones now generated repellent sensations and the affect that was repression. Because Freud believed that individual development recapitulated that of the species, the same mental forces automatically caused repression when the zones were stimulated during childhood.

Sexuality and the Unconscious

There were two main theoretical consequences of Freud's introducing a childhood sexual instinctual drive: repression became a two-stage process, and the unconscious became more than a repository of unwanted thoughts.

Two-Stage Repression

In Freud's new conception, *primary repression* was the first of the two stages and took place when the component drives lying dormant in the phylogenetically abandoned but still-sensitive erotogenic zones were stimulated and released unpleasure. The other was *secondary repression* and took place much later when and if direct or indirect stimulation of those zones again caused unpleasure. As well as explaining why repression acted only on sexual memories, yesteryear's revived phylogenetic memories provided a nucleus around which today's repressed ontogenetic experiences could be grouped.

Of each kind of repression the question may be asked: What *is* it? Obviously Freud tells us what each *does*, but he leaves us ignorant of *how* each produces its effects. Both repressions are therefore uncharacterized theoretical terms, lacking properties of their own. The defect is already found in Freud's earlier and simpler concept: How is an idea stripped of its affect? Converted into muscular innervation? Displaced on to another idea? Similarly, what brings about a reversal of affect?

Another and seemingly fatal objection to both the old and the new concepts is that an idea is lost to consciousness when stripped of its affect but finding it and abreacting it requires the affect still to be present. How can that be? Similarly, the adult perverse activity Freud allowed for, only took place if the sexual instinctual drive was strong enough to override the unpleasure, especially when the 'average uncultivated woman' was led to it. What conditions allow overriding? How is the strength of the overriding force measured independently of its effect?

Drives and Mental Structures

Postulating a sexual drive in childhood required Freud to propose a second drive, one strong enough to effect the repression of sexuality at a time when the ego and its standards were necessarily weak. Freud had always based his explanations on the conflict between two mental forces as,

for example, in the opposition of a conscious system to an unconscious one, and it is not surprising that he now chose an ego-instinctual drive to give the ego the energy to control the sexual drive. The ego-instinctual drive also accounted for the development of a sense of reality, one that found the real objects that satisfied the child's real needs rather than its hallucinated images of them.

By 1923, Freud had drastically modified even this duality and replaced it with another. The ego and sexual drives were fused into a life-instinctual drive, termed Eros, which was challenged by the destructive force of a death-instinctual drive, termed Thanatos. Freud advanced two kinds of arguments for this conceptualization. First, he reinterpreted a number of behaviors as seeking unpleasure rather than avoiding it. Thus, whereas he had originally seen repetitions of unpleasurable traumatic dreams as attempts to master the traumas, he now said they were instances of a repetition compulsion. Similarly, when patients transferred on to the analyst their childhood feelings toward their parents, they were repeating painful experiences that had never brought satisfaction. Transference was now under the death instinct and very far from an attempt to avoid unpleasure. To these examples Freud added the repetition of what he took to be an unpleasant experience in his grandson's play. His 'maturer reflection' now allowed him to conclude that all these behaviors were motivated by the death instinct. Freud also advanced a complex speculative argument that he believed supported the existence of a biologically based instinctual drive that unconsciously sought the individual's death.

Structures and Unconsciousness

These theoretical reconsiderations had consequences for the topographic theory. Without Uncons. where were the new instincts housed? Freud again reinterpreted an early observation, this time on resistance, and drew on an apparently new one. The reinterpretation was that the resistance to recovering associations was unconscious. Paradoxically, although Cons. and Precons. were the organs of consciousness and repression, they were now required to have unconscious components. Freud's apparently new observation was of his patients getting worse precisely after he praised them for improving. Calling the deterioration a 'negative therapeutic reaction,' Freud now attributed it to an unconscious sense of guilt motivated by Thanatos. Even though he had previously said that the sense of guilt was conscious, even 'overstrongly' so in conditions such as depression, he now said it was unconscious.

Unconsciousness and the Structural Theory

Clearly neither Thanatos nor Eros could be easily housed in Cons., Precons., or Uncons. At any level of consciousness their manifestations would have had to be logical and rational, and it would be just as difficult to conceptualize an irrational and illogical system providing organized opposition to them. Freud's solution was to propose three new mental structures – the ego, superego, and id – and house derivatives of the new instinctual drives in them irrespective of whether the ideas they motivated were conscious or unconscious.

The ego drew on a neutral form of energy derived from Eros and had the responsibility for cognitive functions, such as perception, attention, memory, motor control, and for judgments of reality. It was the only structure in which consciousness arose and in which anxiety could be experienced. The superego formed out of the ego and was powered by parts of Thanatos. It contained the individual's moral standards, a conscience that scrutinized behavior against them, and an agency that punished infringements of them. The id was the original energy store. It contained a fused form of Eros and Thanatos and was governed by the primary process. The assumption of fusion was, Freud said, "indispensable to our conception;" only by combining with Eros could the self-destructive aim of Thanatos be deflected from the individual. Fusion, in turn, required the further assumption of defusion, one that "forces itself upon us." Only defused drives could provide energy appropriate to the structures.

Some degree of unconsciousness attended each of the new structures: the id wholly so, but the ego and superego functions most closely connected with it were also partly unconscious. Freud now decided that theoretical emendation was necessary: the conscious or unconscious status of an

idea did not depend on its belonging to systems like Cons., Precons., or Uncons. Granting he had 'managed pretty well' with the ambiguity of both repressed and preconscious ideas being 'unconscious,' he now restricted the terms conscious, preconscious, and unconscious to the descriptive. The structural theory, as it became known, thus completely supplanted the topographic theory with its three systems and instinctual driving force.

The Oedipus Complex and the Structural Theory

During the formation and resolution, a universally present Oedipus complex changes in instinctual drives gave each of the new structures the energy appropriate to it. One part of the complex was based on the child's nonsexual primary identification with its parents: the child wanted to be like them. The other part built on the child's sexual object choice: it wanted to have the parents as sexual objects. This object choice was made only after Eros could be directed from the autoerotic zones on to real objects. The theoretical representation of the complex was, Freud said, "much more difficult" than its description.

There is a similar theoretical complexity about what happened next. Primary identification was a change in the child's ego such that it resembled its parents. In so doing it became an object of Eros, and the change sublimated the energy of Eros into a nonsexual form that then powered the ego's nonsexual and cognitive functions. At the same time the two parental sexual objects were incorporated into the child's ego where their capacity to satisfy Eros made them targets of Thanatos. There were thus positive and negative complexes. In the positive, Eros was directed at the mother-object and Thanatos toward the father; in the negative form the development was reversed.

Castration Anxiety

A universal threat of castration now caused the development of moral standards, a conscience, and a punitive function. Even when not made aggressively, the threat of castration was always experienced against the trace of a phylogenetic memory of a real but historically remote castration. Anxiety generated by the threat caused the forcible withdrawal of the sexual drive from its objects, and all the parental identifications, together with their functions, to be taken into the ego. Thanatos turned inwards as self-hate and powered the superego that formed as a precipitate of the incorporated parental functions. The identifications so taken over were then united "in some way," said Freud, to confront the rest of the ego with Thanatos.

This remnant ego of the structural theory had lost the capacity to repress, which it had had in the topographic theory. In exchange, it gained the ability to experience anxiety. When anxiety signaled an internal threat, it avoided the danger by bringing the superego's repressive forces into action and escaped the external threat through its cognitive and motor functions. Although, Freud said this ego was now only "a poor creature" menaced by dangers arising from the id and superego within and from the external world, it could still defend against both.

Freud's and Other Psychodynamic Theories

We have seen that Freud first thought of unconscious mental processes in a way not very different from his contemporaries. The systems of Freud's first psychodynamic theory comprising an unconscious mind driven by sexuality, and operating with an illogical primary process that a secondary conscious process attempted to control, was completely different. The structural theory with its three agencies of id, ego, and superego was even more so.

Freud's topographic theory also differed markedly from the two main psychodynamic theories which followed more or less immediately from his. Alfred Adler's was the first. He rejected instinctual drives as important determinants of behavior, and with that the concepts of repression, an unconscious mind, and a universal Oedipus complex. He postulated an innate but nonbiologically based drive to mastery that drove development from the individual's initial sense of inferiority toward a self-perfection that overcame it, the behavior of which the individual was unaware did exist, but it was not in the unconscious mind or caused by or related to repression. If an Oedipus complex developed, it was

because of overindulgence, usually by the mother, which prevented normal development.

Carl Gustav Jung's was the second. It was built around an instinctual drive possessing a nonsexual, general form of energy (confusingly called 'libido') located in a collective unconscious that also housed the primordial images or archetypes that regulated it. Repressed and merely forgotten material was located in a personal unconscious. Jung's concept of libido had no autoerotic direction; the moral standards of the ego were innate and drew their energy from part of the nonsexual libido. Jung rejected Freud's understanding of the Oedipus complex, especially of its universality, and the notion that the superego was descended from it.

Adler's and Jung's theories have the same conceptual problem as Freud's. How can theories based on a single motivating force be validated? Can a choice be made among them? Appeal cannot be made to direct observation because all are based on interpretations and reconstructions, which are neither simple, nor direct, nor rule governed. Consider the different interpretations of delusions of grandeur to which each of the three lead: Freud's and Jung's in the case of Schreber, and Adler's in Nijinsky's. Freud said Schreber's delusions came about when his libido was withdrawn from its real-world objects and directed on to the ego. He thereupon lost his sense of reality and his ego became his whole world. Jung said that the whole of Schreber's libidinal drive (as he conceived it) had withdrawn from reality, not just the form affecting his sexuality. Adler proposed Nijinsky's drive had overcome his great sense of inferiority, and had taken him to a grandeur that denied him in real life. How could a choice be made?

Can choice be made among modern psychodynamic theories of the unconscious? To the extent that they are based on similar kinds of evidence, interpretations, reconstructions, and theorizing as Freud's, it cannot. As an illustration, consider the basic choice between Freud's topographic and structural theories. Curiously, although Freud clearly believed that the structural theory had supplanted the topographic, many modern psychoanalytic theorists have made the opposite choice, and therefore of concepts of unconsciousness different from Freud's. Several broad strands of thought, none of which are completely independent, can be identified among those making this choice. There is a substantial group, among them David Rapaport and George Kline, who reject the structures completely on the grounds that they are too abstract. They revert to something like the topographic theory, including an unconscious as a system, because they say it gives a more meaningful representation of what happens in treatment. On the other hand, Joseph and Ann-Marie Sandler revised Uncons. precisely because it needed to be brought more into line with clinical experience. Some reject the structures completely or modify them significantly (Roy Schafer), or the functions Freud assigned them, as does Susan Millar in requiring the superego to love the ego, or like Heinz Kohut incorporate all the functions into a unitary self.

Different concepts of unconscious necessarily follow from the strand that rejects some or all of the instinctual drives (Edward Bibring, Charles Brenner, and Robert Holt), especially of Thanatos (Ernest Jones and David Werman), or the vicissitudes of the instincts (Heinz Hartmann). A related rejection having profound effects on what is to be considered unconscious is of what Freud termed the shibboleth of psychoanalysis – the Oedipus complex – as a phenomenon (Bennett Simon and Jay Greenberg) because it develops quite differently and relatively late (Melanie Klein).

How can a variation like this among psychoanalytic theories of unconsciousness be explained? If free association is not influenced by the analyst's unconscious suggestion, as Freud insisted, each analyst should, broadly speaking, gather the same data. Occasionally one or another modern analyst grants or hints that this is not so (Jacob Arlow and Charles Brenner, Jeremy Nahun), but none challenges the deterministic assumptions basic to the method. If the data are the same, variation in theory must then be due to their being interpreted differently. Studies of the interpretation of dreams and symptoms by different psychoanalysts clearly show there are great differences, and that is also true of interpretations of more general clinical phenomena such as transference. To the extent that reconstructions are complex interpretations, it is true of them also. Can differences like these be reconciled when there are no rules against which interpretations and reconstructions can be judged?

Conclusion

Without rules for guiding interpretations and reconstructions every psychoanalyst can formulate his or her own psychodynamic theory. Following Freud's example, many do so. Freud based many elements of his concepts of unconscious mental processes on such idiosyncrasies as inferring repression from his feeling of personal effort, arguing for a death-instinct based repetition compulsion by simply reversing earlier interpretations and giving more mature thought to them, making indispensable assumptions the basis for instinctual fusion and defusion, and abandoning the topographic theory because of the need to do away with ambiguities with which he previously managed well enough. Feelings and interpretations as idiosyncratic as these are found in modern theorists.

How then did Freud's theory win out over its rivals? There is not much hard evidence, but several factors seem likely to have contributed. First, psychoanalytic concepts seem to provide unlimited explanatory scope and are not limited to pathological conditions. Second, they have what Wittgenstein termed 'charm.' Despite unconscious mentation allegedly being governed by an alien primary process, every unconscious motive, drive, or process is understandable in the same way as its directly experienced conscious counterpart. Third, explanations can be derived from them as readily as explanations of everyday behavior. Without specific training, everyone can apply them and be their own analyst in generating plausible interpretations of their own and their friends' behavior.

Social influences having little to do with validity, but much to do with social movement psychoanalysis cannot be overlooked. The movement came about largely through the proselytizing of A.A. Brill, Ernest Jones, and Freud's controlling of a doctrine through a special committee drawn from among his closest disciples – one that Phyllis Grosskurth has called the 'Secret Ring.' No other depth psychology had that kind of founding impetus, centralized control, or appeal to everyday explanations to maintain that impetus.

Recasting any of the psychodynamic theories of unconscious mental processes into other theoretical frameworks is possible, but can be useful only if there is agreement on the validity of the particular theory being recast. Were those conditions met, a nice illustration of what might be achieved is provided by the radical minimalist connectionist model, Lucynet, implemented by Dan Lloyd, to simulate Freud's case of Lucy R. Lloyd accepted Freud's report at face value, a trust that might be misplaced given his other reports. Lucynet generated Lucy R.'s symptoms without a memory archive, an unconscious, or an explicit function for repression. However enthusiastically cognitive psychologists might greet such a recasting, it is difficult to see that many psychoanalysts would.

Other attempts to reformulate other psychodynamic theories may be more difficult than Lloyd's. Almost all their data and therefore the bases of their interpretations remain hidden under the privileged couch. Consequently the evidence to decide which is worthy of recasting derives from the testimonies of two consciousnesses: the introspecting patient's and the interpreting psychoanalyst's. May Alfred Binet's warning of nearly 120 years ago be disregarded or is suspicion still justified?

See also: Perception: Unconscious Influences on Perceptual Interpretation; Unconscious Cognition; Unconscious Goals and Motivation.

Suggested Readings

Borch-Jacobsen M (1996) Remembering Anna O.: A Century of Mystification. In: Olson K, Callahan X, and Borch-Jacobsen M (trans.). New York: Routledge (Original work published 1995).

Breuer J and Freud S (1895) Studies on hysteria. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 2, pp. 19–305. London: Hogarth.

Cioffi F (1998) Freud and the Question of Pseudoscience. Chicago and La Salle: Open Court.

Ellenberger HF (1970) The Discovery of the Unconscious. New York: Basic Books.

Esterson A (2001) The mythologizing of psychoanalytic history: Deception and self deception in Freud's accounts of the seduction theory episode. History of Psychiatry 12: 329–352.

Freud S (1900) The interpretation of dreams. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vols. 4 and 5, pp. 1–621. London: Hogarth.

Freud S (1916–1917) Introductory lectures on psychoanalysis. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 15 and 16, pp. 9–496. London: Hogarth.

Freud S (1920) Beyond the pleasure principle. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 18, pp. 7–64. London: Hogarth.
Freud S (1923) The ego and the id. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 19, pp. 12–59. London: Hogarth.
Freud S (1933) New introductory lectures on psycho-analysis. In: Strachey J (ed.) The Standard Edition of the Complete Psychological Works of Sigmund Freud, vol. 22, pp. 5–185. London: Hogarth.
Hirschmüller A (1989) The Life and Work of Josef Breuer: Physiology and Psychoanalysis. New York: New York University Press. (Original work published 1978).
Klein DB (1977) The Unconscious—Invention or Discovery?: A Historico-Critical Inquiry. Santa Monica, CA: Goodyear.
Lloyd D (1998) The fables of Lucy R.: Association and dissociation in neural networks. In: Stein D and Ludik J (eds.) Neural Networks and Psychopathology: Connectionist Models in Practice and Research, pp. 248–273. Cambridge: Cambridge University Press.
Macmillan M (1997) Freud Evaluated: The Completed Arc. Cambridge, MA: MIT Press.
Macmillan M (2003) Challenges to psychoanalytic methodology. In: Chung MC and Feltham C (eds.) Psychoanalytic Knowledge and the Nature of Mind, pp. 219–238. London: Palgrave.
Macmillan M and Swales PJ (2003) Observations from the refuse-heap: Freud, Michelangelo's Moses, and psychoanalysis. American Imago 60: 41–104.
Timpanaro S (1976) The Freudian Slip: Psychoanalysis and Textual Criticism. In: Soper K (trans.). London: NLB (Original work published 1974).
Whyte LL (1960) The Unconscious Before Freud. New York: Basic Books.

Biographical Sketch

Malcolm Macmillan graduated with DSc from Monash University (1992), MSc Melbourne University (1964), and BSc from the University of Western Australia (1950). Since 2005, he has been a fellow of the Academy of the Social Sciences in Australia, and during 2004–05 was the president of the International Society for the History of the Neurosciences. His visiting positions include fellowships at Oxford University and the University of Pittsburgh.

Macmillan was a founding member, then fellow (1988), and life member (2005) of the Australian Psychological Society. He was also a founding member of the College of Clinical Psychologists of the APS. Since 1991 he has been a fellow of the Association for Psychological Science. In November 2006, he became one of the three coeditors of the *Journal of the History of the Neurosciences*.

Macmillan is the author of the prize-winning *An Odd Kind of Fame: Stories of Phineas Gage* (MIT Press), and his *Freud Evaluated: The Completed Arc* (MIT Press) has had considerable critical success. He is also the author of some 90 papers, book chapters/encyclopedia entries, and book reviews, and presenter of over 30 conference papers. Currently he is evaluating the work of Alfred Walter Campbell and comparing Campbell's cytoarchitectonic research with Brodmann's.

Psychopathology and Consciousness

S A Spence, University of Sheffield, Sheffield, UK

Glossary

Cognition – Strictly speaking, this means thinking (as in Descartes': *cogito ergo sum*), but the word tends to be used to describe the processing of information by the brain: for example, attention, concentration, and different forms of memory.

Compulsion – A behavior linked to an obsession, the submission to which the subject resists, the prelude to which is increasing anxiety and tension, and the performance of which may lead to temporary relief, for example, compulsive hand-washing.

Delirium – A global impairment of cognitive function, usually acute in onset, fluctuating in intensity, and potentially treatable. Consciousness is characteristically impaired ('clouded').

Delusion – An abnormal belief based upon abnormal inference, incorrigible, and out of keeping with the subject's social, cultural, and religious background.

Dementia – A global impairment of cognitive function, usually chronic in onset, progressive, and irreversible. Consciousness is initially preserved ('clear').

Hallucination – A percept in the absence of a veridical external stimulus, over which the subject has no control.

Hysterical conversion – An unconscious mental conflict is hypothesized to have been resolved through its conversion into a physical symptom, for example, paralysis, for which an explanatory physical cause cannot be found.

Illusion – A misperception of a veridical external stimulus, more frequent in low stimulus intensity environments (e.g., darkness) or when the subject is distracted.

Imagery – Mental percepts (in the absence of external stimuli), over which the subject exerts control, and which they recognize as imagined (e.g., in the mind's eye).

Neurosis – A mental condition, the symptoms of which are exaggerations of those present in normal mental life, for example, anxiety.

Obsession – A recurrent, intrusive thought, usually distressing in content, which the subject recognizes as his own, but which he tries to resist. The subject matter he rejects as being incongruent with his identity (ego-dystonic).

Pseudodementia – An apparent global impairment of cognitive function, subacute in onset, during which the subject fails tasks out of an expressed belief that they cannot do them (they tend to 'give in' too soon); associated with depression or malingering.

Psychosis – A mental condition in which the symptoms present are not held to be features of normal mental life, for example, delusions.

Introduction

Psychopathology concerns the study of abnormal states of mind and their correlates in abnormal patterns of behavior. The term has been used in a purely descriptive sense in the context of phenomenological philosophy, as an aid to interpretation in psychotherapeutic approaches to the mind and, more recently, in an experimental context, most notably in the pursuit of the neural correlates and biological underpinnings of abnormal symptoms and syndromes. Psychopathology necessarily invokes the experience of a conscious subject and the attempts of an observer to understand and describe that experience

as accurately as possible. Nevertheless, psychopathology is prone to the limitations associated with trying to determine what another person is thinking: the third-person 'observer' can never know what the first-person 'subject' is experiencing; much depends upon inference.

This article describes the main symptoms of mental disorders with particular reference to first-person experience and third-person observation. We offer definitions of key psychiatric terms such as hallucination, delusion, delirium, and dementia. Where the biology of such experiences is understood this is touched upon. Throughout, emphasis is placed upon the accurate understanding of terminology, realistic appraisal of clinical contexts, and the limits upon what may be factually stated regarding another's mental state. Finally, we consider interactions where communication may be deliberately subverted.

What Is Psychopathology?

Psychopathology is the study of the experience and expression of abnormalities of mind. Hence, it inevitably involves a dynamic interaction between the person and their surroundings, the subject (or patient) and their witness (the observer).

Psychopathology has been approached in three different ways over the last 150 years:

1. *Descriptive psychopathology* involves the careful description of mental states of closely observed human experiences, without recourse to explanation (this has been called 'phenomenology' and is exemplified by the classic contributions of Karl Jaspers). This form of psychopathology is a prerequisite to accurate diagnosis; it requires time spent listening to subjects; and, by definition, the mental phenomena examined are conscious: they comprise what the subject experiences (and can describe). The formal, structured description of mental symptoms probably began in the 1860s.
2. *Interpretative psychopathology* also involves the careful description and understanding of human subjective experience, but seeks to explain it mainly through recourse to hypothesized unconscious mechanisms (this has been called psychodynamic psychopathology and is exemplified by the classic contributions of Sigmund Freud). Hence, much of what might be discussed in psychodynamic psychopathology concerns what the subject herself is unaware of (that which is 'unconscious' and initially unknown to her); part of the purpose of therapy is to understand recurring patterns of behavior that might, until then, have eluded recognition.
3. *Experimental psychopathology* has sought to understand subjective phenomena mainly in terms of brain processes; it currently constitutes the dominant paradigm and might subsume a host of investigators using a variety of physical methodologies to examine the biological processes supporting disordered mental life (e.g., Wilder Penfield's pioneering use of direct electrical stimulation of the human cerebral cortex to elucidate the focal correlates of subjective experience (voices were heard by subjects when their temporal lobes were stimulated); or the discovery of the huntingtin gene by the Huntington's Disease Collaborative Research Group in 1993, the latter indicative of a similarly mechanistic approach, albeit applied at a radically different level of explanation: genes rather than anatomically specified cerebral systems).

Hence, contemporary psychiatric practice requires the ability to retain an understanding of, and relationship with, individual human subjects, to help them attain autonomy (i.e., to effect change in the real-world) while simultaneously incorporating into their clinical care knowledge gleaned from experiments conducted at many different levels of mechanistic understanding (from genes to communities). This may precipitate tensions and, indeed, within scientific academic psychiatry, there is currently a search for endophenotypes, brain markers that may be objectively diagnostic of specific neuropsychiatric disorders, which could effectively circumvent the need for prolonged, detailed description of individual mental states. These approaches raise ethical issues that have yet to be fully explored.

What Is a Psychiatric Diagnosis?

Most psychiatric conditions comprise syndromes, in other words they are characterized by groups of symptoms that tend to run together. Such syndromes are clearly described in operational terms in documents such as the Diagnostic and Statistical Manual (DSM) of the American Psychiatric Association and the International Classification of Diseases (ICD) of the World Health Organization (WHO). Hence, in schizophrenia (ICD 10, WHO, 1992, category F20) it would not be unusual to encounter a patient who believes that he is being spied upon, that others are reading his mind from a distance, and that there are persecutors who speak to each other about him, whom he can hear (''he is bad, look at the way he holds his fork, he is evil''). These phenomena (persecutory delusions, thought broadcast, and third-person auditory verbal hallucinations, respectively) would satisfy criteria for a diagnosis of schizophrenia, assuming that they had lasted for more than a month and that they occurred in the absence of any other processes (such as cannabis use, F12, or gross brain disease, F06) that might have provided plausible explanatory alternatives.

It follows that most psychiatric diagnoses essentially describe subjective states. In schizophrenia (F20) there is often a theme of persecution, in dementia (F00-F03) the subject is losing awareness of themselves and their surroundings, in depression (F32) they are suffering despair, but in all these states the diagnosis is quintessentially concerned with current consciousness, with what it is like to be that person at that time. For the most part then psychiatric diagnoses differ from many other forms of medical diagnosis, where there are objective, biological pathologies that constitute the evidential basis for diagnosis, and of which the subject need not necessarily be aware: a woman with an ovarian tumor may not know that she has it until it causes pain, a man may have advanced coronary vascular disease long before he is aware of his condition. Hence, these pathologies, these diagnoses, may be 'silent' in a way that psychiatric disorders have not been said to be (to date). This is not to say that things might not change, but it does call into question what one might be treating in a psychiatric patient who did not have symptoms. The treatment might be medical, but would it still be psychiatric?

Of course, there are exceptions on either side of the dualistic (and outmoded) body/mind divide: several 'physical' diagnoses lack a consistent pathophysiology and essentially comprise descriptions of symptoms (e.g., irritable bowel syndrome, fibromyalgia, chronic fatigue syndrome) and it is noticeable that these too constitute descriptions of conscious states that could not be said to occur 'silently' (indeed, they also often cross over the hinterland, into psychiatry, when a sufferer is offered a psychotherapeutic intervention). Conversely, there may be 'psychiatric' disorders, that, while asymptomatic, might be said to be present (in some sense) before their subject knows of them (e.g., in the undiagnosed carrier of the huntingtin gene (F02.2), or the person developing Alzheimer's disease pathology (F00) before their memory impairment is apparent). Nevertheless, a manifest psychiatric disorder seems to require a conscious subject for its expression: we do not currently speak of 'silent schizophrenia.'

Dissecting Phenomenology

Traditionally, psychopathological symptoms (experienced by the subject) and signs (witnessed by the observer) have been described, organized, and addressed according to defined phenomenological categories, for example, the perceptual disturbance of the person with schizophrenia (F20) who hears 'voices' contrasts with the memory disturbance of the person developing Alzheimer's disease (F00). While such a distinction is feasible for certain, discrete diagnostic entities (e.g., Korsakoff's syndrome, F10.6, can occasionally comprise solely a disturbance of memory, thereby constituting a pure amnestic syndrome), most of the time the disorders patients present with affect more than one domain of mental life, for example, the man with schizophrenia may hear voices (a disturbance of perception), while believing that his thoughts are not his own (a disorder of thinking), and feeling 'nothing' (a disturbance of emotion). So, while it might be helpful to categorize disordered phenomena according to discrete phenomenological domains, most neuropsychiatric syndromes impact

more than one domain; hence, a comprehensive account of their pathophysiology would necessarily be required to explain a diverse range of phenomena.

Helpful Dichotomies

Some traditional distinctions have gone in and out of fashion, but may still be useful when thinking about neuropsychiatric disorders. Indeed, they are still employed clinically.

Psychosis versus neurosis: a distinction has been made between psychoses, conditions in which phenomena are experienced that are not present in the normal mental state, where a subject 'loses contact with reality,' and neuroses, which may be understood as incorporating exaggerations of 'normal' mental experience. A good example of the former is schizophrenia, the latter, generalized anxiety disorder. While a so-called neurotic patient may elicit considerable understanding from an interlocutor, because the experience they describe (e.g., anxiety) is not qualitatively different from normal states of consciousness, the situation for the patient experiencing a psychotic illness can be more troublesome: he may be trying to describe phenomena that are totally alien to his listener; indeed, the situation he finds himself in may well be beyond his powers of description (e.g., the man who feels thoughts entering his head through his skull and can point to their site of entry). The psychotic/neurotic distinction provides a useful shorthand for describing patients' situations, although it must be acknowledged that there are limits upon its universal applicability: some neurotic phenomena may be so extreme that few healthy people would really share an insight into what it might be like to experience them (imagine the thoughts of the woman suffering from obsessive–compulsive disorder, F42, who washes her hands throughout the night, with soap and bleach, until her skin splits and her fingers bleed, but she cannot stop because of her fear of contagion); conversely, although hearing voices has long constituted the layperson's notion of the quintessential symptom of madness, more recent epidemiological research suggests that voices are heard by more people than see psychiatrists (or even tell their doctors about them); such phenomena may be common in the context of bereavement, sensory deprivation, or significant trauma. They may also arise commonly in the context of falling asleep or waking up (so-called hypnagogic and hypnopompic hallucinations, respectively), when a subject might hear their name being called. Hence, the neurotic may deviate markedly from normality and those who regard themselves as 'normal' can experience something of what it might be like to be psychotic.

Functional versus organic: Historically, it was always apparent that certain mental disorders were more obviously associated with gross (i.e., obvious) pathology of the central nervous system (CNS) than others. In the eras of Hughlings Jackson, Broca, Wernicke, Charcot, Janet, Freud, Kraepelin, Alzheimer, and many of the other great, early neurologists, neuropathologists, and proto-psychiatrists, it was clear that certain mental disorders (such as general paralysis of the insane (GPI), a late complication of syphilitic infection) had visible correlates in the brains and bodies of those afflicted (in life; though confirmation ultimately arrived at postmortem), while other conditions did not seem to exhibit such correlates (e.g., dementia praecox, the diagnosis that eventually became known as schizophrenia). What emerged was an understanding that certain disorders involved gross organic (usually structural) disturbance of the brain, while others were thought to reflect more subtle functional impairments (which might be understood to be psychological or neurochemical in nature). Hence, schizophrenia (F20) and manic-depressive psychosis (which came to be known as bipolar disorder, F31) were regarded as functional psychoses in contrast to GPI or encephalitis lethargica, which could cause organic psychoses (F06). There are now a number of problems with this distinction. One is that, over time, organic correlates of functional brain disorders such as schizophrenia have become well recognized (e.g., subtle enlargements of the fluid-filled ventricular spaces within the brain, and atrophy of the frontal and medial temporal lobes). Another problem is that the functional/organic distinction has been borrowed to refer to an entirely different set of contrasting conditions. In this second scenario, the patient is one who exhibits unusual symptomatology, where there is an element of doubt concerning the veracity of his symptoms. In conversion

hysteria (F44.4) a man might be unable to lift his arm, yet a biological cause cannot be found. The Freudian, interpretative psychopathological approach to this problem is to look for the causes of his symptom, specifically unconscious psychological causes: Is he unable to lift his arm because to do so would then allow him to strike his wife, whom he now hates on account of her infidelity? In this case, the man's paralysis is functional (psychogenic: caused by his 'mind') not organic (i.e., not a consequence of a stroke or brain tumor, etc.). However, there is a further problem with this use of the word 'functional': in itself it does not distinguish the man with hysterical conversion (an 'unconscious cause' for his symptom, F44.4) from the man who is malingering (consciously producing his symptom for personal gain, F68.1). Hence, the term functional may slide between several different meanings: from subtle brain disturbance (in schizophrenia, F20), to hypothesized psychogenic (unconscious) mechanism (in hysteria, F44.4), to implied deceit (in malingering, F68.1). We need to be especially careful when applying this and certain other psychopathological terms in clinical practice (see Table 1).

Visual versus auditory hallucinations: Similarly, a useful distinction has concerned the modality of hallucinations (perceptions experienced in the absence of a veridical external stimulus). While visual hallucinations are more often encountered in organic brain states, auditory hallucinations (particularly of voices) are more characteristic of the so-called functional disorders such as schizophrenia (F20) and severe (psychotic) depression (F32.3). A good example of an organic state would be that of delirium tremens (F10.4), seen among alcohol-dependent patients withdrawing from alcohol. The street drinker, hallucinating in the doorway, may be pointing and gesticulating at persecutors that only he can see. More frightening still may be the knives and spears he sees penetrating his body. Visual hallucinations should prompt consideration of a treatable, physical disease. The characteristic auditory hallucinations of schizophrenia and depression will be addressed below. However, we should note that though this distinction (between visual and auditory hallucinations) is very helpful clinically, it is still the case that organically ill patients may 'hear things,' and that patients with schizophrenia sometimes 'see things.' It is the relative preponderance of these phenomena that is clinically informative.

Delirium versus dementia: When faced with a patient who is cognitively impaired, for example, disorientated in time and place and unable to account for themselves, it is particularly important to quickly consider two things: the physical health of the patient (they need to be examined by a doctor) and any information that can be gleaned from their family, friends, contacts, or witnesses. When the patient cannot speak for themselves we are crucially reliant upon what witnesses can tell us about them. A global disturbance of cognitive function (meaning: concurrently impaired attention, concentration, memory, perception, language, and spatial awareness) may arise gradually over time or with a rapid, precipitate onset. The speed of onset and the stability of the findings are important for diagnosis, as is the level of consciousness: the alertness of the patient. In delirium, a physical illness (F05), such as a chest infection or withdrawal from alcohol (F10.4) precipitates the patient into an acute state of confusion, and his level of awareness may wax and wane over hours and minutes. He may come and go, get worse overnight, experience moments of apparent lucidity, and then deteriorate again. His sleep-wake cycle is much disturbed. He may be particularly prone to visual illusions (misinterpretations of veridical external stimuli): the coat on the back of the door becomes a demon; delirium is an intensely frightening experience. Hence, it may be defined as an acute or subacute onset of global cognitive impairment, in which the level of consciousness fluctuates, and where the cause is usually potentially treatable. Physical investigation and resuscitation is essential, as is a calming environment, a simple, well-lit room (to reduce the likelihood of visual illusions) and, ideally, the constant attention of the same (i.e., familiar) nursing staff. In contrast, while the dementias also impact global cognitive function, they tend to do so more slowly (over months and years) and, in general, the patient remains in clear consciousness while awake. In addition, while the person with delirium may lack insight into their current difficulties (and hence be at risk of acting on mistaken beliefs; e.g., the man who jumps from

Table 1 Some psychopathological terms that have multiple or disputed meanings

Term	Meaning
Ego-syntonic	Consistent with the subject's personality, their long-term values and behaviors.
Ego-dystonic	Inconsistent with the subject's personality, their long-term values and behaviors.
Formal thought disorder	Indicative of schizophrenia but used in two ways. The first (more accurate application) is to refer to a disordered flow of thinking as expressed in disordered speech: speech that is incoherent and at its worst comprises 'schizophasia' or 'word salad.' The second use actually refers to the content of specific delusions seen in schizophrenia, those which relate to thought, such as thought insertion, thought withdrawal, thought broadcast.
Functional	Has been variously applied to mean: serious disorders that lack a structural brain correlate but which are hypothesized to implicate biochemical or psychological etiologies (e.g., schizophrenia), or a disorder which is purely a product of unconscious mechanisms (as in hysterical conversion) or is malingered (i.e., consciously performed).
Insight	The degree to which the subjects understand their situation: do they realize that they are ill; do they understand the nature of that illness; do they believe it requires treatment; what kinds of treatment would they accept as necessary?
Multiple personality disorder	A contentious diagnosis, suggesting that the subject experiences and exhibits more than one personality, usually emerging in the course of interpretative psychotherapy. One 'personality' (the 'host') may have knowledge of all the others (the 'alters'). Has been deployed as a forensic defense.
Neurosis	A mental disorder where the symptom is an exaggeration of phenomena that might be experienced normally (e.g., anxiety) and where (traditionally) the subjects have not lost contact with reality, that is, they retain insight into their disorder.
Organic	Usually implies a mental disorder with a demonstrable physical cause, which might be structural (such as a brain tumor), pathophysiological in the brain (as in epilepsy), the body (as in hypothyroidism), or exogenous (as in drug-induced psychosis).
Paranoid	In the lay sense, it means feeling picked upon or persecuted, and this meaning is applied in the diagnosis of paranoid personality disorder. However, when applied to schizophrenia, paranoid means deluded, literally out of one's mind. Hence, one may suffer from paranoid schizophrenia, and be deluded, but one's delusions need not be persecutory (one might believe that one is Joseph Stalin or Charlie Parker).
Pseudohallucination	A confusing term, used in two ways. The first is a way of describing a percept occurring in inner space (i.e., not in the outside world), against the subject's will, but lacking the reality of a 'true hallucination' (this form is characteristically seen in people with personality disorders, an inner voice, which the patient may recognize as their own, urging self-harm: ''Jump in the river, jump in the river''). A second, less satisfactory application is in the description of a full hallucination, into which the sufferer has gained insight, in other words they are experiencing a hallucination, but they realize it is a symptom of illness and not 'real.'
Psychopath	To the public, a word often confused with psychotic (out of touch with reality), but really a term describing abnormal personalities. In the legal sense (in the United Kingdom) a psychopath has a severe personality disorder, one justifying involuntary incarceration in hospital; in diagnostic usage, a psychopath describes a particular personality disorder, an extreme variant of the antisocial: one associated with the instrumental manipulation of others, lying, and cheating, premeditated violence, lack of remorse and empathy for victims. According to this usage, a 'successful psychopath' might never come to medical/legal attention.
Psychosis	A mental disorder where the symptoms comprise phenomenology that would not be experienced in normal mental life and where the subject is regarded as having lost contact with reality, that is, they currently lack insight.
Thought broadcast	Psychiatrists use this term in three ways: to describe the experience of others being aware of one's thoughts, at a distance; the experience that one's thoughts are literally broadcast out aloud and therefore audible to others close by; or the experience that others have access to one's thoughts via telepathy. All uses are current!
Schizophrenia	Commonly confused with multiple personality disorder (above). Schizophrenia actually refers to a 'splitting of the mind,' not between identities (persons), but between functions, so that thoughts and perceptions become disturbed and incoherent.

the roof because he thinks he can fly), the person with dementia may migrate from knowledge of their failing powers (early in the process) toward a lack of recognition of themselves or others (toward the end); they are at risk of suicide while they know what is happening to them, at risk of exploitation, neglect, or misadventure when they do not.

Form versus content: Descriptive psychopathology has borrowed this distinction from philosophy and it has considerable bearing upon clinical diagnosis. An experience occurs within a phenomenological domain (e.g., as a thought or percept) and this constitutes its form. The theme or subject matter of that experience constitutes its content. Hence, a man with schizophrenia (F20) may hear voices saying "he is evil, he is dirty"; the form is a third-person auditory verbal hallucination, the content derogatory and persecutory. Contrast this case with that of another man, who suffers from obsessive–compulsive disorder (F42): he experiences recurrent, intrusive thoughts, which he knows to be his own, but which prey upon his mind: "I am evil, I am dirty." Here, the form is an obsessive thought and the content is depressive. While an interpretative psychopathologist might pinpoint the similarity between these cases (i.e., their content is derogatory to the self, the patient is identified as evil), the descriptive psychopathologist will make a diagnostic distinction (the form of the auditory hallucinations is in keeping with schizophrenia, while that of the obsessive thought is consistent with obsessive–compulsive disorder; their treatments are different). Nevertheless, while matters of form are crucial to diagnosis and identifying the appropriate therapy, the question of content remains central to understanding the person, the one who experiences that symptom. In the above examples it would not be surprising if questions of evil or immorality were important to these patients when well, if they had featured in their moral development and earlier life. However, their specific disease processes influence the manifestation of their concerns in the present: as heard voices or intrusive thoughts, respectively.

Reactive versus instrumental: The purview of psychopathology extends beyond the domains of mood and perception, and memory and concentration, to include accounts of human moral behavior, particularly those expressions of the personality that may cause suffering to patients or their society (F60–F69). When the cause of that suffering is violence then it may be helpful to consider whether the violent act arose spontaneously, in reaction to the trigger of an acute situation or whether, instead, it comprised the product of premeditation on the part of the perpetrator. So, for instance, a man who is intoxicated with cocaine (F14.0) might feel acutely irritable and hyperaroused, might be overactive and lacking in foresight regarding the consequences of his impending actions; he might then punch a man who blocks his way while leaving a nightclub. Alternatively, a severely psychotic man, suffering from schizophrenia (F20), may come to believe that the devil is stalking him, so that he carries a knife for his protection, and when he is stopped by a stranger who asks him for directions he stabs the man, believing him to be the devil in disguise. Finally, a sadistic psychopathic pedophile (F65.4) may spend many months grooming a victim over the Internet, through an alias adopted while 'online' in so-called 'chat rooms'; he may have repeatedly masturbated, thinking of what he will do to the victim when he meets her, then he acts out his plan with a precision bordering on the obsessional. This behavior is clearly 'instrumental.'

Levels of Understanding

A feature of the clinical psychopathological discourse is the tendency to arrange diagnostic entities hierarchically. By this we mean that because several neuropsychiatric syndromes may share symptoms and signs, items of psychopathology are rarely pathognomonic (solely diagnostic) of one single disorder and clinicians must arrange differential diagnoses according to an order of preference:

1. Partly on the basis of which single diagnosis might most comprehensively (and parsimoniously) account for all of the phenomena described (i.e., applying Ockham's razor);
2. Partly on the basis of relative frequencies and likelihoods of one condition over another (Alzheimer's disease, F00, is common, Huntington's disease, F02.2, uncommon); but also
3. On the basis of a hierarchy of causation (Table 2).

Hence, while a symptom such as anxiety might be entirely normal prior to exams or stage performances, it might constitute a disorder if it is extreme (persistent, disabling, interfering with vocational function); but although it may comprise both a symptom and a diagnosis (as in 'generalized

Table 2 The phenomenological state of anxiety may have many precursors or causes

Level of explanation	Exemplars
Normal	Exam stress
	Public speaking or performance
Neurosis	Generalized anxiety disorder
	Agoraphobia
	Social phobia
	Posttraumatic stress disorder
	Obsessive–compulsive disorder
Psychosis	Agitated, major depression
	Paranoid schizophrenia
Organic, exogenous	Caffeine
	Alcohol (withdrawal from alcohol)
	Amphetamines
	Paroxetine ('Seroxat')
Organic, endogenous	Brain-related: Aura preceding the ictal phase of a temporal lobe seizure
	Systemic: Thyrotoxicosis
	Hypoglycemia
Personality	Anxious, avoidant personality
	Dependent personality disorder

anxiety disorder,' F41.1), it can also be a comorbid feature of other diagnoses, to which it is contributory if not central. For instance, anxiety may accompany many of the neuroses, for example, agoraphobia, F40.0, and social phobia, F40.1, may arise as part of a depressive syndrome, F41.2, can accompany a psychotic illness such as schizophrenia, F20, but might also be a symptom of a bodily disorder such as thyrotoxicosis (hyperthyroidism) or pheochromocytoma (a tumor of the adrenal gland), F06.4 (Table 2). In all these cases the patient may experience and exhibit anxiety and, while there may be subtle differences in periodicity and duration (short bouts of panic associated with going out in agoraphobia; persistent dread in the severely depressed), the symptom is nevertheless common to all these syndromes. In eliciting a history and examining the patient, the psychiatrist is attempting to answer the following questions: Is this anxiety the totality of the condition (and if so what is its meaning to the patient, what makes it happen?) or is it part of something else?

As will be seen in Table 2, as we move from the level of 'normality' through levels of increasingly organic pathology, we also move from what we might call the very psychological to the very physical. We move from the social environment (the stage or examination room implicated in the healthy subject's anxiety), through the exaggerated impact of social interactions (in social phobia and agoraphobia) toward physical environment (ingested as caffeine or amphetamines, F15), until we encounter effects endogenous to the brain and body of the sufferer (e.g., his endocrine system), and ultimately the genes that impact his personality. Such a series of transitions also influences the therapies that are deemed appropriate under each condition. We might engage with the psychological levels through psychotherapeutic means, while the purely physical requires a different treatment (e.g., excision of a thyroid tumor). Difficulties arise when an inappropriate level of treatment is applied at the level of the patient's problem.

Disorders of Emotion

Emotions are subjective feeling states; yet they also carry physical correlates that may be apparent to observers. Hence, the frightened man experiences a subjective state of panic, an urge to escape, while his body prepares for 'fight or flight,' with arousal of his sympathetic nervous system (increased heart rate, blood pressure, blood flow to the muscles, increased sweating, respiration, and dilatation of the airways); his response may be obvious to those around him. Some emotions are so 'hard-wired' that they are similarly expressed across all human cultures. These, so-called basic, emotions (fear, anger, distress, disgust, surprise, and joy) are associated with distinct facial expressions that may emerge even in the absence of an observer (i.e., the depressed woman may still look sad even when she is alone). Other, so-called cognitive emotions necessarily implicate others: love, guilt, shame, embarrassment, pride, envy, and jealousy; these invoke relationships, whether proximate or past.

Mood and Affect

In psychiatry, it is common to describe emotions using the words mood and affect. Mood describes a subjective state, what the person is feeling. Affect is how they appear to others. Hence, mood and affect may be congruent (the woman with depression is crying) or incongruent (the man with

schizophrenia is laughing while he describes his stepfather raping him). Mood and affect may be abnormally elevated or lowered, may be exaggerated, diminished, or absent. They may alternate between the extreme highs and lows of bipolar disorder (manic depression; F31) or the lesser peaks and troughs experienced in cyclothymia (F34.0; Table 3).

Mania is characterized by extreme changes in mood, affect, behavior, and cognition (F30 and F31.1–2). The mood is elevated, as is the affect (indeed, a subject's humor may be quite contagious). Speech and movement are accelerated, there is no need to eat or sleep; sexual activity increases. Thoughts accelerate and may acquire new significance, so much so that the person loses contact with reality: he believes that he is God, that he controls the world, has acquired new insights, talents and energy and should not be constrained by others (i.e., he becomes grandiose). He may describe frankly psychotic perceptions: voices telling him he is a genius, lights and visions abound. Unfortunately, happiness may quickly give way to irritability and anger, and manic patients may harm themselves or

Table 3 Disorders of mood and affect

Exemplar	Selected features
Mania	Episode of markedly elevated mood, affect, altered behavior, and thinking; speech and movement accelerated; thoughts grandiose, heightened energy and disinhibition. Mania is a psychotic state (delusions of grandeur and hallucinations), hypomania is the nonpsychotic prelude.
Depression	Episode of marked lowering of mood, affect, reduced self-care, and impoverished thinking; may vary from neurotic to psychotic (delusions of guilt, poverty, nihilism). Risk of suicide (especially early on in treatment, as energy returns).
Bipolar affective disorder	Intercurrent episodes of mania and depression, the latter predominating over the longer term. Risk of suicide when depressed, misadventure when 'high.'
Cyclothymia	A longstanding pattern of instability of mood with mild ups and downs, mood and affect are recognized to be changeable over days and weeks. Might be observed to be 'moody.'
Dysthymia	Persistent lowering of the mood, with some features of depression, but insufficient to satisfy the criteria of full depressive syndrome. Affect may often appear 'miserable' over a long term.
Generalized anxiety disorder	Persistent anxiety markedly interrupts normal relationships and vocation. Prominent symptoms and signs of autonomic arousal.
Phobic anxiety	Anxiety precipitated by specific objects (e.g., 'simple phobias' of spiders, snakes, heights, plane travel, etc.) or social situations (e.g., crowded places and being away from home, in agoraphobia; intimate situations where behavior may be observed by others, such as speaking or eating in public, in social phobia).
Panic disorder	Sudden onset of fear and autonomic arousal, coming 'out of the blue'; fear that one may collapse and die. May complicate agoraphobia.
Acute stress reaction	Similar to panic, but in a setting of extreme environmental threat (e.g., postbomb explosion).
Adjustment reaction	A response to a change in life circumstance, such as the end of a relationship, which may be characterized by anxiety or depression.
Posttraumatic stress disorder	Anxiety may be permanent but heightened in response to environmental cues (resembling the initial trauma); recurrent intrusive imagery of the event; bad dreams, phobic avoidance of similar environments (e.g., avoiding tunnels and stair wells after a train crash).
Obsessive–compulsive disorder	Anxiety accompanies ego-dystonic intrusive thoughts and may increase until the tension can be released by performance of a compulsive act (e.g., hand washing, counting rituals, etc.).
Hypochondriasis	Anxiety is associated with the fear of illness; worries or over-valued ideas predominate.
Somatoform disorder	Similar theme to hypochondriasis, but the patient presents on multiple occasions concerning physical symptoms, which are found to be without pathological cause. May constitute an oversensitive awareness of normal bodily sensations.
Blunting of affect	Lack of feelings, persistent absence of responsivity (may be seen in chronic schizophrenia or end stage dementia).
Incongruous affect	Can be an alarming disconnection between affect and reported mood, as seen in schizophrenia, for example, laughing when they should be crying.
Emotional incontinence, hyperemotionalism	Classically associated with cerebrovascular disease in the elderly; the subject may cry easily, may be excessively moved by sentimental topics.
Catastrophic reaction	Sudden rage or panic in someone with dementia who is failing on tests of cognition (best avoided by gradual, sensitive examination of the cognitive state).

others while they are 'high.' They are also at risk of misadventure, behaving uncharacteristically (e.g., through promiscuity or drug taking). Hypomania (F31.0) describes the nonpsychotic prelude to mania; it may be quite a pleasant state, associated with new thoughts and energy. The patient might be reluctant to seek treatment.

In depression (F32 and 33) the mood and affect are lowered; the patient feels sadness, which may be accompanied by tearfulness and slowing, so-called retardation. To the observer, the depressed person appears under-active, hunched and crying, or detached and vacant. However, in the elderly patient there may be pronounced agitation, torment. Appetite and sleep fail, weight may be lost, thinking becomes an effort. Thoughts turn to failure, hopelessness, feelings of worthlessness, and guilt; a minor misdemeanor from the past assumes new significance. If there are psychotic features they are mood congruent: voices are pessimistic, visions are of Hell or disaster, the body feels diseased, the bowels have 'turned to stone' (Cotard's delusion or nihilism; F32.3; Table 3). The depressed patient may kill himself to end his suffering, because he thinks he deserves punishment or because he is a burden to others. The woman with postnatal depression (F53) is at particular risk of killing herself and her child ('to save him from this cruel world').

Anxiety states (F40–48) are states of fear, marked by an awareness of autonomic arousal: racing heart beats, sweating palms, increased respiration, paraesthesia (numbness and tingling) of the fingers and toes. Such a state may be more or less constant (as in generalized anxiety disorder, F41.1) or episodic (precipitated by specific objects in the phobias, e.g., spiders, F40.2, or arising 'out of the blue' in panic disorder, F41.0; Table 3). Sometimes anxiety is accompanied by disgust, as in obsessive–compulsive disorder (F42), when the patient feels contaminated, 'unclean.' Sometimes, anxiety and depression are combined (F41.2), the patient wringing her hands, miserable, and agitated. The most difficult mixed state to diagnose can be that of mixed mania, F31.6, where, despite all the behavioral features of increased activity, rapid speech, and fleeting thoughts, the mood described (subjectively) is one of sadness and dread.

In the *histrionic* individual (F60.4) the expression of emotion may be superficially extreme, but it seems to carry no deeper feeling; it appears exaggerated yet 'shallow,' the performance is similar to that of a bad actor: one notes the caricature of an emotion, but detects no authentic feeling behind it. The patient can flit from one emotion to another without any apparent awareness of their incongruity (e.g., the man intoning slowly, with eyes closed, as he describes the funeral of a friend, with apparent pathos, smiles suddenly as he mentions that he will buy a car with the money bequeathed him). In the *borderline* personality (F60.31) the mood can be labile, changing acutely in response to arguments or disappointments; essentially, self-esteem is unstable, the patient decompensates easily; they may self-harm repeatedly. In these cases the feelings appear 'felt,' but they may be transient. In a nonpejorative sense, they appear 'adolescent': sensitive, quick to appear, flaring up intensely, and then fading. The problem is that the period of intensity may give rise to severe self-harm (e.g., cutting or overdosing).

Deficit states: While one may conceptualize mania, depression, and anxiety as exaggerations of normal human emotions, however much they may have become contextually inappropriate, in some conditions, such as schizophrenia (F20), it is clear that the feeling tone has been lost. The affect can appear 'blunted' as if the patient is incapable of feeling happiness or sadness. When asked about his mood he may simply say that he feels 'blank,' 'vacant,' 'nothing.' Such a deficit may be associated with a lack of will and initiative; the patient says little and stays in bed much of the day. Similarly, some patients who have undergone head injuries or the effects of frontal brain tumors or strokes may lack emotion, appearing grossly apathetic or else facetious, embarking upon silly pranks.

In organic states, *hyperemotionalism* may present as a lowered threshold for tearfulness; a man may notice that his moods are stronger, more vivid; he cannot prevent himself from crying when he encounters family or friends. He cries easily at the news on the television. In some forms of epilepsy intense joy, rage, laughter, or crying may each arise, as a prelude (aura) to a seizure (ictus), though they are likely to be short-lived and incongruent under the circumstances. Nevertheless,

they may be frightening to witness. Ictal violence may be completely disinhibited and very severe. Incongruous affect also appears rather disconcerting and is seen in schizophrenia (where it may be associated with speech disturbance): the patient seems to smile while discussing pain or distress.

Disorders of Thought and Speech

Classically, textbooks of phenomenology describe disorders of thought as a single category; however, it actually constitutes an amalgam of disordered thinking, thoughts, and speech acts. There are two types of information that we are attempting to gather in a descriptive psychopathological interview: what the person tells us about their thinking, that is, what it is that they experience (subjectively); and how their thoughts are organized and expressed to us (objectively): hence, we are (inevitably) drawing inferences about their thinking and their thoughts from listening to their speech (and sometimes reading what they have written). For this reason it is helpful to divide thought disorders into those of:

1. thinking (the process),
2. thought form, and
3. thought content (or theme).

It should be noted that varied vocabularies have been applied, some theoretically derived, some derived from observation; this account will attempt to be as purely descriptive as possible (see Table 4).

Regarding the thinking process, we are interested in the emergence and flow of ideas; the spontaneity exhibited, and the coherence of the stream of thought. If we ignore the content for a moment, we are really listening for whether the flow and cadences of this person's speech sound as they might if we were engaged in a 'normal,' healthy conversation. Clearly, the clinical interaction is not 'normal' for the patient; they may be nervous, guarded, embarrassed, or resentful about 'having to see a psychiatrist.' Nevertheless, we must listen carefully to their speech. In general, in depression and mania, even if the subject matter is distressing it usually remains understandable; and even if we cannot interrupt the manic patient we can probably follow what they are saying. When speech becomes incoherent, muddled, and disjointed, we are more likely to be dealing with schizophrenia or an organic mental state. When we use the specific term 'formal thought disorder' we are denoting the incoherence of schizophrenia (F20). This is one of those psychopathological terms that have to be used very precisely (Table 1).

The form of a thought and its content (theme) are summarized in Table 5. The most important forms of thought that a psychiatrist detects are those of overvalued ideas, obsessions, and delusions. Within each of these there is a theme: the concern of the anorexic teenager with her fear of fatness (F50.0), the obsessional's guilt that he thinks of feces (F42), and the delusional system which transforms the reality of the paranoid patient (e.g., F22). One has to listen, but also to probe ideas, one has to ask questions, even challenge what the other believes, in order to gauge whether they really believe it, whether they might occasionally question it themselves and, perhaps most importantly of all, whether they will act upon it.

Risk

Most psychiatric diagnoses are associated with an increased risk of self-harm. This is most marked in psychoses (especially depression and schizophrenia), and is especially likely among those who are without hope or in pain, who have lost much through their illness and who lack social support and family. One should always enquire about suicidal ideation, its frequency and intensity, any planning that has occurred and what opposing constraints there may be; suicidal risk may be ameliorated somewhat by religious belief, 'something worth living for' and the existence of people whom the patient would not wish to leave behind (bereaved by his death). Self-harm may be 'driven' by certain phenomena: command hallucinations, telling the patient to kill herself, or motor restlessness (akathisia) induced by certain drugs (below); sometimes self-harm is a misadventure driven by a delusion ("I am invincible, I can fly"; "if my hand does not burn, I have been saved").

Harm toward others is greatly stressed by the British media. Particular attention is devoted to 'stranger violence' perpetrated by men with

Table 4 Disorders of the process of thinking and its expression (speech)[a]

Behavior	Possible causes	Definition
Silence	Mutism	Rarely primary (a patient who never acquired the power of speech), usually secondary – 'involuntary' (due to depression, mania, schizophrenia, brain lesion, such as left frontal stroke, hysteria) or 'voluntary' (elective mutism). 'Psychiatric stupor' describes a mute patient whose eyes follow events in the room, cf. 'neurological stupor' in which consciousness is obtunded. From the patient's perspective the problem may be one of a lack of thoughts (depression, schizophrenia, medial frontal stroke) or their multitude ▫ incoherence (mania, schizophrenia).
	Shock	Sudden severe traumatic events may render patients speechless; they may appear psychiatrically stuporose (above).
	Thought block	Rarely a schizophrenia patient may later describe how their thoughts were suddenly 'stopped'; they may implicate a persecutor (someone who 'took them away').
Reduced speech	Inarticulacy	Primary language, not the language of the interview; patients with low intelligence may speak slowly, as may the suspicious, embarrassed or frightened.
	Alogia	Reduced volume (amount) of speech in schizophrenia or depression. Seen with psychomotor poverty or retardation, respectively.
	Nonfluent dysphasias	Left frontal lobe lesions may render the patient mute, or nonfluent (Broca's aphasia), in which case the patient will appear to be trying to speak, though unable to do so, and will seem to perceive their impairment. In transcortical motor aphasia a frontal lesion renders the patient unable to speak spontaneously though he may be able to repeat what others say.
Slurred speech	Dysarthria	May reflect potentially reversible causes such as intoxication, drug overdose, or transient ischemic attack; if permanent, suggests neurological disorder, for example, classically, cerebellar disease, distributed cerebrovascular lesions (strokes), multiple sclerosis, motor neuron disease, Huntington's disease.
Repetition of words	Verbigeration	Repetition of the same words over and over, may indicate confusion, schizophrenia, or noncooperation.
	Palilalia	The patient's repetition of the last word of their own phrase is said to constitute an extrapyramidal sign.
	Echolalia	When the repetition is of the interviewer's words: this may (rarely) be a symptom of schizophrenia or transcortical motor aphasia or else a sign of noncooperation.
Repetition of themes and phrases	Perseveration	Seen in organic brain states (classically frontal lobe syndromes) and chronic schizophrenia. The patient returns to prior themes repeatedly, inappropriately, and out of context.
Poverty of content	Empty speech	Normal amount of speech but the subject says little. May be normal, longwinded, and circumstantial. May have difficulty concentrating; may be easily distracted by environmental cues, as in delirium, schizophrenia, attention deficit hyperactivity disorder. Rarely 'talking past the point' in hysterical states. In schizophrenia, the patient may use stock phrases or excessive, pseudotechnical, 'stilted,' arcane forms of speech, but still convey little. He may lack a goal and wander away from the point and not return. Sometimes speech is so self-referential that it conveys little new information (the patient brings all enquiries back to himself and his concerns).
	Puns and clang associations	Language structured according to its surface qualities (meanings and sounds, respectively), in schizophrenia, hypomania, and sometimes hysteria. ('The present sure is tense,' with apologies to Captain Beefheart.)
Increased rate of speech	Pressure of speech	Speech fast but coherent, may be difficult to interrupt. Tends to become loud and emphatic. Characteristic of hypomania, then mania.
	Flight of ideas	In mania, the patient rushes from topic to topic although he can still be followed.
	Tangentiality	Again, running from thought to thought, but there is an elliptical connection between successive themes. Mention of birds may turn to birds of prey, then to forms of prayer, then to religion.
	Knight's move thinking	As mania speeds up the links between thoughts may become less direct. The interviewer may detect where links were lost. Difficult to keep pace with, best recorded and then described verbatim. In the above example, the spoken (manifest thoughts) might skip from birds to prayers to bells. . .

Continued

Table 4 Continued

Behavior	Possible causes	Definition
Incoherent speech	Derailment	The patient (with schizophrenia) slips between unrelated topics with successive sentences.
	Illogicality	The patient (with schizophrenia) slips between unrelated topics within the space of a sentence. ''In Panama my fisherman's waitress swore yellow glasses in a bird's nest.''
	Paraphasias	Word approximations, may be seen in dementia and schizophrenia, for example, gloves referred to as 'handshoes.' Sometimes the process is one of elimination: ''I went to. . .you know. . .the building, not the school. . .no, the building, not the shop. . ..after the church. . .the hospital. That's right!''
	Neologisms	The patient makes up his own words. Characteristic of schizophrenia (but important to exclude malapropisms and poor educational attainment).
	Word salad	Also called schizophasia, when speech becomes totally incoherent.
	Fluent dysphasis	Left superior temporal lobe lesions may render a patient incoherent but fluent (Wernicke's aphasia), in which case they will not appear to perceive their own impairment.

[a]Always attempt to establish the patient's first language and handedness.

schizophrenia or personality disorders. The risk to others is greatest from those who are paranoid, in the lay sense of the word (Table 1); who abuse illicit substances (especially, crack cocaine and amphetamines); who are nonadherent to conventional medications; and whose personalities might have been prone to violence anyway, even in the absence of a mental illness (men who were 'antisocial' and/or 'paranoid,' in the lay sense, before they ever became mentally ill). Absence of empathy for others or remorse toward victims are poor prognostic indicators. In terms of relationships, jealousy, delusional, or otherwise may lead to violence against a partner, and erotomanic delusions (of being loved by another, at a distance) may occasionally lead to stalking of celebrities, politicians, doctors, and other figures regarded by the patient as hierarchically significant. There is a particular risk to the families and partners of those who are stalked in this way (the deluded individual may believe that their 'relationship' with the beloved might be consummated were it not for the beloved's family).

Ego Boundaries

Thoughts and perceptions may sometimes overlap in the phenomenology of schizophrenia (F20). If a man tells us that he is thinking other people's thoughts and that he feels their point of entry into his skull, is he admitting to thought insertion (a false belief) or a physical hallucination (a false perception) or both? Clinically, it is best to describe the phenomenon as clearly as possible. In terms of theory what his problem exemplifies is the porosity of the so-called 'ego-boundary,' the notion that we may differentiate our own selves from the outside world. In schizophrenia, these territories overlap repeatedly: patients experience others' thoughts and think others access theirs; the physical borders of their body are breached and viscera are touched or moved by agents whom they cannot see. Empirically, it seems likely that such disturbances implicate abnormalities of parietal cortical regions; clinically, what is important is that the patient should be afforded space and not be unnecessarily crowded or touched by those caring for them. Physical contact may be misperceived.

Misidentification Syndromes

The elision of perception and belief is also apparent in the so-called misidentification syndromes. In these, awareness of others' identities is either lost or bestowed inappropriately. In the so-called 'illusion of doubles' (Capgras syndrome) the subject believes that someone known to them (often a spouse or child) has been replaced by someone else, sometimes a real enemy, sometimes a 'robot' or alien. The sufferer agrees that the object of their problem 'looks like' the person they should be but 'something has changed.' Often it is that the person looks less 'real,' less herself; there is a resemblance, but that is all. The phenomenon may be frightening for those involved and may

Table 5 Disordered forms and contents of thought and speech

Form	Description
Over valued ideas	Single themes that can come to dominate a person's life, without being psychotic; often value-laden, for example, the pursuit of thinness in anorexia nervosa, the avoidance of waste in obsessive hoarding. The themes of neuroses may also be of this sort: the fear of illness in hypochondria, over concern with normal physical sensations in somatoform disorder, the cognitive distortions of depression (''People don't like me''), and dysmorphophobia (''My nose is too big'').
Obsessions	Recurrent intrusive thoughts, which the sufferer recognizes as his own, the content of which may be absurd □ distressing, and which he tries to resist/suppress. Common themes include contagion and symmetry; magical thinking may be invoked (''if I walk on the cracks in the pavement my father will get cancer''). Often accompanied by rituals and compulsions, for example, hand-washing. In the addictions, cravings may behave like obsessive thoughts, though the extent to which they are resisted varies with the stage of addiction (''I need a drink'').
Delusions	Abnormal beliefs, acquired through abnormal inferences and out of keeping with the subject's culture. Said to be incorrigible to counterargument. May be primary (rare), coming out of the blue, or secondary to other phenomena, for example, hallucinations or mood states. Characteristic themes: Persecutory delusions – others are against the patient Delusions of reference – others are communicating with him (e.g., the song on the radio is 'meant' for him) Grandiose delusions – mania, schizophrenia, GPI Delusions of guilt – depression Nihilistic delusions – Cotard's delusion Hypochondriacal delusions – depression Religious delusions – may also be grandiose Delusional jealousy – Othello's syndrome, particularly dangerous, commonly seen in alcoholic males who may kill their partners Erotic delusions – De Clerambault's syndrome, also high risk Delusions of control – the patient perceives their thoughts and actions to be under the control of external agencies. Some risk of violence, especially if the perceived controller is identified in proximity to the patient
Beliefs about thought possession	Classic first rank symptoms (Table 6) are thought insertion, withdrawal, and broadcast. Alien thoughts are inserted, one's own thoughts extracted, or known at a distance. Characteristic of schizophrenia. In both schizophrenia and depression there may be a 'milder' experience of 'influence,' so-called 'passivity': the thoughts and feelings are still one's own but they are influenced or 'made' by external agencies.
Shared delusions	Occasionally, people living in close relationship with each other (and relative isolation from others) may come to share the same delusion. Usually, there is a dominant, primary subject who is mentally ill, and in whom the delusion originated, and a passive, secondary subject who has come to share the delusion. Separation of the pair may lead to spontaneous recovery in the second subject. However, they are likely to have developed an enmeshed relationship so that separation may itself be very traumatic. When two are involved, this has been termed folie a deux. Some small sects may exhibit similar dynamics (often it is the leader who is disturbed while the sect 'members' are relatively normal).

constitute a risk for the misperceived (the 'impostor') as the patient may try to kill them.

In Fregoli's syndrome, the problem has been 'reversed': instead of an intimate having been replaced by a stranger, now strangers are replaced by a known (probable) persecutor. The syndrome may involve an elderly woman on a ward, who becomes ill postoperatively and notices that staff have been replaced by a neighbor (in disguise) whom she does not like. Again, there may be assaults upon the misperceived.

The neurological term for the inability to recognize a familiar stimulus, in the presence of preserved sensation and movement, is agnosia, classically a consequence of inferior temporal lobe lesions. Hence, it may be that misidentification syndromes occupy the same perceptual hinterland as 'prosopagnosia' seen in Alzheimer's disease, when a patient no longer recognizes those whom they should, for example, the daughter who cares for them.

Disorders of Perception

In order to understand disorders of perception it may be helpful to consider the normal relationship between our thoughts and experiences.

Top-Down and Bottom-Up Processing

Imagine standing in a very crowded room, among many other people talking. While one is engaged in conversation one may be able to hear one's interlocutor's voice quite clearly, even amidst the din of a hundred other voices. Hence, we may clearly choose to focus our attention upon one stream of incoming information. This has been termed 'top-down' cognition. It implies that higher processes (and centers) in our brains can focus on salient information, despite our presence within a complex environment containing many competing streams of data. We have an element of cognitive control over what we choose to attend to. However, even though we may be focusing upon what our companion is saying there are still moments when we may be suddenly distracted by the mention of our own name by someone else. Thus, even though we were using 'top-down' processes to focus upon one stream of data, somewhere in our brains another process was analyzing those extraneous data emerging around us, so that a single salient stimulus (our name) might gain access to consciousness, 'bottom-up.' This situation exemplifies the relationship between the contents of our awareness, in consciousness, and the streams of stimuli arising within our environment: we need to focus on the salient and while we can do so volitionally, consciously ('top-down') to an extent, we must also rely upon automatic ('bottom-up') mechanisms to detect significant changes in other channels of information. When we are tired, or our hearing is failing, then these processes can become difficult to maintain and their data to disambiguate.

In the above example, the voice of our interlocutor constitutes a true (veridical) stimulus in our environment. While we focus upon what he is saying we can accurately hear his words, even if the room is noisy. However, should we fail to detect his signal properly, should we mishear his speech, then we may experience an illusion. He says ''Chris is retired now''; we hear ''Chris is getting tired now'' or even ''Chris has been fired now.'' Meanwhile, there is the possibility that unattended information in other streams of data might also generate illusions; especially if the illusory material resembles something salient for us ('is that Minnie Driver?').

Top-down processing is all the more apparent in mental imagery (e.g., in our 'mind's eye'), when we 'call up' an image from our past or an imagined future (so-called prospective memory). Because imagery is under top-down control it contrasts with obsessions (thoughts that enter the mind against the subject's will and keep coming back) and also hallucinations (where the percept, e.g., a voice, also acquires consciousness, unbidden). A central problem common to obsessions and hallucinations is that the subject cannot control them using 'top-down' techniques, he cannot 'make them go away.' Now, while illusions and inner imagery are of marginal interest in psychiatry, hallucinations carry much more diagnostic significance.

Hallucinations

The problem for patients with a number of psychiatric disorders is that perceptions emerge into their consciousnesses, which do not originate from real external stimuli and cannot be controlled. It is as if aberrant bottom-up processing cannot be modulated or 'turned down' by top-down processes. Hence, a man with schizophrenia (F20) hears someone speaking about him and he cannot make 'him' stop. Hallucinations have the quality of reality. They may be simple (e.g., tones, lights) or complex (faces, voices). They can arise out of the blue or be carried by environmental signals ('functional hallucinations,' e.g., voices heard in the roar of passing traffic) or else follow on sensations in another modality ('reflex hallucinations,' e.g., a voice that is heard when the skin is touched). Sometimes hallucinations 'could not possibly' be based on real veridical experiences, for example, with extracampine visual hallucinations, the patient 'sees' out of the back of his head; during out of body experiences, he can look down upon his physical body from above. In autoscopic hallucinations he may see his own 'double' walking toward him (such doubles can be mirror images or 'correctly' transposed, that is, with the left ear to right side of the figure).

Auditory hallucinations are particularly characteristic of functional psychoses (above). Some of those described in schizophrenia have acquired special status (as so-called first rank symptoms; Table 6)

Table 6 The first rank symptoms of Schizophrenia

Modality	Symptom
Auditory hallucinations	Third-person comments (including running commentary) upon the patient (''He's looking at the TV; he's changing the channel. . .'').
	Voices conversing between each other.
	Patient's own thoughts being spoken aloud, simultaneously in Gedankenlautwerden, after a delay in echo de la pensee (thought echo)
Bodily hallucinations	Bodily sensations that the patient attributes to external agencies, characteristically affecting internal organs (e.g., ''Professor Hirsch is probing my spine, I can feel his fingers pressing through me. . .'')
Passivity phenomena	The patient experiences his thoughts, feelings, and movements as being under the influence of external entities. They are still his processes, but they are impeded or interfered with by others.
Alien phenomena	Processes are no longer his: thought insertion (he is thinking others' thoughts), thought withdrawal (his thoughts are taken away), thought broadcast (his thoughts are known to others at a distance).
Primary delusions	Rare and diagnostically specific occurrence: includes the so-called delusional perception in which a normal percept, not thematically related to his situation, suddenly precipitates a delusional interpretation by the patient (''I was walking down the street, the traffic lights changed to green, and I knew then that I was the Messiah''); also the autochthonous idea, a delusion which suddenly appears without precedence (''As I sat, thinking, in the library I suddenly realized that I had been chosen to save the world from the masons'').

not because they are confined to this disorder, not because they are the only auditory hallucinations described in it, nor again because every schizophrenia patient experiences them, but because they are 'so unusual' that they are highly likely to be reliably ascertained and agreed upon by different psychiatrists. These auditory hallucinations, described by Kurt Schneider, are listed in Table 6.

Hence, the first rank symptoms have clinical utility because they are so characteristically abnormal. However, people with schizophrenia may also experience second person voices (speaking to, rather than about, them) and, unfortunately, third-person voices can occasionally arise in people with depression, mania, or epilepsy. Again, we are working with the phenomenological nature of psychiatric diagnoses and attempting to compare the constellation of symptoms occurring in a single case with an empirical, syndromal profile, to find the best 'match.'

Generally speaking, in depression and mania, auditory hallucinations are more likely to be second-person in form and mood-congruent in content. While the voices heard in schizophrenia tend to be derogatory and intrusive, seemingly aware of what the patient is thinking, those in depression say depressing things (''The world is ending, you have cancer''), while those heard by the manic may exalt him (''You are a genius, you are cleverer than Ray'').

There are many less common causes of auditory hallucinations and these tend to lack characteristic features. However, there are some notable exceptions: the long-term heavy drinker may develop alcoholic hallucinosis (F10.52), hearing derogatory voices when drinking, in the absence of delusions (i.e., it appears to be a disorder solely of perception); musical hallucinations are another example: they may arise in otherwise healthy elderly people who experience hearing impairments (these patients usually realize that the music is not 'real').

Visual hallucinations are more characteristic of organic mental states; particularly those induced by drugs, for example, hallucinogens such as lysergic acid diethylamide (LSD) (F16), or epilepsy or migraine; but they can also arise in the functional psychoses, delirium and dementia. As with musical hallucinations, occurring in the hard of hearing, elderly people who are visually impaired may sometimes develop visual hallucinations, which may be simple (flashing lights, colors) or complex (small figures, in the so-called Charles–Bonnet syndrome). If cognitively intact, the older patient usually recognizes that such phenomena are not 'real' (hence, in one sense, these constitute 'pseudohallucinations'; Table 1). The greatest diagnostic significance of visual hallucinations is that they prompt one to consider 'organic' disorders (above).

Olfactory and gustatory hallucinations (of smell and taste, respectively) in general suggest temporal or frontal lobe pathology, particularly in the form of

focal partial epilepsies (classically, temporal lobe epilepsy), where unusual smells, such as burning flesh or rubber, may be accompanied by an epigastric aura ('butterflies in the stomach') and facial flushing. In functional psychoses, the schizophrenia patient may also smell burning flesh, with distressing associations (e.g., Hell, the Devil), while the depressed person smells her body rotting or giving off noxious fumes.

Physical hallucinations, that is, disturbances of bodily perception, constitute a wide range of sometimes intermingled phenomena, which go by a variety of names in different texts: tactile, haptic, visceral, somatic, and so on. They pose an interesting philosophical problem in that many physical hallucinations do not concern 'veridical, external stimuli'; in other words, rather like an 'itch,' they cannot be verified by a third-person observer. Nevertheless, what tends to attract attention is the proffered explanation for such phenomena (i.e., what the patient tells us about them). In psychiatric terms, the most telling is a hallucination implicating the viscera, especially the sexual organs; this usually denotes schizophrenia (F20). The patient may say that her persecutors are raping her, that they can change the shape of her spine (and this constitutes a first rank symptom; Table 6). Organic disorders of the parietal cortex, especially on the right side, may sometimes give rise to the experience that the left half of the body is occupied, controlled by someone else ('somatoparaphrenia'). In substance misuse, cocaine is particularly associated with formication, the feeling that there are insects moving under the skin (F14.04). Delusional infestation (with very similar phenomenology) may occasionally arise as an isolated symptom in the elderly (Ekbom's syndrome; F22–23).

Disorders of Cognition

It is a moot point whether we are ever really assessing specific phenomenological domains in isolation from cognition during the psychopathological interview, for it might be argued that cognitive processes underpin and support all of the phenomena described here anyway (e.g., those cognitive processes integral to perception and speech). However, when we talk about cognition we are usually referring to those explicit aspects of intellectual function that the patient might regard as contributory to their 'thinking,' for example, attention, concentration, and memory. It is also clear that as we progress from disorders of emotion and perception toward disordered cognition, we are increasingly dependent upon objective data (e.g., measured intellectual performance) and less upon the subject's point of view. There are several reasons for this: first, the layperson may not necessarily have stopped to assess their own cognitive state (as opposed to their having inevitably experienced their own emotional state); second, the very deficit of cognition that we are investigating might itself have prevented the patient from perceiving that deficit. We rarely 'know what we do not know'; hence, amnesic patients cannot access what it is that they have lost in order to tell us what it is that is missing.

If 'attention' is the ability to focus on information (e.g., the examiner's instructions), and 'concentration' the ability to remain focused (over seconds, minutes, or longer), then it is clear that a great many mental disorders might potentially interfere with such abilities. Some reasons may be 'normal': the patient is bored, tired, uninterested, or more interested by something else happening elsewhere. Other causes might be transitory and, hence, potentially missed or misleading: intoxication, sedation, concussion, delirium, or hallucination (Table 7). Still others might be longstanding and definitely pathological: attention-deficit disorder, dementia or learning disability (Table 8). It follows that while it may be helpful to detect abnormalities of attention and concentration, in order to gauge the patient's safety in their environment and their ability to engage with others, the differential diagnosis for a detectable impairment is very broad indeed. Much depends upon the accompanying data.

When we consider 'orientation,' we mean, 'does the patient know where they are in ''time, place, and person'' '? In most functional mental illnesses, such as the psychoses and neuroses, people retain this type of information (though the institutionalized may occasionally lose track of time; 'everyday is much the same'). People with chronic schizophrenia sometimes underestimate their own age.

Table 7 There are many possible causes of delirium (acute organic brain state)

Potential cause	Exemplars
Drug or toxin	Drug or alcohol intoxication or withdrawal.
Infection	Viral or bacteriological infection affecting the body systemically
	Viral or bacteriological meningitis or encephalitis
	TB, HIV, syphilis, etc.
Endocrine or metabolic	Diabetes mellitus, thyroid disease
	Hypo- or hyperglycemia
	Hypo- or hypercalcemia
	Electrolyte disturbance, renal or hepatic failure
	Porphyria
Nutritional	Thiamine deficiency, as in Wernicke–Korsakoff syndrome. Seen most in alcohol dependence, rarely in malabsorption syndromes, hyperemesis gravidarum.
Autoimmune	Systemic lupus erythematosus, sarcoidosis
	Behcet's syndrome
Neoplastic	CNS tumor (primary or secondary)
	Para-neoplastic syndrome
	Endocrine tumor, for example, pheochromocytoma
Neurological	Traumatic brain injury
	Vascular, for example, transient ischemic attack (TIA), subdural hematoma
	Seizure
	Migraine
	Sleep disorder, for example, narcolepsy

Table 8 There are many possible causes of dementia (chronic organic brain state)

Potential cause	Exemplars
Degenerative	Alzheimer's disease
	FTDs, including Pick's disease and Semantic dementia
	Huntington's disease
	Parkinson's disease, Lewy-body dementia
	Creutzfeldt–Jacob disease
	Normal pressure hydrocephalus
	Multiple sclerosis
Vascular	Multi-infarct dementia
	Chronic subdural hematoma
	Cranial arteritis
	Binswanger's disease
Neoplastic	CNS tumor (primary or secondary)
Traumatic	Severe single head injury
	Repeated head injury in boxers (dementia pugilistica).
Infection	Encephalitis of any sort
	Neurosyphilis
Metabolic	Sustained uremia, hepatic failure, dialysis (aluminum toxicity)
	Wilson's disease
Endocrine	Hypothyroidism
Toxic	Alcohol
	Heavy metal poisoning (lead, arsenic, thallium)
Anoxia	Postcardiac arrest, chronic respiratory failure, carbon monoxide poisoning, anemia
Nutritional	Sustained lack of vitamin B_{12}, folic acid or thiamine

However, in the 'organic' states of delirium and dementia, temporal and spatial information will often be disturbed, less so the subject's identity (though they may confuse and misidentify others; above).

The perception of *time* is a complex process. In psychopathological terms it is of note that time can seem to speed up in certain states (e.g., mania, temporal lobe epilepsy, and LSD intoxication), and slow down in others (especially depression). Patients suffering the consequence of bilateral medial temporal lobe destruction (e.g., following herpes simplex encephalitis) may exist in a perpetual present: they have no recent memory beyond a few minutes ago; they have lost that interval between the onset of their illness and the present ('anterograde amnesia'). Following head injuries, poisoning or anesthesia patients may lose awareness of moments preceding the target event ('retrograde amnesia'). A heavy drinker of alcohol may lose temporal islands in his immediate past ('blackouts'), for example, he cannot recall getting home the night before.

Abnormalities of *place* are not unusual in organic brain states: patients often do not know where they are, indeed one of the presenting complaints of delirium and dementia may be getting lost, not knowing one's way around the neighborhood. In reduplicative paramnesias the subject believes their present physical surroundings to be someplace actually known to them from former experience (probably a complex amalgam of disorientation and illusions and/or hallucinations of the former environment); this is classically associated with delirium tremens (alcohol withdrawal; F10.3). In some forms of epilepsy and migraine, the environment never seen before may nevertheless seem familiar (*déjà vu*) or the known environment

unfamiliar (jamais vu). Sounds may appear inappropriately familiar (deja ecoute) or unfamiliar (jamais ecoute).

It is usually only in advanced dementia that a person loses knowledge of their own identity. If an apparently physically healthy person, who uses language normally and interacts with everyday objects purposefully, says that they do not know who they are, then it is highly likely that the cause is psychogenic (hysterical fugue, F44.1, or malingering, F68.1).

Memory, per se, as opposed to orientation, is a further group of processes, the distinctions between which may be neuroanatomically informative. There are competing methods for classifying memory and its disorders, but a clinically useful schema is shown in Table 9.

Memory can be conceptualized as the storage of information for action. Hence, it follows that information may be brought out of storage, either by the subject accessing the data themselves (through a process termed 'retrieval') or through some external stimulus prompting 'recall.' Retrieval is the more difficult to achieve and may fail earlier in cognitive disorders. It is generally easier to recall something if prompted with a 'cue.' Retrieval can be seen as a relatively 'top-down' process, while recall is more 'bottom-up' or stimulus-driven.

During delirium (by definition) all mental processes may be disturbed (F05). This is also the case in the dementias (F00–03), yet it is clear that at their onsets the dementias may be characterized by relatively discrete psychopathologies that may make them easier to diagnose. For the clinician, there is a noticeable distinction between the dementia of Alzheimer's disease (F00) where the patient at first exhibits symptoms concerned with forgetting everyday subject matter and spatial coordinates (e.g., 'getting lost'), and the frontotemporal dementias (FTDs, e.g., Pick's disease, F02.0) where the first signs might comprise embarrassing, uncharacteristic behavior, for example, a man undoes his fly to urinate at the graveside during a funeral. The psychopathological 'picture' reflects something of the brain regions initially implicated in the pathophysiological process: the temporal and parietal regions in Alzheimer's disease, the frontal regions in FTDs. While most dementia

Table 9 Memory may be divided into multiple systems

Form	Function
Explicit or declarative memory	That which may be spoken about, explicitly. Includes: Prospective memory 'Memory of the future,' really a construction of possible future states (albeit based on known data): thought to implicate prefrontal polar regions. Involved in intentions, plans for future action, strategies. Working memory The contents of consciousness in the immediate present. Classically shown to be able to store a mean of 7 bytes of information (e.g., a telephone number to be dialed). Supported by function of dorsalateral prefrontal executive regions, manipulating data held in posterior association areas 'online.' Episodic (autobiographical) memory Remembered events that the subject has experienced in her life. Implicates medial temporal lobe structures, especially the hippocampus. Semantic memory Shared knowledge of the world, including facts, definitions. Implicates inferolateral temporal regions, also parietal cortices.
Implicit or nondeclarative memory	Forms of memory that cannot be described verbally. Include: Procedural memory Memory for how to perform certain tasks, for example, riding a bike, playing the piano. Implicates the basal ganglia especially. Priming A subject may have learned information without being aware that they were learning, in response to environmental stimuli of which they were unaware, or following which they have no conscious memory, that is, their performance in the present is enhanced by prior experience, of which they have no awareness.

processes elicit a gradual decline in cognitive function, some are temporally distinct: hence, the classic 'stepwise' descent in vascular 'multi-infarct' dementia (F01.1), where new episodes of stroke bring incremental deteriorations; and Creutzfeldt–Jacob disease (CJD, F02.1) and 'new-variant CJD,' where the illness may affect a younger adult and progress is catastrophic, declining to death in as little as a year.

It follows that if memory and cognition are partially 'localizable' to certain brain systems then there may be relative sparing of some faculties (initially), while others are lost. In cortical dementias, for example, Alzheimer's and FTDs, the patient loses episodic memory and visuospatial cognition at first in the former and social control and language function at first, in the latter. However, both may exhibit relatively preserved overlearned manual skills (procedural memory), subserved by the basal ganglia and other 'lower' systems. In contrast, a so-called subcortical dementia (such as Parkinson's, F02.3, or Huntington's diseases, F02.2) may affect the control of movement early on, also slowing cognition (through involvement of the basal ganglia). So, it is the constellation of symptoms seen that aids diagnosis; we are rarely reliant upon a single symptom.

Nevertheless, certain disorders may be restricted to one domain of memory (e.g., the variant of FTD confined to 'semantic dementia'). Clinically, the most common would be Korsakoff's syndrome (F10.6), encountered following the prolonged use of alcohol and the consequences of thiamine deficiency. This is a syndrome in which the subject loses autobiographical memory, exhibiting a dense anterograde amnesia from illness onset forward, and a retrograde amnesia, stretching back from onset 'toward' his younger life (he may retain relatively preserved memory for his childhood). Classically, Korsakoff's is an amnestic syndrome, not a dementia, because only memory is affected, but it is clear that this is not the sole symptom: sometimes the patient 'confabulates,' telling stories to fill in the gaps in his memory. Sometimes the stories can be quite fantastic, and studies have suggested that confabulation indicates frontal impairment (in addition to the usual midbrain pathology of Korsakoff's syndrome).

As dementias progress they become less discretely disorders of memory, and come to disturb all elements of mental life. Patients' personalities may change, their mood states may become labile or apathetic; they may hallucinate and be unable to speak or communicate. Some develop seizures. Behaviors become difficult for carers to cope with: especially screaming, wandering, incontinence, and violence. By this stage, there may be little way of accessing the disordered 'consciousness.'

Visuospatial and Self-Perception

While the emphasis in descriptive psychopathological interviews often focuses upon verbal information, which subjects may describe or remember in words, there is clearly a need to consider nonverbal data, for example, visuospatial information. While in the right-handed subject verbal information is preferentially processed by centers in the left hemisphere, the nonverbal, visuospatial, and acoustic data (e.g., prosody) appear preferentially processed by the right, non-dominant hemisphere. Hence, right-sided parietal lesions may leave the patient particularly impaired in a variety of ways concerning the self's relation to space.

In hemispatial neglect the subject seems unaware of the left side of space; they may eat from only the right side of their plate, neglect to dress the left side of their body, read only half the face of a clock. Initially, the problem may be subtle and revealed only when sensory data are complex (e.g., if there are salient stimuli arising in both halves of the visual field then only those on the right may be perceived). Awareness of the left side of the body may be lost or radically altered: a woman with right parietal epilepsy may report that her left arm is swelling and no longer belongs to her. Following a right parietal stroke a man may report that the left limb in the bed belongs to a relative. Stranger still, following vestibular stimulation (cold water placed in his left ear) he may temporarily regain awareness of his left limb, only to lose it again later.

Given that parietal association cortex synthesizes data from many incoming streams of information it is unsurprising that a great many complex disorders may follow its damage.

If a subject is asked to keep their eyes closed, then somatognosia is a failure to locate an object that touches the skin; finger agnosia is a failure to recognize which of their own fingers was touched; astereognosia describes an inability to identify a three-dimensional form, placed in their hand; while agraphaesthesia is an inability to recognize what has been 'written' on the skin under the same conditions. Anosognosia describes a patient's lack of insight into their own illness. What these phenomena seem to have in common is an impaired appreciation of the state of the body.

In left parietal lesions the deficits are more disparate, including apraxia (an inability to copy a purposeful voluntary movement despite the preservation of normal power and sensation – the deficit implicates the programming of motor sequences); when left angular gyrus is affected, 'Gerstmann's syndrome' may arise, in which the subject exhibits dysgraphia, dyscalculia, finger agnosia, and right-left disorientation.

While the right parietal cortex seems 'interested' in egocentric space – how the world is orientated according to our bodies (to the left and right) – the right hippocampus seems involved in memory for allocentric space (impersonal, map space). Hence, it is particularly implicated in navigation through space.

Insight

It is perhaps fitting to address insight after parietal lobe disorders, given that parietal lesions may deprive patients of awareness of their disabilities (anosognosia). Insight is a multifaceted concept that is deployed in a variety of ways by different disciplines. In terms of descriptive psychopathology, we are primarily concerned with the patient's understanding of the abnormal phenomena they have described to us. At its most basic, the presence or absence of insight is deployed in the diagnosis of neurosis or psychosis, respectively; indeed one often hears the banal declaration that insight is 'intact' or 'absent,' a patient being said to possess 'no insight.' It is most unlikely that insight can ever be said to be entirely present or absent in anyone, let alone someone who is attempting to survive a major change in their mental state.

A more nuanced approach to assessing insight was proposed by Tony David. It involves a staged approach:

1. Does the patient recognize that something is wrong?
2. What do they think is the nature of the problem?
3. Do they think they require help with this problem?
4. What form of help do they think is required?

By the time one has negotiated these steps one is in a far better position to comment upon the insightfulness of the patient-subject.

Indeed, one may find a similar approach to assessing 'capacity' in other areas of life: the capacity to make a will, to advise legal counsel, to marry, to decline medical treatment. Capacity and insight relate to specific decisions; we are all better equipped for undertaking certain judgments than others; in any given important decision in life we need to be able to understand what we are being told, believe the information provided (and our informant), know the consequences of different lines of action, weigh them in the balance, and communicate them with our solicitor, doctor, executor. We are all sifting levels of information.

Disorders of Movement

All movements (and, indeed, even the absence of movement) are of interest phenomenologically; behavior is indicative of mental life, albeit at some remove. Hence, even the mute patient is 'describable' in psychopathological terms. Also, in the normal course of events there are movements that we may take for granted, which are to be remarked upon phenomenologically: the way we use our eyes and hands while talking, the intonations of our voice, the seemingly involuntary shifts we make, on the chair, at the table. Some of these simple forms of information are exaggerated during the psychiatric interview: the depressed person may move very little, offering no expressive aids to speech, while the manic may be standing on the table or trying to climb out of the window. Indeed, the physical space between people is also informative: the frightened man on 'acid' (LSD)

cowers in the farthest corner of the room, while the woman elated on ecstasy sits too close, smiles and flirts; an antisocial male may sit, staring fixedly, deploying a direct gaze that feels intimidating. One of the classic texts on forensic psychiatry invites one to consider whether the hairs upon one's neck stand up in the presence of such a man; the good observer reads their own responses to the situation.

When we turn to disordered movement, we usually mean more pronounced and qualitatively distinct disturbances of behavior. These may be divided into voluntary and involuntary movement disorders.

Voluntary Movements

What can a subject do voluntarily that is abnormal? They can do strange things (jump in the river) or they can do ordinary things in a strange way (stand, waiting for the bus while making a Nazi salute). Patients come to attention in public places usually as a consequence of doing something noteworthy (e.g., standing on the bridge looking at the water; driving the wrong way up the motorway; talking to themselves; exposing themselves or being 'threatening'). In many of these cases there is nothing abnormal about the movements *per se*, it is where their action sequence is leading, the goal of such movements that is abnormal (wanting to drown, crash, or masturbate in public). It is also worth recalling that intoxication (with either alcohol or drugs) may disinhibit and 'magnify' the expression of dysfunction.

However, some voluntary movements do appear unusual. In 'mannerisms' a normal task is performed in an abnormal way, a way that might appear symbolic to some (e.g., beckoning for the bus, making a Nazi salute, above). Mannerisms seem to have a purpose (drinking tea from a cup, with one's little finger raised used to be regarded as quite stylish; certain forms of handshake carry meaning in 'normal' circles). In 'stereotypy' a voluntary behavior is repeatedly performed, yet it seems purposeless (e.g., rocking to and fro in the chair, rocking one's leg at the knee while sitting cross-legged). Sometimes the behavior may merely release tension, as when a patient treated with neuroleptic medications experiences '*akathisia*' (subjective motor restlessness), a state that may be most distressing: never being able to sit still or relax. Sometimes the patient looks restless, but does not feel it; they are rocking to and fro, but cannot say why; when asked to perform a purposeful task while standing, for example, alternating patterns of hand movement, they may 'tramp' on the spot (pseudoakathisia).

Mostly these behaviors are encountered in people with schizophrenia (F20). It is a moot point whether they are entirely attributable to the disorder, partially driven by its treatment (especially neuroleptic-induced akathisia), or whether the more exotic, historic descriptions of catatonic behavior (F20.2) were facilitated by institutionalized living conditions.

Catatonia incorporated a wide variety of signs that are seldom seen today and, when they are seen, rarely coincide as a syndrome: hence, the subject may exhibit 'waxy flexibility,' allowing an examining doctor to move his limbs about, into awkward postures, 'negativism,' resisting the examiner's movements, 'obedience,' doing what the examiner says even when told not to do so, 'mitgehen,' allowing oneself to be moved even when told not to, 'ambitendence,' appearing to alternate between mutually exclusive courses of action (hovering on the threshold, attempting to walk into and out of the room at the same time), 'psychic pillow,' appearing to sleep with the head raised above the bed, 'posturing,' apparently staying in a fixed position for long periods of time.

It is hard to consider these behaviors without noting both the straitened circumstances experienced by such patients, incarcerated within asylums, and the emphasis (in these descriptions) placed upon their obeying or not obeying their examiners' commands.

Other behaviors may look abnormal, but are performed normally, for instance compulsive handwashing (F42.1). The behavior is voluntary, but it is performed under some duress, the patient having tried to resist its performance. Similarly, certain habits can attract psychiatric attention: compulsive hair pulling (F63.3), nail biting, binge eating (F50.2), drinking, and so on. In tics, a recurrent, rapid movement of a group of muscles occurs (F95), sometimes without thinking, sometimes in response to an urge to act. In Tourette's syndrome,

the tics performed vary from those that are simple, for example, coughing, blinking, head turning, to those that are complex: vocal ejaculations of obscenities (coprolalia) or similarly obscene gestures (copropraxia). Sometimes patients can learn to 'hide' the abnormal behavior by morphing it into another gesture or obscuring words with a cough.

Involuntary Movements

The simplest involuntary movement is a reflex, which may or may not be abnormal (the patella tap, the knee-jerk is normal; while a grasp reflex of the fingers is normal in a baby, but not in an adult). Tremor, also, may be normal or abnormal. Thus we may tremble if nervous or frightened, we may shiver when cold. Pronounced tremors may be seen at 'rest' or during action ('intention tremors'). The former include the normal variant essential tremor (snooker players may allay this with alcohol or beta-blocker drugs), and the pathological tremors seen in Parkinson's disease ('pill-rolling tremor') or iatrogenically induced, following neuroleptic medication (e.g., haloperidol). Tremors confined to intention are classically associated with cerebellar disease (e.g., postalcohol dependence).

More pronounced involuntary movements ('dyskinesias') may attract attention to the subject and be very embarrassing. 'Choreiform' movements are proximal (involving shoulder and hip joints), jerky and 'dance-like.' They are seen in Huntington's disease (F02.2), Sydenham's chorea (following streptococcal infection in children) and, rarely, antipsychotic medication. 'Athetoid' movements are distal (involving hands, wrists, feet, ankles), sinuous and snake-like; they may be encountered in cerebral palsy, poststroke, and, rarely, postneuroleptics. Choreiform and athetoid limb movements appear to lack purpose. However, some involuntary hand movements may appear purposeful even if they are not: in the 'anarchic hand' syndrome, a brain lesion 'releases' the contralateral hand to grasp objects in the environment. To the subject the hand appears to have 'a mind of its own.' In a right-handed subject a left premotor lesion may release a grasping right hand, while a right premotor or callosal lesion may release an interfering left hand; the subject exhibits 'intermanual conflict,' the left hand may undo buttons while the normal right hand tries to do them up, the left may put down the cup, which the right is attempting to raise.

Tardive dyskinesias (so-called because they come on slowly after the initiation of neuroleptic medication) may affect the face most noticeably. A patient exhibits abnormal chewing movements, protrusion of the tongue, and head and neck turning. While 'top-down' processes may hold the process in check while the subject is alert and unoccupied, distraction with a difficult task (e.g., naming the months of the year in reverse order) may unmask the abnormal movements (an important element of clinical assessment). Similar dyskinesias may affect the fingers and wrists, repetitive movements being more noticeable while the patient is walking.

Dystonias are abnormalities of posture caused by abnormalities of motor tone. In torticollis, the head is turned to one side because the contralateral neck muscles are in spasm. In writer's cramp the muscles of the hand are affected; in blepharospasm, those external to the eyes.

In Parkinson's disease the increased tone of the muscles gives rise to rigidity, which can be felt upon bending the limbs. These signs may also be found in those receiving neuroleptics, and chronic administration may give rise to unusual truncal postures: Pisa syndrome, leaning to one side, Rabbit syndrome, flexing the trunk and arms together.

It is important to note that although disorders of movement may be most apparent to observers and less noticeable to the patient, some may be very distressing indeed. Hence, acute dystonia, a painful arching of the neck and back, following acute administration of neuroleptic medication, and associated with occulogyric crises (in which the eyes roll upwards under the upper eyelids), is acutely distressing; while akathisia, a sense of motor restlessness, increases the risk of suicide.

Disorders of Personality

While it might be relatively easy to assess a patient's speech and movement upon first meeting them, it is a great deal more difficult to assess their

personality. Indeed, even after repeated contacts and recourse to coinformants one might still not know another person's personality. Descriptive psychopathology tends to adopt a typological approach to personality, categorizing abnormal personalities into 'types,' little more than caricatures really, listing criteria that must be satisfied, though often these are rather blunt terms. More realistic is a dimensional approach to those aspects of feeling, behaving, and relating, which constitute a person. So it would probably be more valid to regard all humans as located on a continuum when it comes to extraversion/introversion, paranoia, empathy, moodiness, and obsessionality, so that each personality might occupy a point in a highly multidimensional space (there being many facets to a personality). However, clinically, we are required to provide sketches, often on the basis of inadequate information.

What is a personality? It is a pattern of temperament, behavior, and relation to others that is probably established by teenage years and which remains generally stable over the long term. Under approximately equal genetic and environmental influence, it seems that we become 'more genetic' as we age, 'more like ourselves,' a finding that has been attributed to the human ability to create our own environments as we become autonomous. It is also clear that, in the philosophical sense, we are subject to 'moral luck': we did not choose our genes or our early environments, so 'moral responsibility' is probably at best a relative condition: some have more or better choices than others, but few might be said to be 'free!'

In terms of psychopathology, a personality disorder represents an abnormal pattern of temperament, behavior, and relation to others, which is usually in place by teenage years, as a consequence of which, either the person or their society suffers. It is a controversial topic. Some might argue that it constitutes the medicalization of 'deviance.' Some might exclude it from psychiatry altogether. Currently, the UK government places great emphasis upon attempts to treat those with what it calls 'dangerous and severe personality disorders.' Evidence of response to treatment is limited.

Table 10 shows those personality disorders currently recognized by the prominent diagnostic systems. They may be 'clustered' into those people who seem cut off from others (Cluster A), those who have tempestuous relationships with them (Cluster B), and those for whom relationships are unsatisfactory, emotionally (Cluster C). This is a rough and ready categorization, but it is interesting to note that those personalities with whom psychiatrists probably have the most contact are those harming themselves (borderline personality disorder, F60.3) and others (antisocial personality disorder, F69.2); both located in Cluster B. Though other personalities may suffer, they tend not to present to psychiatric care as frequently.

Difficult Situations

A recurring theme of this article has been the central problem in psychopathology, which is aligned with the 'problem' of consciousness itself: we do not know what others are thinking. Just as scientists and philosophers cannot explain how a mechanistic, and apparently deterministic, universe generates awareness, or devise a means by which one person could truly know what it is like to be another (or indeed a bat!), thereby limiting discourse to an attempted correlation between 'third-person' observations (of an object) and 'first-person' experiences (of a subject), so also psychopathology attempts to describe and diagnose what it is that someone else is experiencing. In this regard, as I have stated already (above) there is something distinguishing psychiatry from other branches of medicine: we are uniquely concerned with the phenomenological, not merely as a signpost on the way to reaching a diagnosis, but as our primary concern: it is, after all, the mental state that we are attempting to 'treat': we aim to change phenomenology. Admittedly, we have the 'clues' offered by the subject's speech and behavior, but it is obvious that we operate at several removes (and indeed dimensions) from the object of our inquiry, (We are a first-person subject cogitating upon our third-person observations of another first-person subject, possibly supplemented by the evidence of other subjects in our midst). Given that most psychiatric diagnoses describe what people think and the reasons for their actions this represents

Table 10 Clinically encountered personality disorders

Disorder	Features
Cluster A	
Paranoid	Suspicious of others, sensitive to slights, often feels 'put down'; bears grudges. Suffers as a consequence of perceived exclusion.
Schizoid	Cold affect, prefers own company, lives in 'own world' mentally, insensitive to praise or criticism. Not thought to suffer from isolation.
Schizotypal	Controversial concept, possible mild variant of schizophrenia; subject may be particularly interested in occult, UFOs, alternative or eccentric lifestyles, prone to illusions, magical thinking, unusual patterns of speech, may seem cold and aloof.
Cluster B	
Antisocial	Recurrent breaker of society's rules, impulsive, reactive aggression, lacks 'sense of responsibility,' has difficulty sustaining relationships with others. At severe end, psychopaths may engage in instrumental violence, sadism, lying, cheating, and lack remorse for conduct or empathy for victims.
Borderline	Unstable sense of self, frequent decompensations, often connected with relationships, prone to feelings of abandonment, recurrent self-harm, may experience short bouts of very low mood, suicidal ideation, low self-esteem, confusion over sexuality, problems with impulse control: binge eating, use of drugs and alcohol, stealing, promiscuity, gambling.
Histrionic	Superficial expression and experience of emotions ('theatricality'), but appears shallow, needs to be the center of attention, given to displays of decompensation. Suggestible, preoccupied with physical attractiveness.
Narcissistic	Feels greater self-worth than others, sense of entitlement, resentful of others, fantasizes about success, may lead 'Billy Liar' lifestyle, spending money he does not have, pretending to be a very important personage. Lacks empathy and uses others.
Cluster C	
Obsessive–compulsive (anankastic)	Rigid character, stubborn, pedantic, preoccupation with perfectionism that limits completion of tasks, needs routine, unable to adapt to change. May appear 'sticky' in relationships and thinking (things must be done 'his way'), unable to 'let things go.'
Avoidant	Sensitive to criticism, avoids social contact out of fear of rejection. Suffers as a consequence. Anxious. Limits contact to those where 'sure to be liked.'
Dependent	Overly reliant upon others, requires much support. Afraid of being left alone, avoids critique of those relied upon, has others make decisions.

something of an impasse. We are always operating at the level of inference.

It follows then that there may be several ways in which this system of psychopathological communication may break down. Some are obvious: we may speak different languages, come from different cultures, and carry different assumptions about the world. We may be 'there' not because we want to be, but because someone else has requested the meeting: the benefits agency, the insurance company, the custody sergeant, the duty doctor, the prison staff, the defense counsel, or the judge. There may be many reasons why the subject doubts or mistrusts the psychopathologist.

Therefore, there is always a risk of misunderstanding and of deception. Indeed, the frailties of psychiatric diagnoses were exposed by the classic experiment of D.L. Rosenhan who, with others, gained admission to psychiatric hospitals in the United States in the 1970s, while claiming to 'hear voices' (they did not hear them, they were well). They concluded that there was a problem with psychiatric diagnosis, because they were believed, uncritically. More recently, Peter Tyrer and colleagues have been concerned by what they term 'instrumental psychosis,' the apparent relapse into psychoses of people known to have had such illnesses in the past, but for whom the current 'episode' is, what one might call, 'convenient,' for example, it circumvents a court appearance. Is there anything surprising about any of this? No, not really. It is the price that we pay for attempting to understand phenomena that only a 'subject' can ever truly know.

Similar issues are played out in casualty departments at night, whenever a duty psychiatrist attempts to discern whether another human being 'really' intended to 'kill herself': whether she knew that she would be found, whether she knew how many tablets might have proven fatal, whether there was an argument, whether she told anyone what she was thinking of doing, who it was that

called the ambulance, whether there was a note and what it said, did she feel remorse? We have to infer the answers, but we must also be humane. More important than being strictly accurate or 'intellectually correct' is the desire to save life; to admit a patient 'unnecessarily,' to 'err on the safe side,' might not concur with what we actually believe did happen (or what we are encouraged to do by health service managers) but we are there to save a life, not to conduct an experiment.

We played out these issues also in the discussion of the word 'functional,' where we pointed to the difficulty in determining whether physical states, behaviors that resemble illnesses, but for which there is no discernible 'organic' cause, are attributable to the unconscious processes hypothesized in hysteria or the conscious intention of a subject who is deceiving us, malingering. We may try to detect the difference, while diagnostic systems maintain such distinctions, but it is clear that objective proof of one state above another is largely hypothetical.

So, is there such a thing as 'silent schizophrenia'? Only if psychopathology ceases to be about mental states, about what a consciousness perceives. Until then, schizophrenia is an experience, experienced by a subject, the subject whom we meet, in the room at the clinical interview.

See also: Phenomenology of Consciousness; Psychodynamic Theories of the Unconscious.

Suggested Readings

Amador X and David A (eds.) (2004) Insight and Psychosis: Awareness of Illness in Schizophrenia and Related Disorders. Oxford: Oxford University Press.

Berrios GE (1996) The History of Mental Symptoms: Descriptive Psychopathology Since the Nineteenth Century. Cambridge: Cambridge University Press.

Breuer J and Freud S (1974) Studies on Hysteria, Strachey J (trans.), The Penguin Freud Library, vol. 3. London: Penguin.

Cutting J (1987) The phenomenology of acute organic psychosis: Comparison with acute schizophrenia. British Journal of Psychiatry 151: 324–332.

Enoch D and Trethowan W (1991) Uncommon Psychiatric Syndromes, 3rd edn. Oxford: Butterworth-Heinemann Ltd.

Evans D (2001) Emotion: A Very Short Introduction. Oxford: Oxford University Press.

Gelder M, Mayou R, and Cowen P (eds.) (2001) Shorter Oxford Textbook of Psychiatry, 4th edn. Oxford: Oxford University Press.

Hare RD (1999) Without Conscience: The Disturbing World of the Psychopaths Among Us. New York: Guilford Press.

Jaspers K (1963) General Psychopathology. Hoenig J and Hamilton MW (trans.). Manchester: Manchester University Press.

Leudar I and Thomas P (2000) Voices of Reason, Voices of Insanity: Studies of Verbal Hallucinations. London: Routledge.

Pawar A and Spence SA (2003) Defining thought broadcast: Semi-structured literature review. British Journal of Psychiatry 183: 287–291.

McKenna P and Oh T (2005) Schizophrenia Speech: Making Sense of Bathroots and Ponds that Fall in Doorways. Cambridge: Cambridge University Press.

Mullins S and Spence SA (2003) Re-examining thought insertion: Semi-structured literature review and conceptual analysis. British Journal of Psychiatry 182: 293–298.

Rosenhan DL (1973) On being sane in insane places. Science 179: 250–258.

Spence SA (1999) Hysterical paralyses as disorders of action. Cognitive Neuropsychiatry 4: 203–226.

Spence SA (2006) All in the mind? The neural correlates of unexplained physical symptoms. Advances in Psychiatric Treatment 12: 349–358.

Spence SA, Crimlisk HL, Cope H, Ron MA, and Grasby PM (2000) Discrete neurophysiological correlates in prefrontal cortex during hysterical and feigned disorder of movement. Lancet 355: 1243–1244.

Spence SA and Halligan PW (eds.) (2002) Pathologies of Body, Self and Space. Hove: Psychology Press.

Tyrer P, Babidge N, Emmanuel J, Yarger N, and Ranger M (2001) Instrumental psychosis: The Good Soldier Svejk syndrome. Journal of the Royal Society of Medicine 94: 22–25.

World Health Organization (1992) The ICD-10 Classification of Mental and Behavioural Disorders. Geneva: World Health Organization.

Biographical Sketch

Sean A Spence is a professor of general adult psychiatry at the University of Sheffield, where he holds an MRC Career Establishment Grant. He was previously De Witt-Wallace visiting research fellow at Cornell University, New York (1999), and MRC clinical training fellow at the MRC Cyclotron Unit, Hammersmith Hospital (1995–98). He has won the Royal College of Psychiatrists Research Prize and Medal (1997) and Royal Society of Medicine Section of Psychiatry Essay Prize (1997). He studied medicine at Guy's Hospital, London, acquired an intercalated BSc in psychology, and won the Gillespie Prize in psychological medicine (1986). His principal research interest is in the regulation of voluntary behavior (volition) in healthy subjects and those affected by neuropsychiatric disease. He is a coauthor of *Managing Negative Symptoms of Schizophrenia* (Science Press, 2001), a coeditor of *Pathologies of Body, Self and Space* (Psychology Press, 2002), and of *Voices in the Brain: The Cognitive Neuropsychiatry of Auditory Verbal Hallucinations* (Psychology Press, 2004). In 2001 the 2nd Reunion de investigacion en psicopatologia, in Cadiz, was devoted to his volitional studies. In 2007, a television series entitled 'Lie Lab' was devoted to his work on fMRI of deception (Channel 4 UK). His clinical work is with the homeless.

Religious Experience: Psychology and Neurology

H Roggenkamp, M R Waldman, and A B Newberg, University of Pennsylvania, Philadelphia, PA, USA

Glossary

Autonomic nervous system – A part of the nervous system that connects the brain to the body and regulates various organ systems to prepare for an arousal or calming response.
Deafferentation – Blocking input, usually neuronal, into one or more structures in the brain.
Electroencephalography – A method for measuring the electrical changes that occur in the brain.
Functional magnetic resonance imaging (fMRI) – A technique for performing scans to show activation of different parts of the brain using magnetic fields rather than radioactivity.
Neurochemicals/Neurotransmitters – Refers to different substances such as serotonin and dopamine that are chemicals or molecules that are required for the functioning of the brain.
Phenomenology – Relating to the subjective experiences of the world that people have.
Positron emission tomography (PET) – A technique for performing scans of various physiological processes in the brain that requires initial administration of a radioactive material, similar to single photon emission computed tomography (SPECT).
Posterior superior parietal lobe – An area of the brain located toward the back (posterior) and top (superior) that is involved in spatial orientation of the body.
Prefrontal cortex – Part of the brain's frontal lobe that is highly involved in focusing attention and executive functions.
Single photon emission computed tomography (SPECT) – A technique for performing scans of various physiological processes in the brain that requires initial administration of a radioactive material, similar to positron emission tomography (PET).
Thalamus – A key central structure in the brain that is a crucial relay between sensory organs and the brain and between different brain structures to each other.

Introduction

Spiritual Experience as a Means for Studying Consciousness

Religion has served as a vital influence on the lives of humans since the beginning of modern man and a plethora of practices incorporate people's ongoing quest to be closer to what they perceive as God. Religious experiences, particularly those that are transformative, have often been described as involving an alteration in the usual sense of consciousness. The study of religious and spiritual experiences is ultimately the study of complex mental processes. Thus, the study of religious and spiritual experiences is also potentially one of the most important areas of research that may be pursued by science in the next decade. This may not be an understatement since such experiences offer a fascinating window into human consciousness and psychology, the relationship between mental states and body physiology, emotional and cognitive processing, and the biological correlates of religious and spiritual experiences.

At one end of the continuum of religious experiences is the state that might be called 'Absolute Unitary Being' which is described in the mystical literature of all the world's great religions. When a person is in that state he or she loses all sense of

discrete being, even to the point where the difference between self and other is obliterated. There is no sense of the passing of time, and all that remains is a perfect timeless undifferentiated consciousness. When such a state is suffused with positive affect, there is a tendency to describe the experience, after the fact, as personal. Such experiences are often described as a perfect union with God (the 'Unio mystica' of the Christian tradition) or the perfect manifestation of God (the attainment of 'moksha' in the Hindu tradition). When such experiences are accompanied by neutral affect they tend to be described, after the fact, as impersonal. These states are described in concepts such as the Void or Nirvana of Buddhism or the Absolute of a number of philosophical/mystical traditions.

There is no question that whether the experience is interpreted personally as God or impersonally as the Absolute, it nevertheless possesses a quality of altered consciousness – a transcendent wholeness without any temporal or spatial division. There seem to be neurobiological correlates of such states of consciousness, such as the inhibition of sensory input into the posterior superior parietal lobe (PSPL), especially on the right. This area of the brain is responsible for the orientation of objects in three-dimensional space. If it is denied of all sensory input, as a result of mechanisms generated during practices such as profound meditation, the result may be a sense of pure space. Since space has no objective reality (unless it relates things to each other), the subjective experience is one of total spacelessness, or of total perfect unity. Research on brain areas involved during states of absolute unity in the context of meditation will be discussed in more detail later in the article.

Typically, some form of meditation or concentration is used to bring the individual into this altered state of consciousness. Although there are many styles and forms of meditation, we have typically divided such practices into three basic categories. The first category is one in which the subjects simply attempt to clear all thought from their sphere of attention. This form of meditation is an attempt to reach a subjective state characterized by a sense of no space, no time, and no thought. Further, this state is cognitively experienced as fully integrated and unified such that there is no sense of a self and other. The second category is one in which the subjects focus their attention on a particular object, image, phrase, or word, and includes practices such as transcendental meditation (TM) and various forms of Tibetan Buddhism. This form of meditation is designed to lead to a subjective experience of absorption with the object of focus. The third category of meditation can be referred to as mindfulness, in which the person observes without judgment the uninterrupted flow of thoughts and feelings as they consciously rise and fall in the mind. This form of meditation creates what is known as an 'observing self,' which allows the person to feel emotionally neutral. These three forms of meditation tend to overlap, as concentrative practices can also incorporate mindfulness and clearing of thoughts.

Meditation can be practiced in many different conditions. Qigong, an ancient Chinese meditation practice, coordinates breathing and focusing on certain parts of the body, creating a very dynamic physical form of meditation. Qigong is practiced sitting, standing, and during movement, which makes it similar to Yogic meditation. Speaking in tongues (glossolalia) is a religious experience that often involves singing and dancing, further extending the range of activities that can be undertaken to induce altered states of consciousness associated with religious experience.

There is another distinction in which some meditation is guided by following along with a leader who is verbally directing the practitioner either in person or on tape. Others practice the meditation on their own volition. This difference can be seen in specific differences in cerebral activation while practicing volitional or guided meditation. Despite differences in practice methodology, phenomenological analysis suggests that the end result of many religious and spiritual practices is similar. However, this result might be described differently using unique characteristics depending on the culture and individual. Therefore, it seems reasonable that while the initial neurophysiological activation occurring during any given practice may differ, there should eventually be a convergence.

Consciousness and Religious Experience can be Explored via Brain Imaging and Activation Studies

Relevant Imaging and Activation Modalities

For any study of human consciousness, one would have to begin by defining the specific operationalized paradigm through which the experience itself would be explored. Such studies can range from the evaluation of various sensory experiences to determine when, where, and how such experiences enter consciousness. Religious practices such as meditation, which are specifically designed to alter consciousness, can be explored scientifically as well. Different states of consciousness can theoretically be studied using various brain imaging techniques. Depending on what elements of consciousness are the focus of a given study, each of the functional brain imaging techniques provides different methodological advantages and disadvantages. The currently available functional brain imaging techniques include functional magnetic resonance imaging (fMRI), single photon emission computed tomography (SPECT), and positron emission tomography (PET). fMRI requires a large magnet in order to measure changes in cerebral blood flow (CBF) and oxygen content. SPECT and PET require the injection of various types of radioactive tracers that can measure a variety of neurophysiological functions including blood flow, metabolism, and neurotransmitter changes.

fMRI has improved spatial resolution over SPECT, and probably PET, as well as the ability to perform immediate anatomic correlation. Thus, functional MRI can produce functional images that can be directly mapped onto the anatomical structures, thereby helping to pinpoint areas of the brain associated with various activation tasks (PET and SPECT can also be co-registered with structural MRI images, but this often involves complex computer programs to overlay the functional image with the anatomical ones). Another advantage of fMRI is the excellent temporal resolution, so that multiple scans can be made over the course of a single imaging session. However, the major disadvantage of fMRI is that the individual is required to lie down in a confined space and it makes a significant amount of noise, both of which could be very disturbing to various types of practices such as meditation or could interfere with other religious practices that require specific movements or dancing.

PET imaging provides better resolution than SPECT, and has the important ability to make quantitative measures of activity. This could be very important since certain tasks or experiences may either diminish or augment activity in some areas and not others, and this may be missed if absolute quantitation is not performed. However, if one strives to make the environment relatively distraction free, it is sometimes beneficial to perform these studies after hours, which may complicate the use of PET because radiopharmaceuticals such as fluorodeoxyglucose (FDG) may not be readily available. Furthermore, PET is the most expensive of the imaging modalities. One other potential problem is that the most common tracer used in PET imaging, FDG, has an uptake period of approximately 20–30 min, which may prevent the ability to capture a momentary state. SPECT imaging is the most readily available and rivals MRI in low cost per study and also has the advantage, as well as PET, in being able to study subjects outside of the scanner.

For example, SPECT studies can enable the subjects to meditate or pray until they experience a ‘peak’ in their practice. At this point, the subject can be injected with a radioactive tracer through an indwelling intravenous catheter while they continue to practice. The tracer is fixed in the brain at the time of injection so that when the images are acquired approximately 20 min later, the images reflect the CBF during ‘peak’ meditation. Different states can then be compared. These studies have helped toward the development of complex neurobiological models of practices such as meditation, which can dramatically alter a person's consciousness. Current models include not only a variety of structures including the prefrontal cortex (PFC), parietal lobe, limbic system, thalamus, and hypothalamus, but now include neurotransmitters such as glutamate, gamma aminobutyric acid (GABA), dopamine, and serotonin, as well as hormonal, autonomic, and immune changes (see Table 1 – neurochemical table).

Electroencephalography (EEG) is another method for evaluating the brain's activity, although

Table 1 Neurochemically related changes in serum concentration observed during meditation techniques and the central nervous system (CNS) structures typically involved in their production

Neurochemical	Observed change	CNS structure
Arginine vasopressin	Increased	Supraoptic nucleus
GABA	Increased	Thalamus, other inhibitory structures
Melatonin	Increased	Pineal gland
Serotonin	Increased	Dorsal raphe
Cortisol	Decreased	Paraventricular nucleus
Norepinephrine	Decreased	Locus ceruleus
b-Endorphin	Rhythm changed; levels unaltered	Arcuate nucleus

it is not typically considered an imaging technique. EEG reflects cortical activity through recording electrical signals via electrodes placed on the scalp, and primarily measures neuronal activity immediately beneath the cortex. Alpha rhythm waves are 8–12 Hz frequency signals that are related to relaxation and lack of active cognitive processes. Beta wave signals, characterized by rhythms from 18 to 30 Hz, occur when one is alert and focused; frequencies from 30 to 70 Hz are termed gamma activity. Gamma waves seem to be present when the brain is working to integrate a variety of stimuli into a coherent whole and during peak performance. Between the ranges of 4 and 8 Hz one sees theta activity, which is correlated with rapid eye movement sleep, problem solving, overall attention, and hypnosis. Delta activity is seen especially in infants in the first two years of life, and is a low frequency wave of 0.5–4 Hz. It tends to be associated with sleep as well. EEG recordings reflect the difference of voltage between the signals at two electrodes.

Artifacts can confound EEG results very easily: EEG voltages are extremely small, and must be amplified by a factor of one million in order to be measured. Therefore, it is easy to mistake electrical interference that originates from the electrodes, the recorder, or electric lights, as activity in the brain. EEG is best used when paired with neuroimaging modalities such as PET, as it cannot measure activity in deeper areas of the brain such as the brainstem; it primarily shows activity of pyramidal neurons of the cortex.

Several newer methods are also being developed such as optical imaging or near infrared imaging in which a light probe is placed on the skin surface and the change in light received by an adjacent detector reflects changes in the brain's blood flow. This technique has excellent temporal and good spatial resolution and is very portable. Its main disadvantage is that it cannot observe deep structures.

Problems with Relating Neurobiology to Religious Experiences

There are several generic methodological issues with brain imaging studies of various elements of conscious experience. The most problematic is not so much what is measured on the scan, but whether the scan corresponds to the subjective state that one is trying to measure. In the example of studying religious experiences, it typically is not possible to 'interrupt' the practitioner during the practice, and therefore it can never be known if the scan corresponds to a specific subjective state of consciousness or a specific experience. Studies that attempt to measure changes in the brain associated with sensorial experience presenting in consciousness require the subjects to indicate when the experience does in fact enter into their consciousness. The problem here is that there is a necessary delay between when they experience something and when they actually respond. Furthermore, the response itself can alter brain physiology. Thus, neurobiological research into the subjective nature of consciousness is quite difficult and any findings need to be interpreted carefully. On the other hand, neuroimaging techniques continue to make major advances and provide the best window into the underlying neurobiological processes associated with consciousness.

Besides practical questions regarding imaging modalities, there are additional concerns regarding inferences about brain activation data and religious experience. Up to the present, consciousness has been understood as referring to consciousness of something. The existence of a state like pure consciousness or absolute unity, consciousness devoid of content which meditators often report experiencing, has generally not even been entertained as a problem. Such mystical experiences are devoid of the perception of

discrete reality, and have no sense of the passage of time, no sense of the extension of space, and no sense of the self-other dichotomy. What is particularly interesting about the state is that neither during the experiencing of pure consciousness, nor upon subsequent recollection, is this state ever perceived as subjective. Although it is attained by going deeply within the subject, once it is attained, it is perceived as neither subjective nor objective. In other words, the state consists of an absolute sense of unity without thought, without words, without sensation, and not even being sensed to inhere in a subject. Such a state may have a specific neurobiological correlate, supported by preliminary brain imaging research of meditative states, but more data is necessary to explore the implications of this state more deeply. Further, it may still remain to be seen whether the biological correlates explain the experience or are merely evaluating how the brain responds to an altered consciousness state.

Brain Activation in Religious Experiences

The following review is presented as a working hypothesis regarding the brain's function during various religious and spiritual practices and experiences. This hypothesis incorporates recent

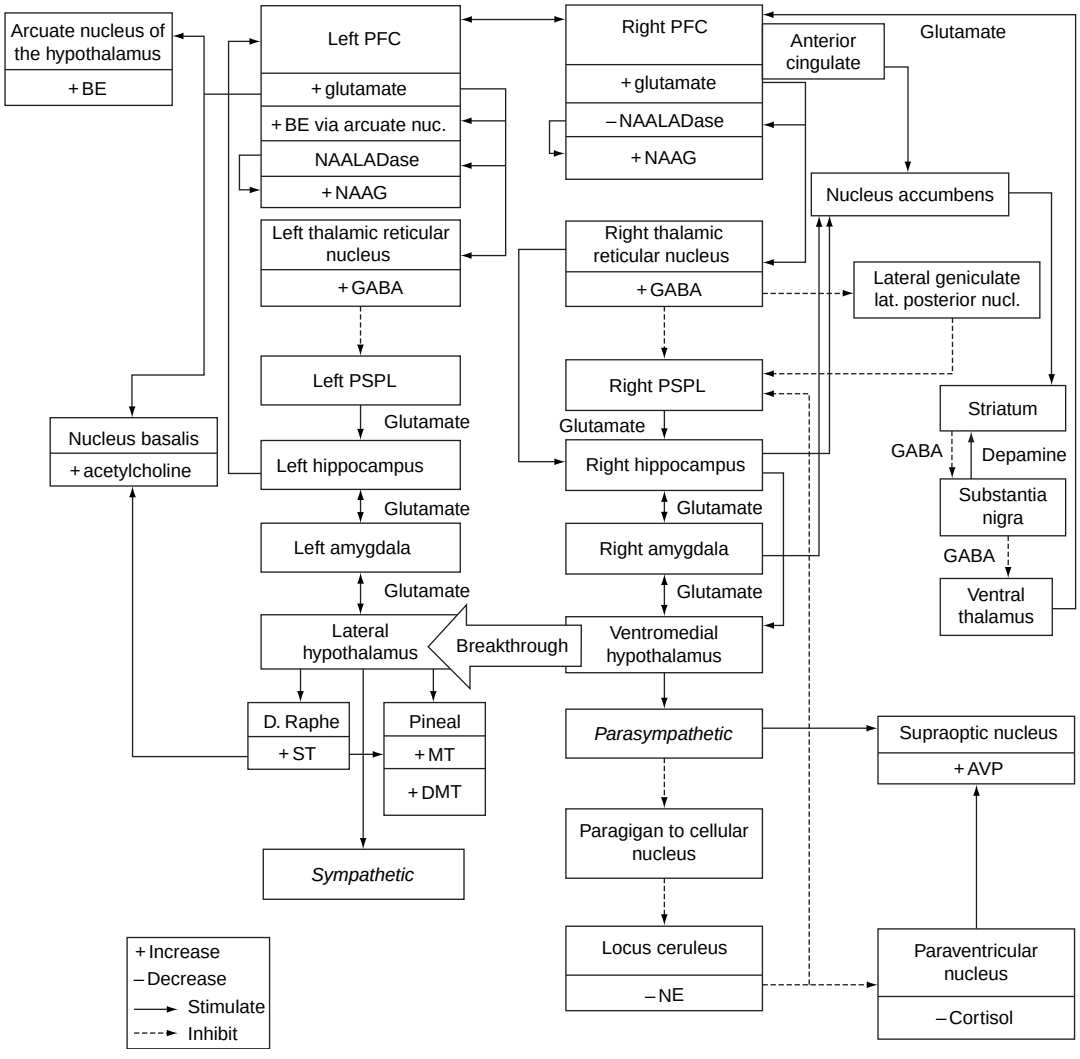


Figure 1 Schematic overview of the neurophysiological network, possibly associated with meditative states and experiences.

neuroimaging, neurochemical, hormonal, and physiological studies. The purpose of this review is to provide a foundation from which many different types of religious experiences and practices can be considered and compared. As shown in Figure 1, the schematic representation of the involved pathways begins with the PFC and suggests a number of complex interactions with the thalamus, PSPL, limbic system, and autonomic nervous system (ANS). It should be noted that not all structures that might be involved in religious and spiritual experiences are included, especially since new structures such as the insula and claustrum are just now being hypothesized to be involved. Furthermore, a number of both excitatory and inhibitory neurotransmitters can now be proposed to play a role in such practices and experiences. Dopamine, serotonin, acetylcholine, and several other molecules may be associated with various phenomenological aspects of such experiences and these are also considered in this model. It would be anticipated that depending upon the specific practice, the ritual, tradition, and individual involved in the specific mechanisms might be somewhat different. However, focusing on the phenomenology of such experiences, this model provides information regarding the diversity of experiences, both sensory, cognitive, and affective, that can be associated with religious and spiritual experiences. The model was initially developed utilizing information from studies primarily on meditative practices due to the relatively large amount of data available. However, this model can likely be applied to many different types of practices and experiences as they result in altered states of consciousness.

Activation of the Prefrontal and Cingulate Cortex

Since some approaches toward attaining spiritual experiences involve practices in which the person on their own volition attempts to enter into such a state, we will begin with the parts of the brain that might be associated with the human will. Brain imaging studies suggest that willful acts and tasks that require sustained attention are initiated via activity in the PFC, particularly in the right hemisphere. The cingulate gyrus has also been shown to be involved in focusing attention, probably in conjunction with the PFC. Since transformative types of practices such as meditation require intense focus of attention, it seems that there should be activation of the PFC as well as the cingulate gyrus. This notion is supported by the increased activity observed in these regions on several of the brain imaging studies of volitional types of meditation including that from our laboratory. In a study of eight Tibetan Buddhist meditators, the subjects had an intravenous line placed and were injected with a CBF tracer while at rest, in order to acquire a 'baseline' image. They then meditated for approximately 1 h when they were again injected with the tracer while they continued to meditate. The tracer was fixed in the brain at the time of injection so that when the images were acquired approximately 20 min later, they reflected the CBF during the meditation state. These images demonstrated increased activity in the PFC bilaterally (greater on the right) and the cingulate gyrus during meditation. Therefore, meditation appears to start by activating the prefrontal and cingulate cortex associated with the will or intent to clear the mind of thoughts or to focus on an object. Studies of other forms of meditation, such as Yoga Nidra, Kundalini Yoga, TM, and Christian prayer show an increase in activation in the frontal and prefrontal cortices as well. Of course, meditation does not necessarily lead to transformation, so that frontal lobe activity alone does not seem to provide a complete understanding of these experiences. In fact, some transformative experiences are associated with feelings in which the self feels 'overtaken' by the experience itself. The self feels as if it becomes absorbed or lost into the experience. If the frontal lobe is involved in a willful activity, then a transformative experience associated with a feeling of losing the self or of the original self being overcome by something spiritual, might be associated with a lack of activity in the frontal lobes. A decrease in frontal lobe activity is seen in Pentecostal Christian glossolalia practitioners, a practice involving speaking in tongues. It would seem that glosssolalia practitioners may not be focusing their attention when speaking in tongues; they report feeling that they can no longer control what is happening. More research is needed that compares different practices and their effects on frontal lobe activity.

Thalamic Activation

Several animal studies have shown that the PFC, when activated, innervates the reticular nucleus of the thalamus, particularly as part of a more global attentional network. Such activation may be accomplished by the PFC's production and distribution of the excitatory neurotransmitter glutamate which the PFC neurons use to communicate among themselves and to innervate other brain structures. The thalamus itself governs the flow of sensory information to cortical processing areas via its interactions with the lateral geniculate and lateral posterior nuclei and also likely uses the glutamate system in order to activate neurons in other structures. The lateral geniculate nucleus receives raw visual data from the optic tract and routes it to the striate cortex for processing. The lateral posterior nucleus of the thalamus provides the posterior superior parietal lobule (PSPL) with the sensory information it needs to determine the body's spatial orientation.

When excited, the reticular nucleus releases the inhibitory neurotransmitter GABA onto the lateral posterior and geniculate nuclei, cutting off input to the PSPL and visual centers in proportion to the reticular activation. During practices such as meditation, because of the increased activity in the PFC, particularly on the right, there should be a concomitant increase in the activity in the reticular nucleus of the thalamus. While brain imaging studies of meditation have not had the resolution to distinguish the reticular nuclei, our recent SPECT study did demonstrate a general increase in thalamic activity that was proportional to the activity levels in the PFC. This is consistent with, but does not confirm the specific interaction between the PFC and reticular nuclei. If the activation of the right PFC causes increased activity in the reticular nucleus during meditation, the result may be decreased sensory input entering into the PSPL. Several studies have demonstrated an increase in serum GABA during meditation, possibly reflecting increased central GABA activity. This functional deafferentation related to increased GABA would mean that fewer distracting outside stimuli would arrive at the visual cortex and PSPL enhancing the sense of focus. During meditative experiences, similar mechanisms might be in place although the interaction between the frontal lobe and the thalamus might be more complex such that some states may be associated with mutual increases, some with mutual decreases, and some with one turned on and the other off. The mechanism of the later is less clear at this time.

It should also be noted that the dopaminergic system, via the basal ganglia, is believed to participate in regulating the glutamatergic system and the interactions between the PFC and subcortical structures. A recent PET study utilizing 11C-Raclopride to measure the dopaminergic tone during Yoga Nidra meditation demonstrated a significant increase in dopamine levels during the meditation practice. They hypothesized that this increase may be associated with the gating of cortical-subcortical interactions that leads to an overall decrease in readiness for action that is associated with this particular type of meditation. Similarly, there might eventually be a reciprocal decrease in PFC activity associated with the inhibitory function of the thalamus. This decrease may effectively take the frontal lobes 'off line' and result in the sense of losing control. It is also well known that the dopaminergic system is involved in euphoric states and thus may be associated with transformative experiences of high emotional states. Future studies will be necessary to elaborate on the role of dopamine in spiritual transformation as well as the interactions between dopamine and other neurotransmitter systems.

Posterior Superior Parietal Lobe Deafferentation

The PSPL is heavily involved in the analysis and integration of higher-order visual, auditory, and somaesthetic information. It is also involved in a complex attentional network that includes the PFC and thalamus. Through the reception of auditory and visual input from the thalamus, the PSPL is able to help generate a three-dimensional image of the body in space, provide a sense of spatial coordinates in which the body is oriented, help distinguish between objects, and exert influences in regard to objects that may be directly grasped and manipulated. These functions of the PSPL might be critical for distinguishing between the self and the external world. It should be noted

that recent studies have suggested that the superior temporal lobe may play a more important role in body spatial representation, although this has not been confirmed by other reports. Regardless of what the actual relationship between the parietal and temporal lobes is, in terms of spatial representation, these areas are likely to play a key role in the altered sense of self and space associated with these unusual states of consciousness.

If, for example, there is deafferentation of the PSPL by the reticular nucleus's GABAergic effects, the person may begin to lose their usual ability to spatially define the self and help to orient the self. This is an experience frequently associated with meditative practices as well as with transformative types of experiences. Deafferentation of the PSPL has also been supported by several imaging studies demonstrating decreased activity in this region during intense meditation.

Hippocampal and Amygdalar Activation

In addition to the complex cortical-thalamic activity, religious experiences might also be associated with altered activity in the limbic system, especially since stimulation of limbic structures is associated with profound emotional responses. The hippocampus acts to modulate and moderate cortical arousal and responsiveness, via rich and extensive interconnections with the PFC, other neocortical areas, the amygdala, and the hypothalamus. Hippocampal stimulation has been shown to diminish cortical responsiveness and arousal; however, if cortical arousal is initially at a low level, then hippocampal stimulation tends to augment cortical activity. The ability of the hippocampus to stimulate or inhibit neuronal activity in other structures likely relies upon the glutamate and GABA systems respectively.

It has been previously suggested that the blocking of sensory information (i.e., deafferentation) of the PSPL might be associated with the loss of the sense of self and/or the orientation of that self to other objects in the world. Since this loss of orientation of the self is often described in religious experiences, the deafferentation of the PSPL might be an important mechanism underlying such experiences. If partial deafferentation of the right PSPL occurs during a transformative experience, the result may be stimulation of the right hippocampus, because of the inverse modulation of the hippocampus in relation to cortical activity. If, in addition, there is simultaneous direct stimulation of the right hippocampus via the thalamus (as part of the known attentional network) and mediated by glutamate, then a powerful recruitment of stimulation of the right hippocampus could occur. Right hippocampal activity may ultimately enhance the stimulatory function of the PFC on the thalamus via the nucleus accumbens, which gates the neural input from the PFC to the thalamus via the neuromodulatory effects of dopamine.

The hippocampus greatly influences the amygdala, such that they complement and interact in the generation of attention, emotion, and certain types of imagery. It seems that much of the prefrontal modulation of emotion is via the hippocampus and its connections with the amygdala. Because of this reciprocal interaction between the amygdala and hippocampus, the activation of the right hippocampus likely stimulates the right lateral amygdala as well. The results of the fMRI study by Lazar *et al.* support the notion of increased activity in the regions of the amygdala and hippocampus during practices such as meditation. However, whether such functional changes are associated with religious experiences remains to be seen.

Language Areas

An increase in the inferior parietal region has been reported in Franciscan nuns during their focus on a bible passage, which is an area commonly associated with language functioning. This type of activation is lacking in meditative practices that do not involve words. One example is Tibetan Buddhists' focus on a visual stimulus, which is unlike the nun's verbal focus. Interestingly, Pentecostal Christians speaking in tongues showed no increase in the inferior parietal region and a decrease in frontal lobe activity, despite the fact that they "speak" during their practice. This finding makes sense when considering that linguistic analyses of glossolalia find that it does not correspond to any linguistic structure. It would seem that the process of speaking in tongues is distinct from speaking a language. More research

of this practice and other verbal practices such as Sufism is needed.

Hypothalamic and Autonomic Nervous System Changes

The hypothalamus is extensively interconnected with the limbic system. Stimulation of the right lateral amygdala has been shown to result in stimulation of the ventromedial portion of the hypothalamus with subsequent stimulation of the peripheral parasympathetic system. Increased parasympathetic activity should be associated with the subjective sensation first of relaxation, and eventually, of a more profound quiescence. Activation of the parasympathetic system would also cause a reduction in heart rate and respiratory rate. All of these physiological responses have been observed during meditation. Religious and spiritual experiences can also have a number of strong emotional responses ranging from bliss to ecstasy. Blissful elements of such experiences may be mediated in part by activity in the parasympathetic system.

Typically, when breathing and heart rate slow down, the paragigantocellular nucleus of the medulla ceases to innervate the locus ceruleus (LC) of the pons. The LC produces and distributes norepinephrine (NE), a neuromodulator that increases the susceptibility of brain regions to sensory input by amplifying strong stimuli, while simultaneously gating out weaker activations and cellular 'noise' that fall below the activation threshold. Decreased stimulation of the LC results in a decrease in the level of NE. The breakdown products of catecholamines such as NE and epinephrine have generally been found to be reduced in the urine and plasma during meditation, which may simply reflect the systemic change in autonomic balance. However, it is not inconsistent with a cerebral decrease in NE levels as well. During a meditative practice, the reduced firing of the paragigantocellular nucleus probably cuts back its innervation of the LC, which densely and specifically supplies the PSPL and the lateral posterior nucleus with NE. Thus, a reduction in NE would decrease the impact of sensory input on the PSPL, contributing to its deafferentation. Thus, religious and spiritual experiences associated with an altered state of consciousness could be facilitated by such neuronal interactions.

The LC would also deliver less NE to the hypothalamic paraventricular nucleus. The paraventricular nucleus of the hypothalamus typically secretes corticotropin-releasing hormone (CRH) in response to innervation by NE from the LC. This CRH stimulates the anterior pituitary to release adrenocorticotropic hormone (ACTH). ACTH, which in turn stimulates the adrenal cortex to produce cortisol, one of the body's stress hormones. Decreasing NE from the LC during meditation would likely decrease the production of CRH by the paraventricular nucleus and ultimately decrease cortisol levels.

A drop in blood pressure associated with parasympathetic activity would be expected to relax the arterial baroreceptors leading the caudal ventral medulla to decrease its GABAergic inhibition of the supraoptic nucleus of the hypothalamus. This lack of inhibition can provoke the supraoptic nucleus to release the vasoconstrictor arginine vasopressin (AVP), thereby tightening the arteries and returning blood pressure to normal. AVP has also been shown to contribute to the general maintenance of positive affect, decrease self-perceived fatigue and arousal, and significantly improve the consolidation of new memories and learning. During meditative practice, plasma AVP has been shown to increase dramatically, although it is not known whether there will be a similar increase associated with more transformative experiences. AVP could help to enhance the memory of a particular experience, perhaps explaining the subjective phenomenon that meditative experiences are remembered and described in very vivid terms.

Prefrontal Cortex Effects on Other Neurochemical Systems

As PFC activity increases, it produces increasing levels of free synaptic glutamate in the brain. Increased glutamate can stimulate the hypothalamic arcuate nucleus to release b-endorphin (BE). BE is an opioid produced primarily by the arcuate nucleus of the medial hypothalamus and distributed to the brain's subcortical areas. BE is known to depress respiration, reduce fear, reduce pain, and produce sensations of joy and euphoria. That such effects have been described during meditation may implicate some degree of BE

release related to the increased PFC activity. Further, the joy and euphoric feelings associated with spiritual experiences might similarly implicate the endogenous opioid system. Meditation has been found to disrupt diurnal rhythms of BE and ACTH, while not affecting diurnal cortisol rhythms. Thus, the relationship between opioid receptors and various spiritual experiences is not clear, especially in light of one very limited study demonstrating that blocking the opiate receptors with naloxone did not affect the experience or EEG associated with meditation.

Glutamate activates N-methyl D-aspartate receptors (NMDAr), but excess glutamate can kill these neurons through excitotoxic processes. It is possible that if glutamate levels approach excitotoxic concentrations during intense experiences, the brain might limit its production of N-acetylated-alpha-linked-acidic dipeptidase, which converts the endogenous NMDAr antagonist N-acetylaspartylglutamate (NAAG) into glutamate. The resultant increase in NAAG would protect the cells from excitotoxic damage. There is an important side effect, however, since the NMDAr inhibitor, NAAG, is functionally analogous to the disassociative hallucinogens ketamine, phencyclidine, and nitrous oxide. These NMDAr antagonists produce a variety of states that may be characterized as either schizophrenomimetic or mystical, such as out-of-body and near-death experiences. Whether such substances are elaborated within the brain during the short interval associated with a religious and spiritual experiences is not yet clear.

Autonomic-Cortical Activity

In the early 1970s, Gellhorn and Kiely developed a model of the physiological processes involved in meditation based almost exclusively on ANS activity, which while somewhat limited, indicated the importance of the ANS during such experiences. These authors suggested that intense stimulation of either the sympathetic or parasympathetic system, if continued, could ultimately result in simultaneous discharge of both systems (what might be considered a 'breakthrough' of the other system). We have suggested that this breakthrough is associated with the most intense, and potentially most transformative types of religious and spiritual experiences. Several studies have demonstrated predominant parasympathetic activity during meditation, associated with decreased heart rate and blood pressure, decreased respiratory rate, and decreased oxygen metabolism. However, a recent study of two separate meditative techniques suggested a mutual activation of parasympathetic and sympathetic systems by demonstrating an increase in the variability of heart rate during meditation. The increased variation in heart rate was hypothesized to reflect activation of both arms of the ANS. Several other studies have also suggested that there may be some type of mutual excitation or rapid oscillation between the sympathetic and parasympathetic arms of the ANS during meditation practices. This notion also fits the characteristic description of meditative states in which there is a sense of overwhelming calmness as well as significant alertness. Also, the notion of mutual activation of both arms of the ANS is consistent with recent developments in the study of autonomic interactions. However, whether such complex interactions actually occur during meditative practices remains to be fully elucidated.

Serotonergic Activity

Activation of the ANS can result in intense stimulation of structures in the lateral hypothalamus and median forebrain bundle that are known to produce both ecstatic and blissful feelings when directly stimulated. Stimulation of the lateral hypothalamus can also result in changes in serotonergic activity. In fact, several studies have shown that after meditation, the breakdown products of serotonin in urine are significantly increased suggesting an overall elevation in serotonin during meditation. Serotonin is a neuromodulator that densely supplies the visual centers of the temporal lobe, where it strongly influences the flow of visual associations generated by this area. The cells of the dorsal raphe produce and distribute serotonin when innervated by the lateral hypothalamus (and also when activated by the PFC). Moderately increased levels of serotonin appear to correlate with the positive affect, while low serotonin often signifies depression. This relationship has clearly been demonstrated with regard to the effects of selective serotonin reuptake inhibitor medications, which are widely used for the treatment of depression.

Increased serotonin levels can affect several other neurochemical systems. An increase in serotonin has a modulatory effect on dopamine, suggesting a link between the serotonergic and dopaminergic system that may enhance feelings of euphoria, frequently described during religious and spiritual states. Serotonin, in conjunction with the increased glutamate, has been shown to stimulate the nucleus basalis to release acetylcholine, which has important modulatory influences throughout the cortex. Increased acetylcholine in the frontal lobes has been shown to augment the attentional system, and in the parietal lobes to enhance orienting without altering the sensory input. Increased serotonin combined with lateral hypothalamic innervation of the pineal gland may lead the latter to increase production of the neurohormone melatonin from the conversion of serotonin. Melatonin has been shown to depress the central nervous system (CNS) and reduce pain sensitivity. During meditation, blood plasma melatonin has been found to increase sharply, which may contribute to feelings of calmness and decreased awareness of pain.

EEG Studies of Spiritual Experiences

EEG can be used to study neuroelectrical changes that occur due to various spiritual practices such as meditation or prayer. For example, over sixty studies have been performed since at least 1957 on the effects of meditation, based on EEG output; practices studied include Kriya Yoga, Transcendental Meditation (TM), Zen Buddhism, Tibetan Buddhism, and Kundalini Yoga. For many years this method was the only tool available to researchers for studying the activity of the mind – the first studies using functional neuroimaging modalities such as PET, SPECT, and fMRI to study meditation were performed in the 1990s. Thus, there is a large amount of data on brain wave changes and meditation, and in this data there is some consideration of the fact that no two practices are necessarily the same. Different practices can induce varying EEG results.

Overall, when changes in alpha waves (8–12 Hz) have been described during meditation practices, the result is an increase in alpha wave power. In addition, this band is stronger at rest in meditators versus nonmeditators. Interestingly, the location of increased alpha power varies across meditation practices. Advanced Qigong meditators have been shown to have increased alpha power only over the frontal cortex, with decreased alpha power over the occipital cortex. TM, Zen meditation, various yoga practices, and even relaxation training have all been observed to affect alpha waves. Alpha wave blocking can occur in meditation as well. In resting states, alpha power is reduced after eyes that were closed are opened, in which case the occipital cortex shows a decrease in alpha waves. This signifies going from a state of relaxation to one of cognitive processing. Therefore, there is a connection between alpha waves and cortical processing as alpha decreases are typically seen after a stimulus is administered. Studies from the 1960s examining Indian yogis while meditating showed no alpha blocking in response to stimuli such as placing their hands in cold water. A more recent study of TM meditators showed no changes in alpha blocking when musical tones were presented. These results suggest that meditation training enables one to maintain levels of relaxation and mental 'emptiness' despite stimuli that are meant to be disruptive.

Increased theta activity (4–8 Hz) can be seen especially in advanced long-term meditators, but might be correlated with maintaining sustained attention, which is generated by the anterior cingulated and medial/dorsolateral prefrontal cortex, and may not necessarily be part of the spiritual experience. An increase in theta waves can be seen in TM, certain types of Yoga meditations, and Tibetan Buddhism, and is especially associated with advanced practitioners. When studying two different types of Qigong meditation it was found that those practicing the more concentrative type showed frontal midline theta activity, while the more passive, mindfulness based type did not illicit such brain activation in practitioners. It is further interesting to note the hypnosis is often associated with a similar increase in theta activity. Trances can be observed in religious rituals and in patients with psychiatric conditions. EEG recordings of a trance called Kerauhan, which is experienced by Indonesians participating in the Calonarang ritual, showed enhanced alpha and theta bands of spontaneous EEG activity. These trance results differ from results of individuals who undergo trances as a

result of schizophrenic or dissociative disorders. Schizophrenic trance states tend to show a decrease in alpha rhythms.

A Tibetan Buddhist practice called gTum-mo generates heat in the body and is characterized by increased beta activity; in addition, increases in finger and toe temperatures were seen at up to 8.3 C in this practice. More recent EEG studies of Tibetan Buddhist meditation have shown increases in gamma power, while a study of Zen meditators showed mostly decreased gamma coherence. Gamma activity may play a role in affective regulation, but more research is needed to elucidate this relationship more clearly.

Cortical synchronization is another aspect of cortical physiological function that EEG can measure. While Sahaja yoga meditators and TM practitioners have been found to show less hemispheric lateralization when compared to controls, subjects trained in mindfulness meditation have shown an increase in right-sided alpha power while not meditating. Subjects in the previously mentioned Tibetan gTum-mo condition showed greater asymmetries during their practice, with an increase in right hemispheric activity. Therefore, no generalizations can be made regarding hemispheric lateralization and meditative practices. In addition, lateralization has been studied outside of the realm of meditation, as left hemispheric dominance has been associated with happier states and traits, and left and right hemisphere interactions are involved with the approach-withdrawal process.

Other Relevant Topics in Researching Religion and Spiritual Experiences

The Role of Psychotropic Drugs in Spiritual Experiences

The effects of dugs on consciousness are discussed in another article of this encyclopedia. However, drugs have been used to induce mystical states, and for the purposes of ritual and religious experience throughout human history. Several studies have linked dimethyl tryptamine (DMT) to a variety of mystical states, including out-of-body experiences, distortion of time and space, and interaction with supernatural entities. Hyperstimulation of the pineal at this step, then, could also lead to DMT production that can be associated with a wide variety of mystical-type experiences associated with hallucinogens. It has been observed that under circumstances of heightened activation, pineal enzymes can also endogenously synthesize DMT. Another hallucinogen, D-lysergic diethylamide (LSD), is thought to work directly on the serotonergic system and use the connections that the system has, in order produce wide-ranging effects, including stimulation of the noradrenergic system. Tryptamine-based psychedelic drugs such as psilocybin and LSD bind to cortical serotonin (ST_2) receptors (especially in the temporal lobes), which can result in a hallucinogenic effect. The mechanism by which this appears to occur is that serotonin inhibits the lateral geniculate nucleus, greatly reducing the amount of visual information that can pass through. If combined with reticular nucleus inhibition of the lateral geniculate, serotonin may increase the fluidity of temporal visual associations in the absence of sensory input, possibly resulting in the internally generated imagery that has been described during certain meditative states and ritual experiences.

As mentioned, psilocybin has similar effects on the serotonergic system as LSD. In a study of 36 volunteers who participated in a double blind psilocybin study, 22 of them reported having a 'complete' mystical experience, while only four reported having a mystical experience with the control drug (methylphenidate). In addition, 67% of the volunteers rated the psilocybin experience as either the single most meaningful experience of their lives, or in their top five meaningful lifetime experiences. However, a recent PET study of the brain during psilocybin use was interpreted as a 'psilocybin model of psychosis.' Interestingly, many of the brain areas activated during psilocybin use are the same as those activated during spiritual experiences. Further research of psychotropic drugs is needed.

A Sample of Relevant Religious and Spiritual Practices

In order to fully understand the implications of the research described above, it is important to have some insight into the meditation practices studied.

One form of Kundalini Yoga meditation includes volitional practices such as one in which practitioners passively observe their breathing and silently repeat the phrases 'sat nam' during their inhalation and 'wahe guru' during exhalation. It is important to note that this practice focuses meditators both on a physical point (their breathing) and a mental point (the phrases). This type of meditation activates neural structures involved in attention (frontal and parietal cortex), which relate to the mental concentration and arousal/autonomic control (pregenual anterior cingulate, amygdala, midbrain, and hypothalamus), relating to physical concentration. A practice of Tibetan Buddhist meditators involves focusing on an image in their mind's eye, and concentrating on increasingly absorbing themselves in this image. Practitioners report clarity of thought and losing a sense of time and space, making this form of meditation, similar to the Kundalini Yoga practice mentioned above, one that focuses intently on mental processes, however, without physical focus. Increases in CBF in the inferior and orbital frontal cortices, dorsolateral prefrontal cortices, the sensorimotor and dorsomedial cortices, the midbrain, cingulate gyri, and the thalamus are often seen here. Note that brain areas important for autonomic arousal are not active here.

A Yoga Nidra guided meditation practice involves a relaxed meditative state in which the consciousness of the sensory world and the consciousness of action are subjectively dissociated, meaning the mind 'withdraws' from wishing to act. Meditators in this tradition tend to report a loss of conscious control of action, and vivid imagery. While Yoga Nidra meditators are instructed to be relaxed, their focus is not physiological in the sense of the focus on breath in Kundalini meditation. Yoga Nidra meditators tend to have increased dopaminergic tone in the ventral striatum during an altered consciousness state. This makes sense when examining circuits in the brain that run from the frontal to the subcortical regions of the brain, in which the ventral striatum is wired, that regulate behavior. Damage to one of these loops can cause anterior cingulate syndrome, which is characterized by an increased sense of apathy and a decrease in speech, motor behavior, and emotionality. What follows from the existence of an anterior cingulate syndrome is that altered activity in the ventral striatum during meditation could, in some form, mirror the effects of an altered frontal subcortical loop. This research suggests that functions known to be mediated by a certain area can be altered through thought processes established in meditation, as seen in a change in the local activity. As is evident, meditation practices vary widely, and have very different effects on brain activation.

Psychological Studies of Spiritual Experiences and Religion

Meditation, and more generally, spiritual practices are quickly spreading into many areas of psychological and health treatments. It appears that even short-term practice of meditation can significantly alter an individuals functioning, both physiologically and psychologically. One of the most agreed upon changes that occur is an increase in attention. This includes increased performance on tests that present distracting stimuli and the ability to sustain attention. However, all types of spiritual practices do not show the same results.

Research on the effects of spirituality and religion on health outcomes is abundant and not the focus of this article. However, meditation is being used to treat psychological effects of disease, and has gained a reputation as a treatment for anxiety and stress. Its effects on the CNS seem to be able to translate into decreasing individuals' perception of their stress levels. Programs that teach mindfulness-based stress reduction are being developed in hospitals and recommended for both the healthy and sick. Another mindfulness-based practice has been integrated into various forms of cognitive therapy to treat depression and anxiety and prevent relapse into these psychological conditions. Mantra recitation has been used in HIV patients to improve their overall quality of life, existential well-being and spiritual faith, even outside of affecting their physiological functioning. In addition to those suffering from illness, caregivers tend to turn to religion and spirituality for coping purposes, and religiosity seems to contribute to their own well-being.

There remains some debate over the idea of religion as a universal buffer against psychological stress, primarily because the health benefits, which are only slightly significant, may be due to the

individual, not to the religious activity itself. Religion has also been associated with fear of death and guilt, and when someone whose religious beliefs differ from those around him or her, religion can lead to stress and tension. Many have suggested a link between religion and happiness, however, because a clear definition of happiness does not exist, this relationship needs to be qualified more exactly. Overall, more research is needed on the effect of religion and psychological well-being.

Future Issues with Studying Spiritual Experiences

It is important to note that while neurophysiological mechanisms may be correlated with a spiritual experience they do not explain why the brain is capable of generating these experiences. Some would assert that this ability proves the existence of a higher power, but this is an opinion that should be considered with great care. There is no reasonable hypothesis to explain conscious awareness or spiritual experience arising out of an electrical input/output system, no matter what its complexity. It is unknown if animals with less complicated nervous systems have spiritual experiences or even have basic consciousness. This and other problems are explored in other articles of this encyclopedia.

Critics of research on religious experience and spiritual experience assert that the field is seeking to measure the immeasurable, and some religious groups argue that the research attempts to reduce religious experience to a function of neurons and neurotransmitters. This leads to the question of whether the deepest spiritual and mystical experiences are externally real or not, that is, solely generated and existing in the minds of the meditators and mystics, or actually part of some external reality/force in an objective sense. However, it might be argued that physiological analyses need not challenge the reality of religious experiences or of a possible reality causing them, any more than it challenges the reality of the physical world, since ultimately, all human experiences are interpreted in some manner by the brain. Another issue with the current research is that the subjects studied tend to be individuals whose practices are on the extreme end of the religious spectrum. Such a small percentage of people feel that they attain these intense spiritual states, that these studies do not address the neurobiology of a typical religious experience. The everyday believer who might go to a church, synagogue or mosque on their holy day, and consider that to be his/her religious activity, is more difficult to evaluate scientifically because of the subtle nature of these experiences. Thus, substantial advances in scientific methodology and an expansion of research to include the enormous variety of religious and spiritual experiences, not just the strongest ones, will be required to fully elucidate the nature of the psychology and neurology of such experiences.

See also: Altered and Exceptional States of Consciousness; The Neurochemistry of Consciousness.

Suggested Readings

Austin J (1999) Zen and the Brain. Cambridge, MA: MIT Press.

Beauregard M and O'Leary D (2007) The Spiritual Brain. New York: HarperOne.

Cahn BR and Polich J (2006) Meditation states and traits: EEG, ERP, and neuroimaging studies. Psychological Bulletin 132(2): 180–211.

Cardena E, Lynn SJ, and Krippner S (2000) Varieties of Anomalous Experience. Washington, DC: APA.

McNamara P (2006) Where God and Science Meet: How Brain and Evolutionary Studies Alter Our Understanding of Religion. Westport, CT: Praeger Publishers.

Newberg AB and Iversen J (2003) The neural basis of the complex mental task of meditation: Neurotransmitter and neurochemical considerations. Medical Hypothesis 61(2): 282–291.

Newberg AB and Waldman MR (2007) Born to Believe. New York: Free Press.

Biographical Sketch

Hannah Roggenkamp, BA is a research assistant in the Department of Radiology, University of Pennsylvania.

Mark Waldman, BA is a therapist and an associate fellow at the Center for Spirituality and the Mind, University of Pennsylvania.

Andrew Newberg, MD is associate professor in the Department of Radiology and Psychiatry, and director of the Center for Spirituality and the Mind, University of Pennsylvania.

Self: Body Awareness and Self-Awareness

J L Bermúdez, Washington University in St. Louis, St. Louis, MO, USA

Glossary

Body awareness, high level – Representations of the body that feed directly into and are informed by central cognitive/affective processes.

Body awareness, low level – Awareness of the structure, layout, and moment-by-moment distribution of body parts derived from mechanisms of somatic proprioception and somatosensation.

Immunity to error through misidentification – Information sources (such as introspection or proprioception) are immune to error through misidentification when the information that they provide can only be about the subject receiving it – when there is no possibility for misidentifying the source of the information.

Self-awareness – Self-awareness is a special type of awareness of the self. It is derived from information sources that are immune to error through misidentification and that have immediate and direct implications for action.

Somatic proprioception – Awareness of limb position and bodily configuration derived from the vestibular (balance) system, the sense of touch, and receptors in the joints, muscles, and tendons.

Introduction

Human beings, and indeed most animals, are aware of their bodies in distinctive ways. We have ways of finding out about our own bodies that are very different from the ways we have of finding out about other physical objects in the world. There are distinctive information channels that allow us directly to monitor both the body's internal states and how it is oriented in space. Some of these information channels are conscious and others unconscious. They all contribute, however, to a distinctive type of experience, viz, the experience of oneself as an embodied agent. It is to this distinctive type of experience that theorists are referring when they talk about body awareness. Body awareness is of fundamental importance for the survival of the organism. It allows us to monitor homeostatic states, such as hunger and thirst, as well as to detect and anticipate bodily damage. It is of fundamental importance in generating and controlling action. Moreover, body awareness has an important role to play in self-awareness. For the purposes of this article, I take self-awareness to be the same as self-consciousness and am understanding it in a specialized sense (explained in the 'Glossary' and in the 'Self-awareness and the elusiveness thesis' and 'Body awareness as self-awareness' sections).

This article begins by outlining the different types of body awareness. High-level forms of body awareness are discussed in 'High-level body awareness' and lower-level types of body awareness in 'Low-level body awareness'. The lower-level types will be primarily emphasized in the article, and the section 'The mechanisms of lower-level body awareness' outlines some of what is known about the physiological mechanisms governing lower-level body awareness. In 'Body awareness in visual perception' section, we consider a type of body awareness derived from visual perception. The sections entitled 'Self-awareness and the elusiveness thesis' and 'Body awareness as self-awareness' explain why body awareness counts as a form of self-awareness. Finally, in 'The space of body awareness' we explore the spatiality of body awareness and the differences between how we experience bodily space and how we experience extra-bodily space.

High-Level Body Awareness

To understand the complex phenomenon of bodily awareness, we need to begin with a very general distinction between high-level and low-level representations of the body. High-level representations of the body feed directly into, and are informed by, central cognitive/affective processes. In contrast, low-level representations of the body (which we will consider in the next section) feed directly into action. The general distinction is sometimes made between the body-image (high-level) and body-schema (low-level). In this section and the next, I extend this basic distinction into a fuller taxonomy of the different types of body awareness.

Within the general category of high-level representations of the body, we can distinguish at least three different types of representation or bodies of information:

Conceptual awareness of the body: The set of beliefs we all have about the structure and nature of our body: how the body fits together, the functions of particular body-parts, their approximate locations, and the sort of things that can go wrong with them.

Semantic awareness of the names of body-parts: Knowledge that interfaces with nonsemantic ways of identifying events in the body to allow us to report on what is going on in our bodies.

Affective awareness of the body: Representations of the body associated with emotional responses to the body.

Conceptual representations of the body are not particularly interesting from the perspective of self awareness. There seems little reason to think that such conceptual representations will be any different in kind from the set of commonsense beliefs that we all have about the physical and social world. They may be based upon particular forms of body awareness, but it does not seem appropriate to describe them as actually being a form of body awareness.

The remaining two types of higher-level representation are more interesting and more clearly types of body awareness. One way of seeing this is to reflect on how these higher-level representations can become distorted. There are identifiable pathologies specific to both semantic and affective representations of the body. The pathologies associated with affective representations of the body are familiar. Bulimia and anorexia are good examples – forms of emotional response based on distorted representations of the body. These are plainly types of body awareness, as is the (relatively) undistorted affective response to the body that normal subjects have.

There are also identifiable pathologies associated with semantic representations of the body. Patients with autotopagnosia have difficulty in naming body-parts or pointing to body-parts identified by name or by the application of some stimulus, either on their own bodies or on a schematic diagram of the body. The problems here are not purely semantic. Semantic representation of the body is not simply a matter of knowing the names of body-parts. Although superficially similar deficits can be found in some aphasic patients, autotopagnosic patients do not have a localized word category deficit. They lack a particular way of representing bodily locations, as we see from the fact that the problem carries across to pointing to body-parts identified by the application of a stimulus. Again, what is at stake here is the distortion of what in normal subjects is a mode of body awareness – an awareness of the body that manifests itself in the ability to identify specific body-parts.

Low-Level Body Awareness

Turning to lower-level representations of the body, here too we find a range of phenomena and associated information channels that need to be distinguished. The first is information about the structure and limits of the body. This type of body-relative information has a number of distinctive pathologies, which have been studied by Ronald Melzack, Vilayanur Ramachandran, and others. The best known is the phenomenon of phantom limb found in many patients with amputated limbs, as well as some with amelia, the congenital absence of limbs. This first category of body-relative information performs two tasks. First, it is responsible for the felt location of sensations. Sensations are referred to specific body-parts in virtue of a body of information about the structure of the body.

Second, the same body of information informs the motor system about the body-parts that are available to be employed in action. Information about the structure and limits of the body is relatively stable, although experiments by James Lackner and others have shown that it can be distorted (as in the aptly named Pinocchio illusion, in which subjects who are touching their nose and have their biceps tendon stimulated with a vibrator feel their noses growing).

This structural type of body-relative information should be distinguished from semantic representations of the body (as discussed in the section 'High-level body awareness'). In deafferented patients they are dissociated in both directions. Deafferented patients have lost peripheral sensations in certain parts of the body. The patient GL studied by Jacques Paillard suffers from almost complete deafferentation from the mouth down, although she retains some sensitivity to thermal stimuli. If a thermal stimulus is delivered to a point on her arm that she is prevented from seeing then, although she is unable to point to the location of the stimulus on her body, she is able to identify the location verbally and on a schematic body diagram. In my terms, she possesses semantic information without body-relative information. The dissociation also holds in the opposite direction. Another of Paillard's patient had a parietal lesion that resulted in central deafferentation of the forearm. Although she could not verbally identify and report on a tactile stimulus delivered to her deafferented hand in a blindfolded condition, she was able to point to the location of the stimulus. (Unlike the very similar and well-documented phenomenon of blindsight, there was no need to force subjects to make a choice.) The dissociation here may well be between an action-based representation of the body and an objective representation of the body. An action-based representation of the body represents body location in a way that feeds directly into action, whether that action is body-directed or world-directed. It is this that is lost in GL, but preserved in the patient with the deafferented forearm. In what I am calling an objective representation of the body, on the other hand, the body does not feature purely as a potentiality for action, but rather as a physical object whose parts stand in certain determinate relations to each other.)

There is a second type of lower-level representation of the body. This is a moment-to-moment representation of the spatial position of the various parts of the body. This moment-to-moment representation of bodily position is essential for the initiation and control of action, and needs to be constantly updated by feedback from moving limbs. This representation has been called the short-term body-image by the philosopher Brian O'Shaughnessy, but the name is misleading, suggesting that there is a single way in which the disposition of body-parts is represented, whereas in fact the spatial location of any given body-part can be coded in three different and independent ways.

The first type of coding is relative to objects in the distal environment. Consider a simple action, such as reaching one's hand out for an object. The success of this action depends upon an accurate computation of the trajectory from the initial position of the hand to the position of the relevant object. This requires the position of the hand and the position of the object to be computed relative to the same frame of reference. I call this object-relative spatial coding. It is most likely that object-relative spatial coding takes place on an egocentric frame of reference – that is to say, a frame of reference whose origin is some body-part. The reason for calling this type of coding object-relative is that it deals primarily with the spatial relations between body-parts and objects in the distal environment.

But many actions are directed toward the body rather than to objects independent of the body. Some of these actions are voluntary, as when I clasp my head in my hands in horror. Some are involuntary, as when I scratch an itch. Many more are somewhere between the two, as when I cross my legs or rub my eyes. Clearly, the possibility of any of these sorts of action rests upon information about the location of the body-parts in question relative to each other. We can call this sort of information body-internal spatial coding. It is information about the moment-by-moment position of body-parts relative to each other.

Body-internal spatial coding is required, not just for body-directed action, but also for many types of action directed towards objects in the distal environment. Psychological studies of action often concentrate on very simple actions, such as grasping objects with one hand. But the vast majority of actions

require the coordination of several body-parts. When I play volleyball, for example, I need to know not just where each of my hands is relative to the ball as it comes over the net, but also where each of my hands is relative to the other hand. Both body-internal and object-relative spatial coding is required.

A third type of information about the moment-to-moment disposition of the body is just as important for the initiation and control of action as the first two. This is information about the orientation of the body as a whole in objective space, primarily involving information about the orientation of the body with respect to supporting surfaces and to the gravitational field. This information comes from the calibration of information from a number of sources. The three principal sources of orientational information are vision, the vestibular system in the inner ear, and the proprioceptive/kineasthetic system (at least two of which must be properly functioning for orientational information to be accurate). I call this orientational coding.

The Mechanisms of Lower-Level Body Awareness

The various types of lower-level body awareness have intricate and highly developed physical underpinnings that are relatively well understood.

Physiologists and neurophysiologists have devoted considerable attention to understanding the mechanisms of proprioception (awareness of limb position and bodily configuration) and somatosensation (bodily sensation). We have a good understanding of how bodily sensations originate in specialized receptors distributed across the surface of the skin and within the deep tissues. Some of these receptors are sensitive to skin and body temperature. Others are pain detectors (nociceptors). There are receptors specialized for mechanic stimuli of various kinds, such as pressure and vibration. Information about muscle stretch comes from muscle spindles. Other receptors monitor stresses and forces at the joints and in the tendons. Information from all of these receptors and nerve endings is carried by the spinal cord to the brain. Once again, the mechanisms here are relatively well understood. It is thought, for example, that there are three different pathways in the spinal cord. One pathway carries information stemming from discriminative touch (which is a label for a complex set of tactile ways of finding out about the shape and texture of physical objects). Another carries information about pain and temperature. The third carries proprioceptive information. Each of these pathways ends up at a different brain area. The discriminative touch pathway travels to the cerebral cortex while the proprioceptive pathway terminates in the cerebellum. The properties of these brain areas have been well studied. We know, for example, that tactile information is processed in the somatosensory cortex, which is located in the parietal lobe. The somatosensory cortex is somatotopically organized, with specific regions representing specific parts of the body. The cortical space assigned to information from each bodily region is a function of the fineness of tactile discrimination within that region (which is itself of course a function of the number of receptors there). Neuropsychologists, neuroimagers, and computational neuroscientists have made considerable progress in understanding how somatosensory and proprioceptive information is processed in the brain and how that processing can be disturbed by brain injury.

Body Awareness in Visual Perception

We can of course perceive the body visually in the same way as we perceive any other object. Looking at ourselves in the mirror would be an example. This does not count as body awareness in any interesting sense. But there are several ways in which vision provides us with a distinctive form of body awareness. These operate through

1. self-specifying structural invariants in the field of vision,
2. visual kinesthesis, and
3. the perception of affordances.

Visual body awareness is derived from various types of self-specifying information available in visual perception. These have been studied most systematically by the perceptual psychologist J. J. Gibson. Unfortunately, Gibson's work in this area is not as widely accepted as it should be, no doubt due to his polemical and controversial criticisms of

standard information-processing approaches to vision. Gibson argued that vision does not involve the type of information-processing standardly studied by cognitive scientists and cognitive neuroscientists. We should understand vision in terms of direct sensitivity to information that is present in the flow of light around the perceiver. Gibson's theory of ecological optics is based on the idea that our perceptual systems resonate to this information. Fortunately, it is possible to learn from Gibson's analysis of the self-specifying information in vision without taking literally his ideas about the direct pickup of information.

We begin with the self-specifying structural invariants. The field of vision contains bodily objects that hide, or occlude, the environment. The nose is an obvious example. It is distinctively present in just about every visual experience. The cheekbones, and perhaps the eyebrows, occupy a slightly less-dominant position in the field of vision. And so too, to a still lesser extent, do the bodily extremities, hands, arms, feet, and legs. They protrude into the field of vision from below in a way that occludes the environment, and yet which differs from the way in which one nonbodily physical object in the field of vision might occlude another. They are, as Gibson points out, quite peculiar objects. All objects, bodily and nonbodily, can present a range of solid angles in the field of vision (where by a solid angle is meant an angle with its apex at the eye and its base at some perceived object), and the size of those angles will of course vary according to the distance of the object from the point of observation. The further away the object is, the smaller the angle will be. This gives rise to a clear, and phenomenologically very salient, difference between bodily and nonbodily physical objects. The solid angles subtended by occluding body-parts cannot be reduced below a certain minimum. Perceived body-parts are, according to Gibson, 'subjective objects' in the content of visual perception.

The mass of constantly changing visual information generated by the subject's motion poses an immense challenge to the perceptual systems. How can the visual experiences generated by motion be decoded so that subjects perceive that they are moving through the world? Gibson's notion of visual kinesthesis is his answer to this traditional problem. Whereas many theorists have assumed that motion perception can only be explained by hypothesizing mechanisms that parse cues in the neutral sensations into information about movement and information about static objects, the crucial idea behind visual kinesthesis is that the patterns of flow in the optic array and the relations between the variant and invariant features make available information about the movement of the perceiver, as well as about the environment. As an example of such a visually kinesthetic invariant, consider that the optical flow in any field of vision starts from a center that is itself stationary. This stationary center specifies the point that is being approached, when the perceiver is moving. The aiming point of locomotion is at the vanishing point of optical flow. These kinesthetic invariants count as forms of bodily awareness because they directly provide information about the body's trajectory through space.

The theory of ecological optics identifies a third form of self-specifying information in the field of vision. This is due to the direct perception of a class of higher-order invariants that Gibson terms affordances. It is in the theory of affordances that we find the most sustained development of the ecological view that the fundamentals of perceptual experience are dictated by the organism's need to navigate and act in its environment. The uncontroversial premise from which the theory of affordances starts is that objects and surfaces in the environment have properties relevant to the abilities of particular animals, in virtue of which they allow different animals to act and react in different ways. According to Gibson, information specifying affordances is available in the structure of light to be picked up by the creature as it moves around the world. The possibilities that the environment affords are not learnt through experience, and nor are they inferred. They are directly perceived as higher-order invariants. The perception of affordances is a form of body awareness, because it contains information about environmental possibilities for action and reaction.

Self-Awareness and the Elusiveness Thesis

The preceding sections have been devoted to exploring the nature and mechanisms of body

awareness. We turn now to the idea that body awareness is a form of self-awareness. In this section, I give some philosophical background that will help in thinking about body awareness and self-awareness.

The first step is to distinguish two different types of awareness – direct awareness and propositional awareness. One can be aware of something (as when I catch sight of someone walking up the garden path) or one can be aware that a particular state of affairs is the case (as when the sound of the doorbell alerts me to the fact that a visitor is at the door). In direct awareness, the object of awareness is a particular thing. In propositional awareness, the direct object of awareness is a proposition or state of affairs (a complex of particular things, properties, and/or relations).

Direct awareness is not sensitive to how one thinks about the object of which one is directly aware. All that is required for me to be directly aware of something is that I be able to discriminate it. I do not need to know what it is, or to conceptualize it in any way. Propositional awareness, however, is not like this. I can be aware that a state of affairs holds when it is conceptualized in one way, but be unaware that it holds under a different conceptualization. I might be propositionally aware that Bob Dylan is balding, in virtue of seeing that the person on the stage in front of me is losing his hair and know that that person is Bob Dylan, without being propositionally aware that Robert Zimmerman is balding, since I have no idea that Bob Dylan is Robert Zimmerman. In contrast, if I am directly aware of Bob Dylan, then I am directly aware of Robert Zimmerman, even if I have no idea who either of them is.

Applying this distinction gives us two ways of thinking about self-awareness. Self-awareness can be understood either in terms of direct awareness of the self or in terms of propositional awareness that the self has such-and-such a property, or stands in such-and-such relations. Although it might seem obvious that we can be directly aware of ourselves, some philosophers have maintained that there can be no such thing as direct self-awareness at the level of direct awareness. The most famous example of this is David Hume's strong claim in his *Treatise of Human Nature* that the self cannot be directly perceived.

> For my part, when I enter most intimately into what I call myself, I always stumble on some perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I never catch myself at any time without a perception, and never can observe anything but the perception.

Hume's claim, in effect, is that we cannot be directly aware of the self in introspection (taking "introspection" here to pick out the means by which we enjoy first-person access to our own mental states). This claim, which seems itself to be derived from introspection, is generally known as Hume's elusiveness thesis.

At the time at which Hume was writing, the consensus view among philosophers and scientists was that the self is a purely psychological entity. This explains the weight attached to the elusiveness thesis. If the self is a purely psychological entity, then introspection is the only possible source for direct awareness of the self, and the impossibility of introspective awareness of the self would cast significant doubt on whether any such purely psychological entity exists (particularly for empiricists such as Hume, since empiricists typically hold that all knowledge must come either through the senses or from introspection). Since there are few if any surviving dualists, Hume's elusiveness thesis does not now have the same force that it once did. In fact, one might reasonably think that the forms of body awareness we have been discussing in this article all count as forms of direct awareness of the embodied self – and hence as counterexamples to the elusiveness thesis. And then, one might go on to think, body awareness is a form of self-awareness (self-consciousness).

Nonetheless, things are not quite as straightforward as this. Not all forms of direct awareness of the self count as forms of self-awareness. If I catch sight of myself in a shop window but do not realize that the person in the window is me, then I am certainly directly aware of myself, but I am not self-aware in any interesting sense. So we need to ask whether body awareness is like the shop window case, or whether it counts as a genuine form of self-awareness. The first step in answering this is to get a clearer idea of what counts as genuine self-awareness.

Most philosophers would think that there are at least two minimal conditions on genuine

self-awareness over and above its involving a direct awareness of the self. The first emerges when we think about an obvious feature of the shop window scenario. One of the things that makes the shop window case so obviously not an example of genuine self-awareness is the fact that I fail to identify myself. I correctly identify certain bodily properties, but fail to recognize that they are properties of my own body and misattribute them to somebody else. We do not want to collapse direct awareness into propositional awareness, and so it cannot be a requirement upon direct awareness of the self that it involves any sort of judgment (such as the judgment: "I am that person"). Nonetheless, we can reasonably demand that for direct awareness of the self to count as genuine self-awareness it should be derived from information channels that do not yield information about anybody's bodily properties except one's own (just as introspection does not yield information about anybody's psychological properties except one's own). In philosophy this is often put in rather cumbersome terms. In what has become the standard terminology, what I am calling genuine self-awareness is based upon information sources that have the property of being immune to error through misidentification relative to the first-person pronoun.

A second requirement upon genuine self-awareness is that it feeds directly into action. Again, we can appreciate this by noting how it is absent from the shop window case. Suppose I see that the person in the window is about to walk into a lamppost. In the scenario I am envisaging I may laugh, or I may be concerned. But, since I do not realize that I am the person who is about to walk into a lamppost, I will not take evasive action. Again, we do not want to make it a requirement upon genuine self-awareness that it involves any sort of judgment (such as: "I am about to walk into the lamppost"). But it seems reasonable to demand that, for a form of direct awareness of the self to count as genuine self-awareness, it should be based on forms of information that can feed directly into action

Body Awareness as Self-Awareness

It is plain that body awareness provides direct awareness of the embodied self. But does it provide a form of self-awareness? At the end of the previous section, we identified two basic requirements for a form of awareness of the self to count as genuine self-awareness. In this section, we begin by seeing how these two basic requirements are satisfied by body awareness. We then go on to explore a further sense in which body awareness yields awareness of the self.

The first requirement for direct awareness of the self to count as genuine self-awareness is that it should be based upon sources of information that do not yield information about anybody's bodily properties except my own (just as introspection does not yield information about anybody's psychological properties except my own). Here we need to focus primarily on two of the types of body awareness distinguished earlier – one operating through the mechanisms of somatic proprioception and somatosensation, and the other operating through self-specifying information in vision. Both of these satisfy the first requirement. Somatic proprioception is the most straightforward. It seems to follow from the simple fact that I somatically proprioceive particular bodily properties that those bodily properties are my own. We are quite simply "wired up" in such a way that proprioceptively derived information has the property of being immune to error through misidentification. Of course, this is not a logical truth. And it may not even be a nomological truth (that is, a truth dictated by the laws of nature). It is certainly conceivable that we could be hooked up to other people's bodies in such a way that we receive proprioceptive information from their bodies as well as (or even instead of) from our own. But since, as things currently stand, that possibility is not a 'relevant alternative,' it does not need to be taken into account. I am always justified in taking proprioceptively derived information to be information about my own body.

The same holds for self-specifying information derived through vision. The forms of self-specifying information identified earlier collectively create, for each of us, a distinctive and unique visual perspective on the world – and of course on ourselves, as physical objects navigating the world. This direct awareness of ourselves obtained through vision satisfies the first requirement for genuine self-awareness because one's perceptually derived information about where

one is relative to objects in the perceived environment could not possibly be about anyone but oneself. Of course, as the mirror and shop window cases show, not all visually derived awareness of the self counts as self-awareness. We need to restrict ourselves to the types of information identified earlier, such as that gained through visual kinesthesis and self-specifying structural invariants.

Moving on to the second requirement for genuine self-awareness, it is clear that direct awareness of the self through somatic proprioception and somatosensation does have immediate and direct implications for action. This holds true both of self-directed action (such as scratching itches) and of action whose target is not the body. One index of this is the role of the various somatic information systems in providing feedback about limb position and balance, as was brought out in our earlier discussion of the different types of low-level body awareness. The action-guiding role of self-specifying information in vision is equally clear. Gibson's analysis of vision focused primarily on how action is guided by information available in the optic flow – information both about the perceiver's movement relative to landmarks in the environment and about the possibilities for action and reaction that the environment affords. The important point is that the implications for action are direct – precisely because this is information that cannot be but about the perceiver. This is the crucial contrast with information derived from mirrors, shop windows, and so forth. This information has implications for action only after one has identified the person seen as oneself.

The direct awareness of the self obtained from the two types of body awareness that we have discussed do seem, therefore, to count as genuine forms of self-awareness, according to the two criteria we have identified. But there is, in addition, a further sense in which body awareness counts as a form of self-awareness. A crucial element in self-awareness is what developmental psychologists call self-world dualism. All subjects properly described as self-aware must be able to register the distinction between themselves and the world. This is a distinction that can be registered in a variety of different ways, and at different degrees of sophistication. Body awareness provides a way, perhaps the most primitive way, of registering the distinction between self and nonself. This is a weaker distinction than the distinction between self and world, of course, but it is certainly a necessary component of it.

There are two key elements to the distinction between self and nonself yielded by body awareness. The first element is an awareness of the limits of the body. This comes from a number of sources. The sense of touch has an important part to play, as the felt boundaries of the body define the limits between self and nonself. This offers a way of grasping the body as a spatially extended and bounded object. But self-specifying information in vision also has a role to play. As we saw earlier, body-parts appear in vision in a special way (as what Gibson evocatively calls 'subjective objects'), since their size can vary in only a limited way.

A second element in understanding the distinction between self and nonself is being able to distinguish between what is, and what is not, responsive to the will. Here somatic proprioception is particularly important. It is the feedback gained through kinesthesia, joint position sense and the vestibular system, which explains how one is aware that the body is responding to motor commands. It is true that much of the body is not at all responsive to the will – the internal organs are obvious examples. But the scope of the will does encompass all bodily surfaces and extremities (whether directly or indirectly). Although not every portion of the bodily surface can be moved at will, it is nonetheless the case that every portion of the bodily surface can be experienced as moving in response to an act of the will. Take an arbitrary area on the top of the head, for example. Although I cannot move that area at will (in isolation), I can nonetheless experience it as moving when I move my head as a whole. The limits of the will mark the distinction between the self and the nonself just as much as does the skin, although in a different way.

The Space of Body Awareness

Almost all existing discussions of the spatiality of body awareness have presupposed that exteroceptive

perception, body awareness, and the intentions controlling basic bodily actions must all code spatial information on comparable frames of reference (where a frame of reference allows locations to be identified relative to axes centered on an object). This is a natural assumption, given that action clearly requires integrating motor intentions and commands with perceptual information and several different types of lower-level information about the body (particularly what we earlier termed the short-term body-image). Since the spatial locations of perceived objects and objects featuring in the contents of intentions are given relative to axes whose origin lies in the body – in an egocentric frame of reference – it is natural to suggest that the axes determining particular proprioceptive frames of reference are centered on particular body-parts, just as are the axes determining the frames of reference for perceptual content and basic intentions. The picture that emerges, therefore, is of a number of different representations of space, within each of which we find representations both of bodily and of nonbodily location. So, for example, we might imagine reaching behavior to be controlled by an egocentric frame of reference centered at some location on the hand – a frame of reference relative to which both bodily location (such as the mosquito bite on my arm) and nonbodily location (such as the cup on the table) can be identified.

Despite its appealing economy, however, this account is ultimately unacceptable, because of a fundamental disanalogy between the bodily space of lower-level body awareness and the egocentric space of perception and action. In the case of vision or exteroceptive touch, there is a perceptual field bounded in a way that determines a particular point as its origin. Since the visual field is essentially the solid angle of light picked up by the visual system, the origin of the visual field is the apex of that solid angle. Similarly, the origin of the frame of reference for exploratory touch could be a point in the center of the palm of the relevant hand. But our awareness of our own bodies is not like this at all. It is not clear what possible reason there could be for offering one part of the body as the origin of the frame of reference that fixes locations in bodily space.

There are certain spatial notions that are not applicable to body awareness. For any two objects that are visually perceived, it makes obvious sense to ask both of the following questions:

(a) Which of these two objects is further away?
(b) Do these objects lie in the same direction?

The possibility of asking and answering these questions is closely bound up with the fact that visual perception has an origin-based frame of reference. Question (a) basically asks whether a line between the origin and one object would be longer or shorter than a corresponding line between the origin and the other object. Question (b) is just the question whether, if a line were drawn from the origin to the object that is furthest away, it would pass through the nearer object.

Neither question makes sense with respect to body awareness. One cannot ask whether this proprioceptively detected hand movement is farther away than this itch, nor whether this pain is in the same direction as that pain. What I am really asking when I ask which of two objects is further away is which of the two objects is further away from me, and a similar tacit self-reference is included when I ask whether two objects are in the same direction. But through somatic proprioception one learns about events taking place within the confines of the body, and there is no privileged part of the body that counts as me for the purpose of discussing the spatial relations they bear to each other.

To get a firmer grip on the distinctiveness of the frame of reference of bodily awareness, one need only contrast the bodily experience of normal subjects with that of completely deafferented subjects, such as Jonathan Cole's patient IW. The moment-to-moment information about their bodies that deafferented patients possess is almost exclusively derived from vision. Their awareness of their own body is continuous with their experience of the extra-bodily world. They are aware of their bodies only from the same third-person person perspective that they have on nonbodily physical objects. The frame of reference for their bodily awareness does indeed have an origin – the eyes – and for this reason both of the two questions mentioned make perfect sense. But this is not at all how we experience our bodies from a first-person perspective.

The conclusion to draw from this is that the spatial content of bodily awareness cannot be specified within a Cartesian frame of reference

that takes the form of axes centerd on an origin. But then how is it to be specified?

We can start from the basic thought that an account of the spatiality of bodily awareness must provide criteria for sameness of place. In the case of somatic proprioception, this means criteria for sameness of bodily location. But there are several different forms of criteria for sameness of bodily location. Consider the following two situations:

(i) I have a pain at a point in my right ankle when I am standing up and my right foot is resting on the ground in front of me.
(ii) I have a pain at the same point in my ankle when I am sitting down and my right ankle is resting on my left knee.

According to one set of criteria the pain is in the same bodily location in (i) and (ii) – that is to say, it is at a given point in my right ankle. According to another set of criteria, however, the pain is in different bodily locations in (i) and (ii), because my ankle has moved relative to other body-parts. Let me term these A-location and B-location, respectively. Note, moreover, that B-location is independent of the actual location of the pain in objective space. The B-location of the pain in (ii) would be the same if I happened to be sitting in the same posture five feet to the left.

Both A-location and B-location need to be specified relative to a frame of reference. In thinking about this we need to bear in mind that the human body has both moveable and (relatively) immoveable body-parts. On a large scale the human body can be viewed as an immoveable torso to which are appended moveable limbs – the head, arms, and legs. Within the moveable limbs there are small-scale body-parts that can be directly moved in response to the will (such as the fingers, the toes, and the lower jaw) and others that cannot (such as the base of the skull). A joint is a body-part that affords the possibility of moving a further body-part, such as the neck, the elbow, or the ankle. In the human body, the relatively immoveable torso is linked by joints to five moveable limbs (the head, two legs, and two arms), each of which is further segmented by means of further joints. These joints provide the fixed points in terms of which the particular A-location and B-location of individual body-parts at a time can be given.

A particular bodily A-location is given relative to the joints that bound the body-part within which it is located. A particular point in the forearm is specified relative to the elbow and the wrist. It will be the point that lies on the surface of the skin at such-and-such a distance and direction from the wrist and such-and-such a distance and direction from the elbow. This mode of determining A-location secures the defining feature of A-location, which is that a given point within a given body-part will have the same A-location irrespective of how the body as a whole moves, or of how the relevant body-part moves relative to other body-parts. The A-location of a given point within a given body-part will remain constant in both those movements, because neither of those movements will bring about any changes in its distance and direction from the relevant joints.

The general model for identifying B-locations is as follows. A particular constant A-location is determined relative to the joints that bound the body-part within which it falls. A-location will either fall within the (relatively) immoveable torso or it will fall within a moveable limb. If it falls within the (relatively) immoveable torso, then its B-location will also be fixed relative to the joints that bound the torso (neck, shoulders, and leg sockets) – that is to say, A-location and B-location will coincide. If, however, that A-location falls within a moveable limb, then its B-location will be fixed recursively relative to the joints that lie between it and the immoveable torso. The B-location will be specified in terms of the angles of the joints that lie between it and the immovable torso. Some of these joint angles will be rotational (as with the elbow joint, for example). Others will be translational (as with the middle finger joint).

This way of specifying A-location and B-location seems to do justice to how we experience our bodies. In particular,

> We do not experience peripheral body-parts in isolation, but rather as attached to other body-parts. Part of what it is to experience my hand as being located at a certain place is to experience that disposition of arm-segments in virtue of which it is at that place.
> It is part of the phenomenology of bodily awareness that sensations are always experienced

within the limits of the body. This is exactly what one would expect given the coding in terms of A-location and B-location. There are no points in (nonpathological) body-space that do not fall within the body.

Although B-location is specified recursively in terms of the series of joint angles between a given A-location and the immovable torso, the torso does not function as the origin of a Cartesian frame of reference.

These aspects of the phenomenology of body awareness (of how we experience our bodies) are important factors in explaining why body awareness is a genuine form of self-awareness. The distinctiveness of the spatiality of body awareness is an important part of what underwrites the agent's sense of the distinction between self and nonself.

Conclusion

Philosophers dealing with bodily awareness have to balance two considerations that pull in opposite directions. There is, first, the obvious fact that the body is a physical object in the world – which generates the plausible thought that awareness of the body must be somehow awareness of the body as an object in the world. Second, and equally obvious, is the fact that the body is (at least from a first-person perspective) quite unlike any other physical object – generating the equally plausible thought that awareness of the body is somehow fundamentally different from awareness of the body as an object in the world. The points about body awareness that have been discussed in this article offer a way of reconciling the apparent conflict. The types of body awareness that we have discussed jointly provide an awareness of the self as a spatially extended and bounded physical object that is distinctive in being responsive to the will.

See also: The MindBody Problem; Philosophical Accounts of Self-Awareness and Introspection; Self: Personal Identity; Self: The Unity of Self, Self-Consistency.

Suggested Readings*

Bermudez JL, Marcel AJ, and Eilan N (eds.) (1995) The Body and the Self. Cambridge, MA: MIT Press.

Bermudez JL (1998) The Paradox of Self-Consciousness. Cambridge, MA: MIT Press.

Cassam Q (1997) Self and World. Oxford: Oxford University Press.

Cole J and Paillard J (1995) Living without touch and peripheral information about body position and movement: Information from deaferented subjects. In: Bermudez JL, Marcel AJ, and Eilan N (eds.) The Body and the Self. Cambridge, MA: MIT Press.

Evans G (1982) The Varieties of Reference. Oxford: Oxford University Press.

Gallagher S (2005) How the Body Shapes the Mind. New York: Oxford University Press.

Gibson JJ (1979) The Ecological Approach to Visual Perception. Boston: Houghton Mifflin.

Knoblich G, Thornton I, Grosjean M, and Shiffrar M (2005) Human Body Perception from the Inside Out. Oxford: Oxford University Press.

Lackner JR (1988) Some proprioceptive influences on the perceptual representation of body shape and orientation. Brain 111: 281–297.

Melzack R (1992) Phantom limbs. Scientific American 266: 120–126.

O'Shaugnessy B (1995) Proprioception and the body image. In: Bermudez JL, Marcel AJ, and Eilan N (eds.) The Body and the Self. Cambridge, MA: MIT Press.

Ramachandran VS and Blakeslee S (1998) Phantoms in the Brain: Probing the Mysteries of the Human Mind. New York: William Morrow.

Shoemaker S (1967) Self-reference and self-awareness. Journal of Philosophy 65(19): 555–567.

* This article draws upon material from my book The Paradox of Self-Consciousness (Cambridge MA: MIT Press. 1998) and from my article 'The phenomenology of bodily awareness' in Phenomenology and Philosophy of Mind, edited by D. Woodruff Smith and A. L. Thomasson (Oxford: Oxford University Press. 2005).

Biographical Sketch

José Luis Bermúdez is a professor of philosophy at Washington University of St. Louis, where he is the director of the Philosophy–Neuroscience–Psychology program and the director of the Center for Programs in Arts and Sciences. Dr. Bermúdez graduated from Cambridge University in 1988 and received his PhD in 1992. Before coming to Washington University in St. Louis, he was professor and chair of the philosophy department at the University of Stirling, UK. Professor Bermúdez's research interests are primarily in interdisciplinary philosophy of mind and philosophy of psychology. Topics of recent interest include the nature of mental content, models of psychological explanation, the role and origins of self-consciousness, and the possibility of thought without language. His books include *The Paradox of Self-Consciousness* (MIT, 1998), *Thinking without Words* (Oxford, 2003), *Philosophy of Psychology: A Contemporary Introduction* (Routledge, 2005), and *Decision Theory and Rationality* (Oxford, 2009).

Self: Personal Identity

E T Olson, University of Sheffield, Sheffield, UK

Glossary

Cerebrum – The largest and uppermost part of the brain, responsible for higher cognitive functions.

Dissociative-identity disorder – A rare condition in which a human being acts as if he or she were inhabited alternately by a number of people, with different personalities, memories, and cognitive abilities.

Functionalist theory of mind – A theory that characterizes mental states in terms of their typical causes and effects, including their interactions with other mental states.

Hemispherectomy – The surgical removal of a cerebral hemisphere.

Persistence conditions – The conditions logically necessary and sufficient for a thing to continue existing, as opposed to ceasing to exist.

Personal ontology – The basic metaphysical nature of human people, such as whether we are material or immaterial.

Persistent vegetative state – A long-term condition, caused by brain damage, in which one's life-sustaining functions continue spontaneously but one is permanently unconscious and unable to act purposefully.

Temporal part – A part of an object that has the same spatial extent as that object has during a period of time, and the same temporary properties, but a shorter temporal extent: for example, the first half of a football match.

Introduction

Personal identity deals with philosophical questions about ourselves that arise by virtue of our being people (or persons). Some of these questions are familiar ones that occur to all of us: What am I? When did I begin? What will happen to me when I die? Others are more abstruse.

Many philosophers, following Locke and Hume, give consciousness a central role in answering questions of personal identity. They say that many of these questions are nothing more than questions about the unity and continuity of consciousness, or at least that facts about consciousness and related psychological matters suffice to settle the facts about personal identity. Others disagree, saying that consciousness is irrelevant to most questions of personal identity.

(A terminological note: this article is about the self in the sense that my self is simply myself – me, the author – and not about other senses of the word self.)

The Problems of Personal Identity

There is no single problem of personal identity, but rather a wide range of loosely connected questions. Here are some of the most central and important ones.

The Persistence Question

One question is what it takes for a person (or a human person) to persist or survive from one time to another. What sorts of adventures could you possibly survive, in the broadest sense of the word possible? What sort of event would necessarily bring your existence to an end? What determines which past or future being is you? What, in other words, does our persistence through time consist in?

Suppose you point to a boy or girl in an old class photograph and say, ''That's me.'' What makes you that one, rather than one of the others? What is it about the way that person then relates to you that makes her you? For that matter, what makes it the case that anyone at all who existed back then is you? Is a certain boy or girl you by virtue of

psychological connections to you? Is it, for instance, because you are now able to remember things that happened to her back then? Or should we look instead to brute physical or biological relations? Might that person be you because she is the same biological organism as you are, or because you and she have the same body? Or does the answer lie somewhere else entirely? This is the question of personal identity over time, or the persistence question. It applies to our survival into the future as well as our existence in the past.

Historically this question arises when people wonder whether we might continue to exist after we die (in Plato's *Phaedo*, for instance). Whether this could happen depends on whether biological death necessarily brings one's existence to an end. Imagine that after your death there really will be someone, in the next world or in this one, who is in some ways like you. How would that being have to relate to you for it actually to be you, rather than someone else? What would the higher powers have to do to keep you in existence after your death? Or is there anything they could do? The answer to these questions depends on the answer to the persistence question.

The Population Question

We can think of the persistence question as asking which of the characters introduced at the beginning of a story have survived to become the characters at the end of it. But we can also ask how many characters are on the stage at any one time. What determines how many of us there are now? If there are some six billion people on the earth at present, what facts – biological, psychological, or what have you – make this the case? The question is not what causes there to be a certain number of people at a given time, but rather what there being that number of people consists in. It is like asking what sort of configuration of pieces amounts to black's winning a game of chess, rather than like asking what sorts of moves might lead to its winning. This is the population question. (It is sometimes called the question of synchronic identity, as opposed to the diachronic identity of the persistence question.)

One possible answer is that the number of people at any given time is simply the number of human organisms there are then, which is determined by brute biology. (We may have to discount human organisms in an immature or radically defective state – human embryos and human vegetables, for instance – as many deny that they count as genuine people.) Another answer, proposed by Hume and further developed in Kant, is that the number of people is determined by facts about psychological unity and disunity. These answers may agree in ordinary cases, but diverge in unusual ones. Other answers are also possible.

Commissurotomy – the severing of the corpus callosum connecting the cerebral hemispheres – can produce a certain degree of mental disunity, illustrated in extreme cases by such peculiar behavior as simultaneously pulling one's trousers up with one hand and pulling them down with the other. Some even say that commissurotomy can produce two separate streams of consciousness. There is dispute about how much mental disunity there really is in these cases. But many philosophers say that if there were enough disunity, of the right sort, there would be two people – two thinking, conscious beings – sharing a single organism, each with its own unified consciousness and its own set of beliefs, experiences, and actions. Dissociative-identity disorder (multiple personality) raises similar issues: we can ask whether Dr. Jekyll and Mr. Hyde would be two people or one. (At least this is so if one of them continues to exist while the other is active; otherwise this case is a puzzle for the persistence question.) Does the existence of two or more separate streams of consciousness entail that there are two or more conscious beings?

In these cases, or in imaginary but possible extensions of them, there can be psychological disunity of the sort that is ordinarily present in cases where there are two or more people: there may be two independent sets of beliefs and experiences, and two separate streams of consciousness. Yet at the same time there is biological unity of the sort we find when there is just one person: there is always just one biological organism present, and one brain. If the number of people or thinking beings is determined by facts about psychological unity (e.g., unity of consciousness), there will be two or more people in these cases. If it is determined by nonpsychological facts, such as biological

unity, there will be one. Philosophers disagree about this, though the psychological-unity account is currently more popular.

The Personal-Ontology Question

More generally, we can ask about our basic metaphysical nature. What are the most general and fundamental characteristics of human people? This is one of the more abstruse questions about personal identity, but it is of central importance. We can call it the personal-ontology question. It is really a collection of more specific questions, such as these:

What, at the most basic level, are we made of? Are we made up entirely of matter, as stones are? Or are we made up at least partly of something other than matter?

If we are made up of matter, what matter is it? Is it just the matter that makes up our bodies? Do we extend all the way out to our skin and no further, or might we be larger or smaller than our bodies? Where, in other words, do our spatial boundaries lie? More fundamentally, what fixes those boundaries? If we do extend right out to our skin, why that far and no further? Does this have to do with the extent of a certain sort of conscious awareness, as Locke thought, or does the answer lie in brute biology?

Are we ontologically independent beings? Or is each of us a mere state or aspect of something else? Think of a knot in a rope. It is not ontologically independent. It cannot exist without the rope: you can't take it away in your pocket and leave the entire rope behind. It is a state or aspect of the rope. The same goes for events, such as the rope's gradually wearing out. The rope itself, though, does not seem to be a state or an aspect of anything else: nothing appears to stand to the rope as the rope stands to the knot. It seems to be an ontologically independent being – what metaphysicians call a substance. The question, then, is whether we are substances, like ropes, or whether we are states or aspects or events, like knots. Is there something – an organism or a mass of matter, perhaps – that stands to you as a rope stands to a knot?

An answer to these questions and others like them would tell us our basic metaphysical nature. The persistence and population questions also fall under the personal-ontology question. Here are the main proposed answers to this question:

We might be biological organisms (and thus material substances). This view, held by Aristotle, is currently defended under the name of animalism. Most philosophers, both past and future, however, have rejected it.

We might be immaterial substances or souls, as Plato, Descartes, and Leibniz held. This view faces notorious difficulties in explaining how the soul relates to the body – what enables my soul to animate this particular organism and no other, for instance – and has few advocates today.

Many contemporary philosophers say that we are material things but not organisms. One such view is that we are spatial parts of organisms, such as brains. A more popular view is that we are temporal parts of organisms: you are spatially the same size as your animal body but temporally shorter, in that the animal extends further into the past or future than you do. Yet another currently popular view is that we are nonorganisms made of just the same matter as our animal bodies. The thinking behind this is that the same matter can make up two different objects at once. Specifically, the matter making up a typical human organism also makes up a certain nonorganism, and such things are what we are.

Hume proposed that we are bundles of mental states and events. Our parts are not cells or atoms, he thought, but memories and dreams. We are not substances, but events or processes: Hume compares a person to a theater production. Advocates of this view need to explain which mental states make up a person, or a conscious being more generally: what ties the bundles together, as it were. They must also defend the baffling claim that a bundle of mental states is the sort of thing that can think or be conscious.

A few philosophers espouse the paradoxical view that we don't really exist at all. When we say 'I,' we fail to refer to anything. Your atoms may be real enough; perhaps even the thoughts and experiences we call yours are real; but those atoms or mental states are not parts or states of any conscious being.

To answer the personal-ontology question, we need to know what sort of thing it is – if

anything – that thinks and is conscious. When I wonder what I am, what is it that wonders?

What Matters in Identity

Then there is the question of the practical importance of facts about our persistence through time. We ordinarily care deeply about what happens to us in the future, differently from the way we care about the future welfare of others. Why is this? More specifically, what reason do we have for this special selfish concern? What justifies it? This is the question of what matters in identity.

Imagine that surgeons are going to destroy all of you but your brain and then transplant that organ into my head, and that neither of us has any choice about this. The resulting person, we may suppose, will have all your memories and other psychological features and none of mine, and will think he is you. To make things simple, suppose that we are both entirely selfish. Which of us will have a reason to care about the welfare of the resulting being? Suppose he will be in terrible pain after the operation unless one of us pays a large sum in advance. Which of us will have a reason to pay? And why? (In the same way, we can ask whether the resulting being will be morally responsible for my actions or for yours – or perhaps both, or neither.)

The answer may seem to turn simply on whether the resulting person would be you or I: if he is you, you have a reason to care about his welfare and I don't; if he is me, the reverse is true. And the reason for this may seem blindingly obvious: I care in a special way about what happens to myself in the future simply because he is me. Each person has a special reason to care about her own future welfare, and hers alone. The only one whose future welfare I cannot rationally ignore is myself. Likewise, only I myself can be morally responsible for my past actions. What matters in identity, we might say, is identity.

But some deny this. They say that I could have an entirely selfish reason to care about someone else's welfare, for his own sake. The reason why I care, selfishly, about a certain person's future welfare is not that he is me, but that he relates to me in some other way. It may be, for instance, that I care selfishly about a certain future person because he has inherited my mental life, and can remember my actions and no one else's. Ordinarily, the only person who has inherited my mental life and can remember my actions is me, and so I ordinarily care selfishly only about myself. But the fact that he is me is not the reason why I care about him. I care about him because of his psychological connections with me. If someone else, other than me, were connected with me in that psychological way, he would have what matters in identity, and I ought to transfer my selfish concern to him. Perhaps that person would then also be responsible for my actions. We will return to this theme in a later section.

Understanding the Persistence Question

Let us turn now to the persistence question: what is necessary and sufficient for a past or future being to be you, or what it takes for a human person to continue existing from one time to another. The concept of identity over time, or persistence, is a notorious source of confusion, and the persistence question is often misunderstood.

The question asks what it takes for a being existing at another time to be numerically identical with you. For this and that to be numerically identical is for them to be one and the same: one thing rather than two. Numerical identity is different from qualitative identity or exact similarity: identical twins may be qualitatively identical (nearly), but not numerically identical. A past or future person need not, at that past or future time, be exactly like you are now in order to be you, that is, in order to be numerically identical with you. You don't remain qualitatively the same throughout your life. You change.

Of course, someone might say (as Hume did) that a past or future being could not be you unless he or she were then qualitatively just like you are now. This would mean that no person could survive any change whatever: even blinking your eyes would be fatal, resulting in your ceasing to exist and being replaced by someone else. In that case there would be no point in asking the persistence

question. Virtually all discussions of personal identity over time assume that it is possible for a person to change.

This point enables us to distinguish the persistence question from another one that sounds similar. We sometimes ask what it takes for someone to remain the same person from one time to another – as opposed to simply asking what it is for someone to continue existing. If I were to change in certain ways – if I lost a great deal of memory, say, or my personality changed dramatically – we might describe this by saying that I should no longer be the person I was before, but a different one.

The question of what it takes for someone to remain the same person in this sense is not the persistence question. It is not about numerical identity at all. If it were, it would answer itself, for I could not possibly come to be a numerically different person from the one I am now. Nothing can start out as one thing and end up as another thing: a thing can change qualitatively, but it cannot change numerically. This has nothing to do with personal identity in particular, but reflects the logic of identity in general. Those who say that certain events could make you a different person from the one you were before mean that you would still exist, but would have changed qualitatively in some profound and important way. They mean that you might come to be a radically different kind of person. If the resulting person were not numerically identical with you, it would not be the case that you yourself were a different person; rather, you would have ceased to exist and been replaced by someone else.

Proposed Answers to the Persistence Question

There have been three main sorts of answers to the persistence question. The most popular, the psychological-continuity view, says that some psychological relation is both necessary and sufficient for one to persist. You are that future being that in some sense inherits its mental features – beliefs, memories, preferences, the capacity for rational thought and consciousness, or the like – from you; and you are that past being whose mental features you have inherited in this way.

This view comes in different versions, varying in the sort of inheritance that figures in them: whether a future being, in order to be you, needs to acquire its mental features from you via a continuously functioning brain, for instance, or whether some less direct transfer, such as Star Trek teleportation, might suffice. They also vary in the sort of mental features that need to be inherited – whether mental contents, such as memories, or core mental capacities such as the capacity for thought and consciousness, for instance. And they vary in how much of your mental life they require to be preserved in order for you to survive. Presumably it would not suffice, for some future person to be me, for him to inherit only one of my belief-states. But how much is enough? Yet another disagreement has to do with fission cases, of which more later. Most philosophers writing on the persistence question since the mid-twentieth century have followed Locke in endorsing some version of the psychological-continuity view.

The second answer is the brute-physical view or bodily criterion, according to which our identity through time consists in some brute physical relation. You are that past or future being that has your body, or that is the same biological organism as you are, or the like. Whether you survive or perish has nothing to do with psychological facts. This is typically held in combination with the view that we are biological organisms (animalism), for organisms, including human animals, appear to persist by virtue of brute physical facts. Here too there are variants.

Hybrid views are also possible: one might propose that we need both mental and physical continuity to survive, or that either would suffice without the other. For present purposes we can treat these as versions of the psychological-continuity view.

Both the psychological-continuity and brute-physical views agree that there is something that it takes for us to persist – that our identity through time consists in or necessarily follows from something other than itself. A third view denies this. Mental and physical continuity are evidence for

our persistence, it says, but do not always guarantee it, and are not required. No sort of continuity is both absolutely necessary and completely sufficient for us to persist. There are no informative and nontrivial persistence conditions for people. This is called the simple view or anticriterialism. It is often, though not always, combined with the view that we are immaterial souls.

The Psychological-Continuity View

The psychological-continuity view is intuitively appealing. Suppose your brain, or your cerebrum, were transplanted into my head, resulting in a being with most of your mental features and few if any of mine. He would, of course, believe that he was you and not me. Many people find it obvious that he would be right, precisely because of his psychological connections with you. (The brute-physical view, by contrast, seems to imply that he would be me, and that you would stay behind with an empty head if your cerebrum were transplanted. Most people find this strongly counterintuitive.) But it is notoriously difficult to get from this conviction to a plausible answer to the persistence question.

What psychological relation might our identity through time consist in? Memory seems to play an important role: the one who got your transplanted brain would seem to be you at least partly because he or she would remember your life. So one version of the psychological-continuity view is that a past or future being is you just in the case that you can now remember an experience that being had then, or that being can then remember an experience you are having now. More precisely,

> Necessarily, a person x existing at time t is identical to a being y existing at another time, t*, if and only if x can remember, at t, an experience y has at t* or y can remember, at t*, an experience x has at t.

This sort of view, the memory criterion, is often attributed to Locke (though it is doubtful whether he or anyone else actually held it).

The memory criterion faces two well-known problems. First, suppose a young student is arrested for drunken excesses. As a middle-aged neuroscientist, she retains a vivid memory of this event. In her dotage, however, she remembers her science career, but has entirely forgotten the arrest and all the other events of her youth. According to the memory criterion, the young student would be the middle-aged scientist, the scientist would be the old woman, but the old woman would not be the young student. This is an impossible result: if x and y are one and y and z are one, x and z cannot be two.

The second problem is that it seems to belong to the very idea of remembering an experience that you can remember only your own. You can no more remember someone else's experiences than you could be a married bachelor. To remember being arrested (or the experience of it) is to remember yourself being arrested. That makes it trivial and uninformative to say that you are the person whose experiences you can remember – that memory continuity is sufficient for personal identity. It is uninformative because you could not know whether someone genuinely remembered a past experience without already knowing whether she was the one who had it. Suppose we want to know whether Blott, who exists now, is the same as Clott, whom we know to have existed at some time in the past. The memory criterion tells us that Blott is Clott if Blott can now remember an experience of Clott's that occurred at that past time. But even if Blott seems to remember one of Clott's experiences from that time, this counts as genuine memory only if Blott really is Clott. We should already have to know who is who before we could apply the theory that is supposed to tell us.

(Note, however, that this is a problem only for the claim that memory connections are sufficient for identity, not for the claim that they are necessary. There is nothing trivial or uninformative in saying that a future being can be you only if he or she can then remember an experience you are having now.)

One response to the first problem is to modify the memory criterion by switching from direct to indirect memory connections: the old woman is the young student because in her old age she can recall experiences the scientist had at a time when the scientist remembered the student's life. The second problem is traditionally (and controversially) met by inventing a new concept, quasimemory, which is just like memory but without the identity requirement: even if it is self-contradictory to say that

I remember an experience of someone else's, I could still quasiremember it. Neither move gets us far, however, for even the modified memory criterion faces a more obvious objection: there are many times in my past that I cannot remember or quasiremember at all, and to which I am not linked even indirectly by an overlapping chain of memories. There is no time when I could recall anything that happened to me while I was dreamlessly sleeping last night. The memory criterion has the absurd implication that I did not exist at any time when I was completely unconscious.

A better solution appeals to causal dependence. We can define two notions, psychological connectedness and psychological continuity. A being is psychologically connected, at some future time, with me as I am now just if he is in the psychological states he is in then in large part because of the psychological states I am in now. Having a current memory (or quasimemory) of an earlier experience is one sort of psychological connection – the experience causes the memory of it – but there are others. Importantly, one's current mental states can be caused in part by mental states one was in at a time when one was unconscious: for example, most of your current beliefs are the same ones you had while you slept last night. We can then define the second notion thus: I am now psychologically continuous with a past or future being just if my current mental states relate to those he is in then by a chain of psychological connections.

This enables us to avoid the most obvious objections to the memory criterion by saying that a person who exists at one time is identical with something existing at another time if and only if the first is, at the first time, psychologically continuous with the second as she is at the second time.

Fission

A serious difficulty for the psychological-continuity view is the fact that you could be psychologically continuous with two future people at once. If your cerebrum were transplanted, the resulting person would be psychologically continuous with you (even if, as recent neuroscience has shown, there would also be important psychological differences). If we destroyed one of your cerebral hemispheres, the resulting being would also be psychologically continuous with you. (Hemispherectomy – even the removal of the left hemisphere, which controls speech – is considered a drastic but acceptable treatment for otherwise-inoperable brain tumors.) What if we did both at once, destroying one of your cerebral hemispheres and transplanting the other? Then too, the one who got the transplanted hemisphere would be psychologically continuous with you, and according to the psychological-continuity view he or she would be you.

But now let both hemispheres be transplanted, each into a different empty head. The resulting beings – call them Lefty and Righty – will each be psychologically continuous with you. The psychological-continuity view as we have stated it implies that any future being who is psychologically continuous with you must be you. It follows that you are Lefty and also that you are Righty. But that cannot be, for Lefty and Righty are two, and one thing cannot be numerically identical with two things. (If you and Lefty are one, and you and Righty are one, Lefty and Righty cannot be two.) This is the fission problem.

Psychological-continuity theorists have proposed two different solutions to this problem. One, sometimes called the multiple-occupancy view, says that if there is fission in your future, then there are, so to speak, two of you even now. What we think of as you is really two people, who are now exactly similar and located in the same place, doing the same things and thinking the same thoughts. The surgeons merely separate them.

The multiple-occupancy view is almost invariably combined with the thesis that people and other persisting things are made up of temporal parts. For each person there is, for instance, such a thing as her first half, which is just like the person only briefer – something analogous to the first half of a football match. On this account, the multiple-occupancy view is that Lefty and Righty coincide before the operation by sharing their preoperative temporal parts, and diverge later by having different temporal parts located afterward. Lefty and Righty are like two roads that coincide for a stretch and then fork, sharing some of their spatial parts

but not others. At the places where the roads overlap, they will look just like one road. Likewise, the idea goes, at the times before the operation when Lefty and Righty share their temporal parts, they will look just like one person – even to themselves. Whether people really are made up of temporal parts, however, is a disputed metaphysical question.

The other solution to the fission problem abandons the intuitive claim that psychological continuity by itself suffices for one to persist. It says, rather, that you are identical with a past or future being who is psychologically continuous with you only if no other being is then psychologically continuous with you. (There is no circularity in this. We need not know the answer to the persistence question in order to know how many people there are at any one time; that comes under the population question.) This means that neither Lefty nor Righty is you. They both come into existence when your cerebrum is divided. If both your cerebral hemispheres are transplanted, you cease to exist – though you would survive if only one were transplanted and the other destroyed.

This proposal, the nonbranching view, has the surprising consequence that if your brain is divided, you will survive if only one half is preserved, but you will die if both halves are. That is just the opposite of what most of us expect: if your survival depends on the functioning of your cerebrum (because that is what underlies psychological continuity), then the more of that organ we preserve, the greater ought to be your chance of survival. In fact the nonbranching view implies you would perish if one of your hemispheres were transplanted and the other left in place: you can survive hemispherectomy only if the excised hemisphere is immediately destroyed.

The nonbranching view makes the question of what matters in identity especially acute. Faced with the prospect of having one of your cerebral hemispheres transplanted, there would seem to be no reason to prefer that the other be destroyed. Most of us would rather have both preserved, even if they end up in different heads. Yet on the nonbranching view that is to prefer death over continued existence. Some philosophers infer from this that that is precisely what we ought to prefer. Insofar as we are rational, we don't want to continue existing. Or, at least we don't want it for its own sake. I only want there to be, in the future, someone psychologically continuous with me, whether or not he is strictly me. Likewise, even the most selfish person has a reason to care about the welfare of the beings who would result from her undergoing cerebral fission, even if, as the nonbranching view implies, neither would be her. In the fission case, the sorts of practical concerns you ordinarily have for yourself seem to apply to someone who isn't strictly you. This suggests more generally that facts about who is numerically identical with whom have no practical importance. What matters practically is, rather, who is psychologically continuous with whom.

The Beginning and End of Life

Advocates of the psychological-continuity view commonly discuss its implications about who is who in imaginary science-fiction scenarios. But the view also has important implications about when we begin and end in real life.

We ordinarily suppose that each of us was once an embryo and then a fetus. We suppose that the fetus my mother once carried was born, grew up, and later wrote this article. (This does not imply that a fetus in the womb is already a person, but at most that it is a potential person: something that can come to be a person. We might start out as nonpeople and only later become people, just as we are first children and later adults.) But the psychological-continuity view implies that none of us was ever an embryo. Whatever exactly psychological continuity amounts to, you can never be psychologically continuous with a being that has no psychological states at all: you cannot inherit your current psychological states from a being that has none to pass on. And the embryo from which you developed had no psychological states. If psychological continuity is necessary for you to persist, as the psychological-continuity view says, then you were never an embryo. The embryo from which you developed was not you. Not only was it not a person; it was not even something that could come to be a person. You did not come into being when the embryo did, but some time later.

How much later? The psychological-continuity view implies that you could not have existed at a

time when you had no psychological states or properties at all. The earliest you could have begun to exist is when the fetus you developed from acquired its first mental states. No one is certain when that was, but embryologists agree that the fetus has no mental states – not even the capacity to feel pain – before mid-gestation at the earliest. So the psychological-continuity view implies that none of us was ever a 4-month-old fetus. The fetus you see in an ultrasound scan is not something that can one day come to be a person.

According to the psychological-continuity view, when you and I come into existence depends on what sort of psychological continuity our identity over time consists in. Many psychological-continuity theorists say that we do not come into being when the first mental capacities develop in the course of fetal development, but only with the advent of more sophisticated mental properties, such as the capacity for self-consciousness. On their view, we do not begin to exist until we are born, or perhaps several months or even years afterward.

The psychological-continuity view also has important implications about when we come to an end. It implies, for instance, that you could not exist in a persistent vegetative state. A being in such a state has no mental states, and thus has not in any way inherited yours. There is complete psychological discontinuity between you as you are now and the human vegetable that would result if you were to lapse into such a state. If all your mental states were destroyed, you would cease to exist, even if the rest of you kept on working as before. The resulting being, despite being alive by most definitions of the word (and in all current legal systems), would not be you. If the doctors withdraw the tubes that feed it, they do not kill anyone, or allow anyone to die. Some being or other dies when the feeding tubes are withdrawn, of course, but not anything that was ever a person. This too goes against what many of us believe.

The reason we tend to think that each of us was once a fetus and might end up in a vegetative state is that in each of these cases there is a good deal of brute physical continuity of an appropriate sort: the same sort of physical continuity we normally exhibit from one day to the next. So these considerations seem to support the brute-physical view as against the psychological-continuity view. More generally, the brute-physical view appears to give the right verdicts about who is who in all real-life cases.

The Persistence Question and Personal Ontology

We have been evaluating answers to the persistence question by considering their implications about who is who in cases where they disagree. We saw that the psychological-continuity view gives attractive results in science fiction stories (such as the cerebrum transplant), but has implausible consequences about when we begin and end in real life. With the brute-physical view it is the other way round.

But this sort of debate over cases is not the only relevant consideration. A good answer to the persistence question must also be compatible with an acceptable answer to the personal-ontology question. What sort of things could we be if the psychological-continuity view were true? That is, what sort of things could persist by virtue of psychological continuity?

The psychological-continuity view appears to rule out our being organisms. It says that our persistence consists in some sort of psychological continuity. As we have seen, this means that you would go along with your transplanted cerebrum, because the one who ended up with that organ, and no one else, would be psychologically continuous with you. And it implies that if you were to lapse into a persistent vegetative state, you would cease to exist, because no one would then be psychologically continuous with you. But the persistence of a human organism does not consist in any sort of psychological continuity. If we transplanted your cerebrum, the human organism – your body – would not go along with that organ. It would stay behind with an empty head. If you were an organism, then you would stay behind with an empty head, contrary to the psychological-continuity view. Likewise, no human organism ceases to exist by lapsing into a persistent vegetative state. If you were an organism, you could survive as a human vegetable. What the psychological-continuity view says about our persistence through time is not true of human organisms. So if that view is true, we could not be organisms.

Psychological-continuity theorists must deny that we are organisms, and find another account of what we are. Some say that we are temporal parts of organisms. (If your cerebrum were transplanted, you would made up of temporal parts of two organisms.) Others say that we are material things made of the same matter as organisms, or bundles of mental states. But the view that we are not organisms faces an awkward problem.

The Too-Many-Minds Problem

Even if you are not an organism, your body is. That organism – a human animal – thinks and is conscious. In fact, it would seem to be psychologically indistinguishable from you. So if you are not that animal, but something else, it follows that there is a conscious, intelligent being other than you, now sitting in your chair and thinking your thoughts. This means that there are at least twice as many thinking beings as the census reports: for each of us, there is another thinking, conscious being, namely the animal we call one's body. Worse, you ought to wonder which of the two thinkers is you. You may believe that you are the nonanimal (because you accept the psychological-continuity view, perhaps). But the animal has the same grounds for believing that it is a nonanimal as you have for supposing that you are. Yet it is mistaken. For all you know, you might be the one making this mistake. If you were the animal and not the person, you would never be any the wiser.

An analogy may help here. Imagine a three-dimensional duplicating machine. When you step into the in box, it reads off your information and assembles a perfect duplicate of you in the out box. The process causes temporary unconsciousness, but is otherwise harmless. Two beings wake up, one in each box. The boxes are indistinguishable. Because each being will have the same apparent memories and perceive identical surroundings, each will think that he or she is you, and will have the same evidence for this belief. But only one will be right. How could you ever know, afterward, whether you were the original or the newly created duplicate?

In the same way, the psychological-continuity view appears to leave you without any grounds for supposing that you are the being who persists by virtue of psychological continuity, rather than the animal. So even if the psychological-continuity view is true, it seems that you could never know whether it applied to you: for all you can tell, you may instead be an organism with brute physical persistence conditions. This is the 'too-many-minds' or 'too-many-thinkers' problem. It threatens to make any view according to which we are not organisms look absurd. Only animalism, the view that we are organisms (and that there are no beings who persist by virtue of psychological continuity), appears to escape it.

Psychological-continuity theorists have proposed two ways of solving the problem. One is to argue that despite appearances, human animals are not psychologically indistinguishable from us. Although our animal bodies share our brains, are physically just like us, and show all the outward signs of consciousness and intelligence, they themselves do not think and are not conscious.

If human organisms cannot be conscious, then presumably no biological organism of any sort could have any mental properties at all. Why not? It may be, as Descartes and Leibniz argued, because organisms are material things. Only an immaterial thing could think or be conscious. You and I must therefore be immaterial. Though this would solve the too-many-thinkers problem, it raises many others, and few philosophers nowadays accept it.

One notable attempt to explain why organisms should be unable to think that is compatible with our being material things, appeals to the nature of mental states and properties. It says that whatever thinks or is conscious must persist by virtue of psychological continuity. That is because it belongs to the nature of a mental state that it tend to have certain characteristic causes and effects in the being that is in the state, and not in any other being. (This is a version of the functionalist theory of mind.)

For instance, your preference for chocolate over vanilla must tend to cause you, and no one else, to choose chocolate. Now if an organism were to have such a preference, that state might cause another being to choose chocolate, for an organism's cerebrum might be transplanted into another organism. That would violate the proposed account of mental

states. It follows, on that account, that no organism could have a preference; and similar reasoning goes for other mental states. The persistence conditions of organisms are incompatible with their having mental properties. But a material thing that would go along with its transplanted cerebrum – a being of which the psychological-continuity view was true – could have mental states. It would follow that you and I, who obviously have mental states, persist by virtue of psychological continuity, and thus are not organisms. This would solve the 'too-many-thinkers' problem, and at the same time show that the psychological-continuity view is true. It is, however, a minority view.

The other alternative for psychological-continuity theorists is to concede that human organisms think as we do, so that you are one of two beings now thinking your thoughts, but try to explain how we can still know that we are not those organisms. One strategy for doing this focuses on the nature of personhood and first-person reference. It proposes that not just any being with mental properties of the sort that you and I have – rationality and self-consciousness, for instance – counts as a person. A person must also persist by virtue of psychological continuity. It follows from this that human animals, despite being psychologically just like us, are not people.

The proposal goes on to say that personal pronouns such as 'I' refer only to people. This means that when your animal body says or thinks 'I,' it does not refer to itself. Rather, it refers to you, the person who says it at the same time. When the animal says 'I am a person,' it does not thereby express the false belief that it is a person, but rather the true belief that you are. It follows that the animal is not mistaken about which thing it is, because it has no first-person beliefs about itself at all. And you are not mistaken either. You can infer that you are a person from the linguistic facts that you are whatever you refer to when you say 'I,' and that 'I' never refers to anything but a person. You can know that you are not the animal thinking your thoughts because it is not a person, and personal pronouns never refer to nonpeople.

Although this proposal avoids the surprising claim that organisms cannot have mental properties, it gives a highly counterintuitive view of what it is to be a person. And it still implies that there are twice as many intelligent, conscious beings as we thought. It remains a controversial view.

Conclusion

The answer to the persistence and personal-ontology questions is likely to depend on more general metaphysical matters. Which account of personal identity you find attractive will depend on your view about the metaphysics of material things in general.

Consider, for example, the view that all persisting things are made up of temporal parts, mentioned earlier. As we saw, psychological-continuity theorists who hold this view can say that if your cerebrum were divided and each half transplanted into a different head, you would survive twice over, as it were. They need not accept the nonbranching view, according to which you would cease to exist, but would survive if just one cerebral hemisphere were preserved. So temporal-parts theorists will find the psychological-continuity view more attractive than those who reject temporal parts.

The temporal-parts theory also implies that animalists face their own version of the too-many-thinkers problem. Suppose you are an animal, and that you are made up of temporal parts. Then there is a temporal part of you that is just like you are now, except that it extends in time only from midnight last night until midnight tonight: your today-part. It is now sitting in your chair, fully conscious and thinking your thoughts. How could you ever know that you are not your today-part, and have only hours to live? This looks just as troubling for the animalist as the problem of how you can know you are not your animal body is for the psychological-continuity theorist. And any solution the animalist can give to her version of the problem will suit the psychological-continuity theorist equally well for hers. (Temporal-parts theorists solve the too-many-thinkers problem by appealing to the account of personhood and first-person reference sketched in the previous section.)

So if the temporal-parts theory is true, the problems facing the psychological-continuity view appear considerably less than they would

otherwise be. Whether persisting things really do have temporal parts, however, is not a question about personal identity. It is a general question about the fundamental nature of the world. The main problems of personal identity cannot be debated in isolation.

See also: Autobiographical Memory and Consciousness; The MindBody Problem; Philosophical Accounts of Self-Awareness and Introspection; Self: Body Awareness and Self-Awareness; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Baker LR (2000) Persons and Bodies: A Constitution View. Cambridge: Cambridge University Press.
Garrett B (1998) Personal Identity and Self-Consciousness. London and New York: Routledge.
Hudson H (2001) A Materialist Metaphysics of the Human Person. Ithaca, NY: Cornell University Press.
Kolak D and Martin R (eds.) (1991) Self and Identity: Contemporary Philosophical Issues. New York: Macmillan.
Martin R (2006) The Rise and Fall of Soul and Self: An Intellectual History of Personal Identity. New York: Columbia University Press.
Martin R and Barresi J (eds.) (2003) Personal Identity. Oxford: Blackwell.
Noonan H (2003) Personal Identity. 2nd edn. London and New York: Routledge.
Olson ET (1997) The Human Animal: Personal Identity Without Psychology. New York: Oxford University Press.
Olson ET (2007) What are We? A Study in Personal Ontology. New York: Oxford University Press.
Parfit D (1984) Reasons and Persons. Oxford: Oxford University Press.
Perry J (ed.) (1975) Personal Identity. Berkeley, CA: University of California Press.
Shoemaker S (1999) Self, body and coincidence. Proceedings of the Aristotelian Society Supplementary Volume 73: 287–306.
Shoemaker S and Swinburne R (1984) Personal Identity. Oxford: Blackwell.
Unger P (1990) Identity, Consciousness and Value. New York: Oxford University Press.

Biographical Sketch

Eric Olson was born in 1963 and grew up in Washington State, USA. He was educated at Reed College and Syracuse University. From 1995 to 2003 he was a lecturer in philosophy at the University of Cambridge, and is currently a professor of philosophy at the University of Sheffield.

Self: The Unity of Self, Self-Consistency

R R Vallacher, Florida Atlantic University, Boca Raton, FL, USA

Glossary

Emergence – Emergence occurs when the individual elements of a dynamical system achieve organization by means of their mutual influence. The higher-order properties that result from the mutual adjustment among lower-level elements provide coordination for the lower-level elements. The higher-level state is said to be emergent because it was not inherent in the properties of the lower-level elements and because it was not imposed on the system from forces outside the system. Emergence provides for substantial growth in the complexity of a system's processes and properties. Because of emergence, very complex systems can often be described by very simple models.

Self-concept certainty – The degree to which people have an internally consistent and confident assessment of themselves. Certainty can vary across different aspects of self-concept, although there is a tendency toward global certainty versus uncertainty that transcends different regions of self-structure. Self-concept certainty is the subjective experience of self-concept coherence and unity.

Self-concept differentiation – The tendency for people to form different assessments of themselves in different roles or with respect to different characteristics and competencies. Differentiation results from the tendency toward progressive integration in self-structure. Because complete integration is impossible to attain, the press toward integration stalls at a level beneath global self-understanding. The differentiated aspects of self-concept reflect locally coherent and unified regions of the person's self-structure.

Self-esteem – Global self-evaluation that transcends specific aspects of a person's self-concept. Self-esteem provides the highest level of unity in self-understanding. It is the most investigated dimension of self-concept and has been related to many social psychological phenomena. However, self-esteem often lacks the specificity required for action and decision making with respect to particular domains of personal and interpersonal functioning.

Self-evaluation maintenance – The tendency by people to protect their self-evaluation, either in global terms or with respect to specific dimensions of self-understanding. Self-evaluation maintenance is manifest as a wide variety of specific cognitive and behavioral mechanisms. These mechanisms are interchangeable and can be substituted for one another depending on their relative effectiveness in a given context. Because most people have a favorable self-concept, self-evaluation maintenance often represents an attempt to maintain high self-esteem. For people with an unfavorable self-assessment, though, self-evaluation maintenance may be manifest as an attempt to maintain low self-esteem – a tendency referred to as self-verification.

Self-handicapping – A special case of self-evaluation maintenance characterized by a concern with avoiding circumstances that pose a risk to lowering the assessment of one's self-perceived competencies. Self-handicapping is commonly conceptualized as any action or choice of performance setting that maximizes the opportunity for internalizing success or externalizing failure on a task relevant to one's self-concept.

Threatened egotism – A special case of self-evaluation maintenance characterized by intolerance of criticism and a potential for aggression against those who are perceived to provide criticism. This tendency is observed among people with high self-esteem but low self-concept certainty. It thus represents a defensive reaction to information that would undermine a person's self-concept coherence.

Introduction

The self is the largest structure in the individual's mental system. Virtually every experience from infancy to the present moment has some degree of relevance for a person's understanding of him or herself. The self is also the most chronically accessible mental structure. Even the most important people and issues may remain out of awareness for extended periods of time, but a person's sense of self is never more than a stranger's glance from becoming the focus of conscious attention. In view of the amount, diversity, and accessibility of self-relevant information, the task of integrating this information to form a stable and unified self-concept would seem to be a daunting, if not an impossible task. Yet, people manage to develop a fairly coherent representation of themselves that provides a sense of personal continuity despite changing circumstances and that functions as a frame of reference for decision making and other forms of self-regulation.

Self-Concept Coherence

Most theories of self-concept formation emphasize the internalization of self-relevant information derived from social feedback, social comparison, and various forms of socialization. This internalized representation stabilizes the person's global self-view and enables him or her to respond selectively to new information experienced in daily life. The specifics of the internalization process, however, are not well defined in most accounts. In recent years, greater specificity has been provided by models built on principles of complexity and dynamical systems. Support for this class of models is provided by computer simulations and, to a lesser extent, by new lines of empirical research. In this approach, the generation of a relatively coherent self-concept conforms to the same general rules that are responsible for the emergence of higher order states in a wide variety of complex systems in nature and the physical sciences.

Self-Organization

A person's self-concept represents a large and diverse set of elements of self-relevant information derived in the course of his or her social interactions and other experiences. In a single day, for example, a person may receive explicit or implicit feedback about his or her actions or appearance from several people, with each instance of feedback conveying a somewhat different perspective regarding the person's characteristics. From day to day, week to week, and year to year, there is a steady onslaught of such feedback that is recognized and processed in some fashion by the person.

If the myriad elements of self-relevant information existed in isolation from one another, each specific self-relevant thought that arose in consciousness might subvert the platform for action established by the preceding self-relevant thought. The person's sense of self would be highly fragmented, providing an erratic frame of reference, at best, for interpreting subsequent experiences and deciding among action possibilities. The elements of self-relevant information do not exist in isolation, but rather are compared, adjusted, and assembled into higher-order units, each providing a semblance of integration in self-understanding. The self-concept, in other words, represents a relatively unified structure within which specific thoughts, images, and memories pertaining to the self become mutually supportive.

The process of self-organization, a fundamental feature of contemporary accounts of complexity and dynamical systems, provides a plausible mechanism for the integration of self-relevant information. In a dynamical system consisting of interconnected elements, the state of each element adjusts to the current state of other elements to which it is connected. Because of this mutual influence, elements that are initially different achieve commonality with respect to some parameter.

As a result, higher-order units are created that represent, and provide coordination for, the set of individual elements. No outside agent or external influence is required for such order to emerge. Hence, the process is referred to as self-organization.

The attainment of self-concept coherence by means of self-organization is possible because the elements of self-structure, though diverse by many criteria, can be scaled with respect to the common parameter of evaluation. Self-relevant information runs the gamut of possible self-evaluation, from memories of misdeeds and perceived character flaws to memorable accomplishments and firmly held values. The elements of self-structure do not have a fixed valence, however. Rather, the self-evaluation associated with an element of self-relevant information is malleable and open to influence from the valence of other elements. An element incongruent with neighboring (thematically related) elements may change its valence or change the valence of its neighbors, so as to establish evaluative coherence with the related elements. 'Distractable,' for example, may initially be viewed as a negative feature of the self, but take on positive valence in the context of other self-perceived qualities that together convey an image of oneself as a creative scientist.

Whereas individual thoughts about the self may be malleable, the self-organization of many thoughts into higher-order units (e.g., personality traits, competencies, and social roles) provides a coherent sense of the self that promotes stability in thought, emotion, and action. When confronted with new and incongruent self-relevant information, the mutual influences among the elements comprising a higher-order unit support one another and thus promote resistance to the potential disconfirmation. In more mechanistic terms, a stable and coherent self-structure emerges because the separate elements of self-relevant information influence each other via multiple feedback loops. Congruent elements provide cross-validation for each other, whereas incongruent elements set in motion mechanisms designed to eliminate the incoherence or isolate the incongruent elements. In this scenario, no higher order agent or homunculus is necessary to dictate the integration of self-relevant information.

Self-Concept Differentiation

Once a higher-order unit of self-understanding emerges, it takes on the status of an element that is amenable to integration with other higher-order units. If such integration occurs, the resultant higher-order unit functions as an element that can be then integrated with other functional units, and so on, in a scenario of progressive integration. A person, for example, may initially integrate specific thoughts about him or herself in an academic context in developing a coherent self-view as a 'good test-taker.' This emergent self-view might then be integrated with other locally coherent self-views ('good study habits,' 'good term-paper writer,' etc.) to promote a yet higher-order self-concept as a 'good student.' The 'student' self-view might then become integrated with coherent self-views in other domains (e.g., employee, friend, son, or daughter) to generate a yet higher-order sense of self.

In principle, the tendency for progressive integration could result in a globally coherent self-concept, with elements of self-relevant information across the entire self-structure providing cross-validation for one another. In reality, complete unity in self-concept is unlikely to be observed. Rather, the press for progressive integration is likely to stall at some level, promoting islands of local coherence. Some higher-order aspects of the self may be difficult or impossible to integrate, or they may rarely become salient at the same time and in the same context. A person may have a coherent and positive view of him or herself as a parent, for example, but this self-view may be compartmentalized and not come into contact with other self-views, such as athlete or science fiction reader. Because of the limits to progressive integration, the self-structure becomes differentiated rather than globally unified, with different sets of elements stabilizing on different values of self-evaluation. A person may have a coherent and positive view of him or herself as a parent, for example, and an equally coherent but less flattering view of him or herself as an athlete.

The emergence of a differentiated (locally coherent) self-structure has been modeled on a cellular automata platform. Cellular automata are programmable dynamical systems. In this approach, a finite

set of elements is specified, each of which can adopt a finite number of discrete states. The elements are arranged in a specific spatial configuration that usually takes the form of a two-dimensional lattice or grid. The location of each element on this grid specifies the element's neighborhood. The elements evolve in discrete units of time, such that the state of an element at $t + 1$ depends on the states of the neighboring elements at time t. The dynamics of cellular automata depend on the nature of the updating rule and on the format of the grid dictating the neighborhood structure. Repeated iterations of the updating rule invariably produce locally coherent clusters of elements sharing a common state.

The application of cellular automata to the emergence of self-structure is straightforward. A person's self-concept is conceptualized as a system consisting of n elements, each reflecting a specific aspect of the self, which are represented as cells arranged on a two-dimensional grid (see Figure 1). The physical proximity of any two elements represents their degree of relatedness. Each element is characterized with respect to its current evaluation, which is either positive (denoted by light gray) or negative (dark gray). Some elements are more important than others and have greater weight in self-evaluation. An element's importance, which remains constant in the course of simulation, is denoted by its height. Each element influences and is influenced by its eight neighboring elements (four on the adjacent sides and four on the connecting diagonals).

In the course of simulation, an element chosen at random tries to adjust to its neighboring elements by checking how much influence it receives from them. This process involves weighting the valence of each neighbor by the neighbor's importance. The weighted sum of evaluations of the neighboring elements that results from this process is then compared to the current valence of the element. If the element's valence agrees with the overall evaluation suggested by its neighbors, the valence does not change. If the valence of the element differs from the overall evaluation suggested by its neighbors, the element changes evaluation only if the combined weight of evaluation from the neighboring elements is greater than the element's own weighted evaluation. This means that is relatively easy for neighboring elements to change the evaluation of a relatively unimportant element, but it is difficult to change the evaluation of a more important piece of self-relevant information. This process is repeated for another randomly chosen element, and then again for another element, and so on, until each element has been chosen. In the next simulation step, each element has a chance to adjust its state again. The simulation steps are repeated until the system reaches an asymptote, indicating no further changes in the state of elements (i.e., static equilibrium) or a stable pattern of changes in the system (i.e., dynamic equilibrium).

The process of mutual adjustment among elements promotes the emergence of clusters, as depicted in Figure 1. Self-relevant information that is randomly distributed at the outset (Figure 1(a)) forms well-defined domains composed of elements that share a common valence (Figure 1(b)). The self-system also becomes more polarized in overall evaluation, with more negative elements switching to positive valence than vice versa. In a disordered system, the proportion of positive and negative elements in a region roughly corresponds to the

(a)

(b)

Figure 1 The emergence of coherence in self-structure by means of self-organization. Reproduced from Nowak A, Vallacher RR, Tesser A, and Borkowski W (2000) Society of self: The emergence of collective properties in self-structure. Psychological Review 107, 39–61, with permission from American Psychological Association (APA).

proportion of positive and negative elements in the entire structure. Hence, any given element is likely to be surrounded by more positive than negative elements and thus is likely to experience greater influence in the positive direction. Once the self-structure has become clustered, however, most elements are surrounded by elements of the same valence, so that only the elements on the border of a cluster are subjected to conflicting influences.

The emergence of locally coherent regions tends to stabilize the self-system. The joint influence of all the elements in a region support the current state of any single element, which promotes resistance to change and enables an element to return to its original value if it is changed. In an incoherent region, however, the current state of an element is supported by some elements but undermined by others. Hence, when an element is influenced from the outside, some of the surrounding elements help the element resist the influence, whereas other elements work in the direction of the influence. If an element is overwhelmed by outside influence, there will be little tendency for it to return to its original state because some of the neighboring elements are likely to support the new state.

Global Self-Evaluation

Despite the limits to global self-understanding, people are able to characterize themselves with respect to an overall self-evaluation. Even if a person makes a clear distinction between his or her competence in parenting and athletics, for example, he or she is likely to experience little hesitation in answering questions pertaining to his or her general self-worth compared to other people. Global self-evaluation – commonly referred to as self-esteem – is in fact the most investigated aspect of self-concept. Numerous tests have been developed to measure self-esteem and each has shown that people can be reliably scored on this trait. People differ in their respective level of self-esteem (although most people consider themselves 'better than average') and their level of self-esteem is relatively stable over time.

There are two ways of reconciling self-concept differentiation and the predilection for global self-evaluation. One account suggests that people simply average the self-evaluations associated with specific regions of the self to come up with a global sense of self-worth. A person might feel very positive about his or her intellectual ability, moderately positive about his or her social skills, and have a negative assessment of his or her athletic competence. If the person simply averaged these self-assessments (excellent, good, and bad), his or her self-esteem would be moderately positive. Some self-aspects are more important than others, so variations on this perspective suggest that greater weight is given to the more important aspects when generating a global self-evaluation. The person depicted above should express fairly high self-esteem if he or she attached greater importance to intellectual prowess but fairly low self-esteem if he or she attached greater weight to athletic prowess.

In the other account, self-esteem is assumed to be independent of specific self-views, reflecting instead a basic appraisal of oneself that derives from the nature of one's social experiences and relationships. A secure childhood, for example, may establish an unquestioned view of oneself as accepted and desirable, regardless of how one's specific competencies are evaluated. Conversely, a childhood associated with neglect and parental inconsistency may create a foundation for self-doubt that persists throughout life, even if the person develops competencies in many areas of life. A recent variation on this idea is the sociometer hypothesis, which assumes that humans have a fundamental need to belong to social groups and to have meaningful social relations. A person's global self-evaluation reflects the extent to which these social belongingness needs are fulfilled: the more completely the person's needs are met, the higher the person's self-esteem.

In support of the sociometer hypothesis, high and low self-esteem people do not differ in many ways other than in their respective social skills and social relations. High self-esteem people are not necessarily brighter, more athletic, more skilled, more physically attractive, or even healthier than are people with lower self-esteem. These findings, however, represent generalizations across large samples of people. For any one person, one or more of these dimensions could be vital to his or her self-esteem. But if these self-defining dimensions differ across people, one would not expect to find a connection in the population between any single self-view and people's level of self-esteem.

Self-Concept Certainty

Although there is a tendency toward integration in self-concept, not everyone achieves the same level of differentiation. Some people achieve a fairly integrated self-structure composed of coherent regions corresponding to roles, broad traits, competencies, and values. But other people do not attain this level of integration and have an overly differentiated and fragmented self-concept. People with an optimally differentiated view of themselves have a coherent view of themselves in different roles and settings, and this stable way of thinking about themselves provides a clear sense of how they should behave in different contexts. People with a fragmented sense of self, however, cannot maintain a stable view of themselves, are torn by indecision and ambivalence, and are very susceptible to the feedback about themselves provided by others. Everyone has negative thoughts and feelings about the self, but people with a differentiated self-concept can compartmentalize these elements into separate regions. People who have a fragmented self-concept, however, cannot think about any aspect of the self without encountering a mix of positive and negative thoughts.

An unintegrated sense of self is experienced as self-concept uncertainty. When asked to indicate what they are like on trait dimensions (e.g., sociable vs. unsociable, energetic vs. lazy), people lacking integration can provide an answer (moderately sociable, fairly energetic, and so on). However, when they are asked later to indicate how certain they are about their trait ratings, these people tend to express uncertainty. They may express fairly high certainty when rating their global self-esteem, but when asked to reflect on themselves for a period of time, their moment-to-moment self-evaluation tends to be erratic, revealing the unintegrated state of their self-structure. People with a more certain view of themselves, in contrast, are likely to express thoughts about themselves that are relatively stable in evaluation over time. They may express negative self-evaluative comments, but these comments tend to be clustered in time, separated from periods during which they express primarily positive comments about themselves.

There are two ways to think about self-concept uncertainty. Perhaps people with an uncertain sense of themselves simply do not have a clear view of what they are like and entertain a wide range of possibilities, none of which prove particularly convincing or provide a great deal of stability. The other perspective is that uncertainty does not reflect an erratic view of the self, but rather two (or more) fairly coherent self-views that are mutually inconsistent in the self-evaluation they suggest. In both cases, the person can only be certain of the basic actions he or she has done or is inclined to do. The difference centers on what the person does with this low-level information. In the first case, the person does not know how to integrate these elements into a coherent view of the self; in the second case, the person has two (or more) ways of integrating this information, and these conflicting self-views vie for ascendance and promote ambivalence in self-evaluation.

Insight into this issue is provided by research employing the implicit association test (IAT), a measurement procedure that exposes unconscious attitudes that go undetected by traditional procedures that ask people to provide explicit reports of their attitudes. A white person might indicate positive attitudes toward members of a minority group, for example, but harbor a stereotyped view of the group that is latent and unexpressed most of the time. The IAT has been employed in recent years to expose implicit self-esteem – people's latent and unexpressed self-evaluation. When asked to evaluate themselves on self-report measures such as the Rosenberg self-esteem test, most people portray a relatively positive ('better than average') self-image. Despite having high explicit self-esteem, however, some people harbor a negative self-view that is out of conscious awareness (and expression) most of the time.

A person's implicit self-esteem is revealed in the IAT procedure by assessing how long it takes him or her to associate 'self' words (I, my, me, and mine) with pleasant words (e.g., kind and wonderful) versus unpleasant words (e.g., filthy and stinky). If the person takes considerably longer to associate the self words with unpleasant than with pleasant words, his or her implicit self-esteem is high. But if the person takes relatively little time to associate the self words with the unpleasant words, he or she is said to have implicit self-esteem that is low.

Explicit and implicit self-esteem are correlated but not very strongly, so some people can be characterized as having high explicit self-esteem and low implicit self-esteem.

There is evidence that this combination of conflicting self-views underlies self-concept uncertainty. People characterized by a discrepancy between explicit and implicit self-esteem tend to show relatively high variability over time in global self-evaluation when they generate stream-of-thought narratives about themselves. Thus, a positive self-evaluative comment is as likely to be followed by a negative self-assessment as by another positive comment. As noted above, such variability in momentary self-evaluation is associated with low self-concept certainty. In contrast, people whose implicit self-esteem is congruent with the level of self-regard expressed on an explicit self-report measure tend to show relatively low moment-to-moment variability in their verbalized self-evaluative statements when asked to reflect on themselves for several minutes. Positive comments tend to be followed by other positive comments, and negative self-evaluative remarks, although less frequent, tend to be followed by other negative self-evaluative remarks.

Maintenance of Self-Concept Coherence

Everyday life provides a continuous influx of self-relevant information, and not all of it is favorable or mutually consistent. Some instances of social feedback and social comparison can provide a direct challenge to a specific self-view (e.g., one's social charm) or to global self-regard. Yet people usually recover from inconsistent self-relevant information and maintain their self-view in specific areas as well as their overall self-esteem. As noted earlier, self-relevant information does not have a fixed evaluation, but rather is open to interpretation. Behaving in a critical fashion, for example, is often considered negative (e.g., as harmful or insensitive), but it can also be viewed in positive terms (e.g., as constructive criticism or 'tough love'). The flexibility of self-relevant information, which is instrumental in the formation of a self-concept, is also critical to the defense of a self-concept once it has formed.

Self-Evaluation Maintenance

People are motivated to protect their global self-esteem, as well as their specific self-views in different regions of self-structure, from events or new information that would call the validity of their self-assessments into question. Because self-concept is inherently an evaluative structure, its protection is tantamount to maintaining a characteristic self-evaluative assessment. People prefer to think about themselves in positive rather than in negative terms, and most people develop a generally flattering view of their characteristics, seeing themselves as above average on many different dimensions. Hence, the maintenance of a self-concept for most people is synonymous with maintaining a relatively high level of self-esteem. When confronted with negative self-relevant information, then, people are motivated to process the disconfirmatory information in a manner that renders it consistent with the positive tone of the self-view at issue.

There are a variety of means by which people can change the implications of self-relevant information for purposes of self-protection. Consider, for example, a person with a positive self-assessment of his or her social skills who is called a crashing bore at a party. The person might experience a temporary setback in his or her self-regard, but is likely to mount a defense to restore his or her self-view in this area. He or she might challenge the critic's credibility, for example, noting that the critic is simply a negative person or perhaps a bad judge of character. Alternatively, he or she might consider the behavior that prompted the critique to be unrepresentative of the way he or she normally behaves in this context. Yet another line of defense is to question the critic's motive for the unflattering feedback. The person might conclude that the critic is envious of him or her and is trying to be hurtful rather than informative. Or perhaps the person decides that the critic really likes him or her but fears that the feeling is not reciprocated and so is engaging in a 'reject before you're rejected' strategy to protect his or her own self-esteem.

The mind is remarkable at processing information in service of maintaining a point of view or belief concerning the self. Protection of one's self-concept, even if it means censoring some

information and engaging in outright distortion of other information, goes right to the heart of selfhood. When in self-protection mode, the self can be looked upon as a totalitarian ego, a set of processes whose primary aim is protecting the head of (mental) state. The variety of means available for self-protection has been likened to a self-defense zoo. These means appear to be interchangeable. If discounting the credibility of the source does not work, for example, a person might justify the behavior that was criticized or remind him or herself how unrepresentative the behavior was.

An especially well-researched strategy for protecting one's self-concept is known as the self-serving bias. This simply means that people accept credit for their successes, but blame their failures on external factors or bad luck. Students who perform well on exam, for example, tend to see the grade as a reflection of their intellect, their knowledge of the material, or their preparation. When they do poorly, they see things far differently: the exam was unfair and not a valid measure of their competence.

Intuition suggests that the most vigorous defenses of the self would occur when dealing with potential enemies or people with whom people have little in common. If a person is criticized by a stranger at a party, for example, he or she would probably mount a particularly strong defense. Although this is undoubtedly true, there are circumstances in which the greatest threat to self-esteem comes from those who are closest to the person and with whom he or she has a great deal in common. Insight into this state of affairs is provided by Tesser's model of self-evaluation maintenance (SEM). The SEM model is primarily concerned with social comparison information. Two factors are critical in determining how we respond to such information: the 'closeness of the comparison person' and the 'personal relevance of the behavior' in question. Normally, people hope that someone who is close to them will excel at the activities in which he or she is engaged. People are proud and delighted when their friends or siblings win a tennis tournament or earn straight A's in college.

This is clearly the case when this person is trying to achieve something that is not directly relevant to one's own personal ambitions. The problem for the self arises when the other person excels at something that one also aspires to be good at. Imagine, for example, a person who is a budding tennis star or trying to stand out in college. A close friend or sibling who wins a tennis championship or earns a perfect GPA can make the person uncomfortable because the success represents a direct comparison with your success in that area. According to SEM reasoning, the person might go out of his or her way to find other ways to maintain a positive self-evaluation, or perhaps change his or her pursuits in order to avoid the inevitable comparisons. This scenario can be observed among friends, whose success in a personally relevant area can prove to be highly threatening. Marital partners, too, can experience this sort of difficulty if they have identical career goals. It is unlikely that their respective careers will proceed in lockstep with one another. The first one to receive recognition in the career can pose a serious self-concept threat to the partner whose time has yet to come.

From the perspective of SEM, there are three choices available when someone close to a person experiences success in an area that is personally relevant to him or her. He or she can rethink the importance of that area, distance him or herself from the other person, or sabotage the other person's performance. The partners to a marriage, for example, can decide to follow different career paths. Alternatively, they could reduce their closeness, which is hardly a desirable solution for a married couple. Finally, the threatened partner could try to hinder the loved one's success in various ways.

There is evidence that these options are exercised. However, the mind's ability to process information in service of self-esteem protection can render these strategies unnecessary. The self-defense zoo, alluded to above, enables people to maintain healthy relations with one another when social comparisons prove to be uncomfortable. Brothers and sisters, husbands and wives, and close friends do not have to denigrate one another, sabotage their success, or change their careers when one of them outperforms the other. Instead, they can come up with a clever rationalization or perhaps invoke a self-serving bias to make sense of their successes and failures. If their partners

are truly close, they will assist in this process and help draft an interpretation that keeps both parties happy.

Self-Verification

Defending the self is often synonymous with protecting a positive view of one's self. Negative social feedback and unfavorable social comparisons engage a host of self-defense strategies, which makes sense if one assumes that people are motivated to develop and maintain high self-esteem. Most people do have relatively high self-esteem and feel that they are better than average on almost every dimension. But not everyone has a flattering view of him or herself in every area, nor does everyone have a global view of the self that is highly positive. When such people protect their self-concept, what are they protecting? A positive self-view or the view they actually have of themselves? Imagine a young man with serious doubts about his social charm being told by a young female that he is the most exciting person she has met in years. Would the male feel comfortable with this feedback? If positive self-evaluation were the issue, he might enthusiastically embrace the feedback. If, however, the real issue centered on maintaining a coherent self-concept, he might discount or even reject the feedback as inconsistent with the way he sees himself.

The issue can be framed in terms of positive evaluation versus coherence. Is it more important to feel good about oneself or to feel coherent in one's self-knowledge? Considerable insight into this issue has been generated under the guise of self-verification theory, a model of self-concept developed and researched by William Swann and his colleagues. Swann's basic answer is that coherence trumps evaluation – but not without a struggle. There is evidence that people choose, like, and retain partners (e.g., roommates and friends) who perceive and evaluate them in a way that matches their own self-view. People prefer romantic partners who see them as they see themselves, even if the self-view at stake has some glaring weaknesses. Having a coherent view of the self, even if it is not all that positive, enables one to predict how one will fare in social settings and thus allows one to decide what to do, what contexts to seek out or avoid, and which people to trust.

The tendency to react based on feelings as well as on thinking complicates the picture slightly. Positive feedback obviously feels better than negative feedback, and because feelings occur more quickly and automatically than thoughts, the first reaction of a low self-esteem person to compliments and flattery is acceptance. And if the person is distracted and unable to marshal his or her cognitive resources to mount a defense against the flattery, he or she may continue to prefer hearing things that are positive but inconsistent with his or her self-view. Over time, though, people's thoughts tend to overtake their feelings. For the low self-esteem person, this means that the bubble eventually bursts and he or she will engage in a variety of actions to reestablish coherence in his or her self-concept – even if this means sabotaging a relationship with someone who considers the person wonderful.

Self-Concept Change

If people's efforts were directed solely at maintaining an existing sense of self, they would never display growth or have the potential for benefiting from experience. For that matter, people would never develop a coherent and meaningful self-concept in the first place if their only concern was protecting what they already knew and believed about themselves in childhood. Clearly, the strategies devoted to self-defense must be overcome in some fashion by forces that prompt a reconsideration of what one is like.

Changing a self-concept is not easy. A person can learn from his or her experiences and update his or her self-assessments, but such changes are largely confined to fairly peripheral as opposed to central aspects of the self. If a person thinks of him or herself in terms of intellectual ability but is less invested in his or her athletic prowess, for example, social comparison information is unlikely to have much impact on the former quality but can promote substantial change in the latter. Thus, the person might rationalize receiving a bad grade in a course (e.g., 'the tests were unfair') but view a poor showing on the racquetball court as credible information regarding his or her ability. With this in mind, one might think a substantial and hard-to-rationalize experience is necessary to induce

change in a central aspect of a person's self-concept. To attack someone's intellectual ability, for example, one must present a strong argument that cannot be easily refuted. Such an approach might work if done with sufficient force and without allowing the person much time to launch a counteroffensive – but then again it might only promote strong resistance and perhaps aggression.

A far different strategy follows from the process of self-organization. As discussed earlier, self-understanding is characterized by progressively higher levels of integration. This suggests that when a person thinks about the self in fairly specific or detailed terms, he or she is open to new information that could tie together the low-level information in a meaningful fashion. This idea has interesting implications for self-concept change. Rather than focusing attention directly on a person's integrated self-view, successful agents of influence (e.g., clinical psychologists, friends, or con artists) are likely to experience greater success by inducing the person to think about him or herself in detailed, fragmented terms. In effect, the strategy is to disassemble the person's integrated structure and then provide clues about how to put the pieces together again. In the routine and familiar contexts of everyday life, self-concept change is rare because people are likely to have an integrated sense of what their behavior represents. Such high-level understanding provides an effective 'shield' against new ways of thinking about the self. For self-concept to change, a crucial precondition for emergence must occur – the disassembly of integrated action understanding into more specific ways of thinking about the behavior.

Vallacher and Wegner's action identification theory offers a plausible scenario for the disassembly and reconfiguration process in self-concept change. The theory holds that any action can be identified at different levels, from concrete, mechanistic representations (lower-level identities) to comprehensive representations reflecting goals, values, consequences, and self-evaluative implications (higher-level identities). The act of taking a test, for example, can be identified in low-level terms as 'answering questions' or in higher-level terms as 'showing knowledge,' 'earning a grade,' or 'demonstrating conscientiousness as a student.' There is a preference for higher-level identification, but there are well-documented factors that engender lower-levels of identification. Simply asking a person to recount the details of what he or she has done, for example, can induce a lower level of identification than would otherwise be the case. Interruption of ongoing action, a fairly common feature of everyday life, can also make a person sensitive to the lower-level aspects of his or her behavior.

Because of the preference for higher-level understanding, the disassembly of action into its lower-level identities creates the potential for emergence and thus renders the person susceptible to new ways of thinking about the action's broader meaning. A student induced to identify taking a test as 'answering questions,' for example, might be open to a new understanding of what test-taking means. To the extent that an emergent higher-level identity has implications for self-understanding (e.g., traits, goals, and values), the person's self-concept may undergo a transformation in one or more relevant domains. The test-taker's self-concept as a student, for example, may change from 'seeking knowledge' to 'earning a degree,' perhaps with corresponding changes in related aspects of his or her self-concept (e.g., from 'intrinsically motivated' to 'motivated by outcomes').

Coherence and Adaptation

Self-concept coherence is considered essential for effective and autonomous functioning. The failure to maintain a coherent self-concept is associated with identity diffusion, promotes uncertainty and ambivalence in social relations, and undermines commitment to long-term goals and persistence in effortful task performance. A stable self-concept is also fundamental for the self-regulation of thought, mood, and action. Because self-concept is defined to a large degree in terms of values, it provides a frame of reference for evaluating courses of action and thus enables the person to control impulses, resist temptation, and behave independently of peer pressure. But effective functioning also requires learning and adaptation to changing circumstances. An inability to update one's self-understanding can prevent such learning and adaptation. Coherence is an important feature of self-concept, then, but this must be balanced by

openness to experience that promotes refinement and change in one's representation of the self. The balance between coherence and flexibility can be difficult to achieve and maintain. An imbalance in either direction – excessive flexibility or excessive coherence – is at the root of important dimensions of individual variation in self-concept.

Nonoptimal Flexibility

Self-regulation would be impossible without a relatively unified self-concept in the domain for which self-control or decision-making is at issue. Research has established a connection between self-concept uncertainty – a hallmark of low coherence – and a tendency to identify one's action at a relatively low level of identification. For example, a person with a highly certain sense of self might identify his or her behavior in social interaction as 'being friendly' or 'demonstrating sensitivity,' whereas someone with a less certain self-concept might identify his or her behavior in the same context as 'speaking rapidly,' 'smiling,' and 'listening carefully.' The higher-level identities have direct relevance to self-concept and provide a clear frame of reference for how to behave in other social contexts. The lower-level identities, in contrast, are elastic in meaning, consistent with widely divergent self-views, and thus provide an equivocal basis for how to behave in other contexts. Speaking rapidly and smiling could indicate friendliness or sensitivity, but they could also support a self-view of ingratiation, submissiveness, or anxiety. In line with the emergence scenario of action identification theory, low-level identification renders the person prone to impulsive action as opposed to maintenance of action plans, and open to social influence in deciding what to do.

By the same token, an incoherent self-concept is vulnerable to momentary social feedback, social comparison, and performance setbacks. Everyday life provides a continuous flow of such information and much of this feedback is mutually contradictory. Without a unified sense of one's enduring characteristics, the erratic nature of incoming self-relevant information would promote volatility in thoughts, feelings, and action tendencies. Research has documented that people who lack self-concept certainty and clarity tend to have unstable self-esteem, react strongly to the self-appraisals provided by other people, experience mood fluctuations, and express highly variable self-evaluation when asked to reflect on themselves in a stream-of-thought manner. Recent clinical research has also identified a subset of people with bipolar depressive disorder who lack stable standards for self-evaluation. These people are at higher risk for hospitalization and suicide ideation than are bipolar depressives who oscillate between well-defined criteria (positive and negative) for self-evaluation. The lack of stable frames of reference undermines effective self-regulation and promotes volatility in mood, cognition, and behavior.

Nonoptimal Consistency

Several lines of theory and research suggest that a preoccupation with maintaining an internally consistent self-view can undermine adaptive personal and interpersonal functioning. Two scenarios in particular have been identified and investigated. First, a person may encounter self-relevant information that is inconsistent with his or her prevailing self-view. Rather than accommodate the self-view to such information, the person rejects this information and the information source. This reaction is typically manifest as cognitive processes that reinterpret, rationalize, or discount the meaning and relevance of the information. The processes of self-evaluation maintenance and self-verification, described above, represent this response to inconsistent information. These processes are essential to the maintenance of a coherent self-concept, but become problematic and a sign of rigidity and defensiveness when they prevent any new information from being considered on its merits.

The rejection of inconsistent self-relevant information can also be manifest behaviorally. Research on self-verification has shown that people will sometimes react to inconsistent social feedback by exaggerating the self-perceived quality in question in an attempt to convince the feedback source that his or her perception is inaccurate. This can result in a counterintuitive situation, in which a person with a negative self-view tries to change the opinion of an observer or a relationship partner who views the person positively. Research on

threatened egotism, meanwhile, shows that the behavioral rejection of inconsistent information can be manifest as antisocial action. When a person is concerned with protecting an insecure self-view, a person who challenges that self-view is considered a threat and may be subject to physical aggression. This tendency is especially likely when the self-view in question is unrealistically positive, suggesting that the combination of high (inflated) self-esteem and low self-concept certainty is a risk factor for violence in interpersonal relations.

The second scenario of nonoptimal coherence concerns achievement motivation. People generally seek out performance settings that match their self-perceived capacities or that provide an opportunity to refine or validate these self-views. In some situations, however, the person may have relatively low certainty regarding the self-perceived capacity in question. In such cases, the person is inclined to avoid opportunities that call that capacity into question. If the performance setting cannot be avoided, the person might sabotage his or her performance by creating obstacles that render success unlikely. Prior to a test of intellectual performance, for example, a person might turn down the chance to engage in practice sessions that could facilitate his or her performance. Sabotage might also take the form of using performing-inhibiting drugs. The rationale for such self-handicapping is that the person is more concerned with maintaining a self-view than with engaging in actions that might result in failure and thus call into question the validity of the self-view. The likelihood and manifestation of self-handicapping shows individual variation and, like individual variation in the tendency toward threatened egotism, is associated with low self-concept certainty.

Summary and Conclusions

The development and maintenance of a self-concept is exceedingly rare in the animal kingdom. A few other species (the great apes and dolphins) reach the cognitive threshold required for self-awareness, but humans alone have the capacity to develop a mental representation that consists of personal characteristics ranging from physical features to personality traits, beliefs, and values. This human tendency is generally considered to have important evolutionary advantages, but the expression of these advantages requires a self-concept that is sufficiently unified and coherent to provide a platform for regulating other psychological processes.

Attention is focused on one's own characteristics and processes when one is faced with real or perceived problems centering on self-control and decision-making. The self is rarely activated when other psychological processes are effective and running smoothly in accordance with environmental demands, personal and interpersonal expectations, and social rules. The restricted range of relevance suggests that the basic function of the self is to identify and repair processes that are not performing at a level that is personally adaptive. It is noteworthy in this regard that self-focused attention is associated with performance impairment, critical evaluation by other people, and a set of unique 'self-conscious emotions' that typically have negative rather positive valence.

The self-regulatory function of explicit self-focused attention has been explored and documented within several research paradigms over the past two decades. A common thread to these accounts is the assumption that people compare their past, present, or possible action to a personal standard or characteristic when evaluating the action. If this comparison process suggests that the action is inconsistent with the regulatory standard, a variety of cognitive, affective, and behavioral mechanisms are engaged to eliminate or reduce the inconsistency.

Self-regulation would be impossible without a relatively coherent self-concept in the domain for which self-control or decision making is at issue. To control one's impulses, resist temptations, delay immediate gratification in favor of a distant reward, take on challenges and risks, and behave in an altruistic manner require a stable and unambiguous frame of reference for deciding the personally appropriate course of action. An internally inconsistent view of the self in the relevant domain would lead to equivocation, ambivalence, indecision, and cognitive distortion in service of the line of least resistance.

The self-concept must also retain sufficient flexibility to accommodate the inherent dynamism

and complexity of everyday life. The same processes that promote the formation of an integrated self-concept are at work when the self-concept undergoes transformation in response to new demands and challenges. In effect, transformation involves the reverse engineering of a self-structure, such that coherent regions are disassembled into their component elements. Because of the press for higher-order integration, the elements comprising the disassembled region are susceptible to influences that suggest a new configuration of self-understanding. With repeated iterations of the assembly and disassembly scenario, the self-structure represents a dynamic equilibrium that provides both unity and flexibility in people's self-understanding.

See also: Meta-Awareness; Philosophical Accounts of Self-Awareness and Introspection; Self: Body Awareness and Self-Awareness; Self: Personal Identity.

Suggested Readings

Baumeister RF (ed.) (1999) The Self in Social Psychology: Key Readings in Social Psychology. Philadelphia, PA: Psychology Press/Taylor & Francis.

Baumeister RF, Smart L, and Boden JM (1996) Relation of threatened egotism to violence and aggression: The dark side of high self-esteem. Psychological Review 103: 5–33.

Baumeister RF and Twenge JM (2003) The social self. In: Weiner I (series ed.) & Millon T and Lerner MJ (vol. eds.) Handbook of Psychology, Vol. 5: Personality and Social Psychology, pp. 327–352. New York: Wiley.

Campbell JD, Trapnell PD, Heine SJ, Katz IM, Lavallee LF, and Lehman DR (1996) Self-concept clarity: Measurement, personality correlates, and cultural boundaries. Journal of Personality & Social Psychology 70: 141–156.

Carver CS and Scheier MF (1999) Themes and issues in the self-regulation of behavior. In: Klyer RS Jr. (ed.) Advances in Social Cognition, vol. 12, pp. 1–105. Mahwah, NJ: Erlbaum.

Leary MR and Tagney J (eds.) (2003) Handbook of Self and Identity. New York: Guilford Publications.

Nowak A, Vallacher RR, Tesser A, and Borkowski W (2000) Society of self: The emergence of collective properties in self-structure. Psychological Review 107: 39–61.

Swann WB Jr. (1996) Self Traps: The Elusive Quest for Higher Self-Esteem. New York: Freeman.

Tangney JP and Fischer KW (eds.) (1995) The Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride. New York: Guilford.

Tesser A (1988) Toward a self-evaluation maintenance model of social behavior. In: Berkowitz L (ed.) Advances in Experimental Social Psychology, vol. 21, pp. 181–227. San Diego: Academic Press.

Tesser A, Felson RB, and Suls JM (eds.) (2000) Psychological Perspectives on Self and Identity. Washington, DC: American Psychological Association.

Vallacher RR, Nowak A, Froehlich M, and Rockloff M (2002) The dynamics of self-evaluation. Personality and Social Psychology Review 6: 370–379.

Vallacher RR and Wegner DM (1987) What do people think they're doing? Action identification and human behavior. Psychological Review 94: 1–15.

Biographical Sketch

Robin Vallacher is a professor of psychology at Florida Atlantic University, and research affiliate at the Center for Complex Systems, Warsaw University. He has been a visiting scholar at the University of Bern, Switzerland, and at the Max Planck Institute for Psychological Research in Munich. Dr. Vallacher has published research on a wide variety of topics in social psychology, from individual-level processes such as self-concept, self-regulation, and social judgment, to interpersonal and collective phenomena

Sensory and Immediate Memory

N Cowan, University of Missouri, Columbia, MO, USA

Glossary

Auditory stream segregation – The tendency for two types of sound that are presented in an intermixed fashion to be perceived as coming from separate sources, or perceptual streams; this is more likely to occur at faster presentation rates.

Continuous moment hypothesis – The hypothesis that events are perceived together when they fall into the same sliding window of time; this window is continually updated.

Discrete moment hypothesis – The hypothesis that events are perceived together in nonoverlapping bins of time; any events falling into the same bin are perceived as simultaneous.

Immediate memory – Temporary information about a stimulus set, such as a word list or object array, that can be retained long enough to allow correct performance on a test shortly afterward, with no intervening interference or distraction. As used here, it includes both sensory memory and temporary memory for concepts.

Perceptual memory – A term that can be considered a synonym for sensory memory. See the introduction for possible differences.

Perceptual moment – Also sometimes called the psychological moment; the small amount of time in which two successive events seem simultaneous, which happens only when they are within about 200 ms of one another.

Psychological present – A period of up to a second or so, during which a single event seems to take place.

Sensory memory – Temporarily remembering how certain things look, sound, feel, taste, or smell.

Short-term memory – A term that can be considered a synonym for immediate memory. See the introduction for possible differences.

Working memory – A term that can be considered a synonym for immediate memory.

Introduction to Concepts and Terms

Sensory memory is temporarily remembering how certain things look, sound, feel, taste, or smell. Immediate memory refers to information about a stimulus set, such as a word list or object array, that can be retained temporarily to allow correct performance on a test shortly afterward, with no intervening interference or distraction. Researchers differ in their definitions but, as used here, immediate memory includes both sensory memory and also the short-lived memory of concepts. The concepts can come either from stimuli or from thoughts.

What is distinct about immediate memory is that it is short-lived; it has a certain richness that fades with time or is eliminated by subsequent stimuli that cause distraction or interference. What is left when it is gone is a more impoverished, yet still critically important, long-term memory representation of sensations, events, and concepts from the past.

Researchers also vary in the particular terms they use. For some, a synonym for sensory memory is perceptual memory. Other researchers would think of perceptual memory to include identification of the objects that are being perceived (e.g., ''I am seeing a chair'' or ''I am hearing a bird''), but would exclude that information from sensory memory. For some, synonyms for immediate memory are short-term memory or working

memory. Other researchers would think of short-term memory as referring only to the information automatically held, whereas working memory or immediate memory would include that information but also information held with the benefit of attention and mnemonic strategies, such as covert verbal rehearsal. We will circumvent these disagreements and stick to the terms sensory and immediate memory.

At any moment, you are likely to be aware of some of the information contributing to sensory and immediate memory, but not other information. Imagine that you are at a large gathering of friends in a large room, with refreshments. The vast field of chatter in the room impinges on your auditory system, but you cannot make out all of the conversations and, mostly, your awareness of this chatter recedes into the background. You are holding a conversation with one person, on financing a new automobile, and several other people nearby are talking about a topic that interests you, a new political sex scandal. It is impossible for you to attend fully to both of these things at once. As you start picking up more about the sex scandal, your understanding of what you are being told about automobile financing starts to wane, as you struggle to stay polite. You notice your conversational partner making a distressed facial expression as he sees your increasing boredom with the topic and, to allay his distress, you say something to demonstrate that you are in fact paying attention. You do not remember much of the content of the recent discussion, but you still can recall the last few words that he said, which you repeat in an interested, questioning tone of voice. The general chatter in the room, the snippets from the sex conversation, and the monologue about car finance are all part of your sensory, perceptual, and immediate memory.

Attention and Immediate Memory

A very important distinction is between information to which you pay attention and information that you ignore. The ignored information does not disappear immediately. Think of the last few words of the car financing conversation that you could repeat although you were not fully listening. Similarly, while you are engrossed in a novel, someone may ask you what time it is. You ask them to repeat what they said but, before they do repeat it, you are able to recall what was asked. How did you do that? You pulled it from sensory memory into a more conceptual immediate memory.

Donald Broadbent studied this kind of issue in the 1940s and 1950s, helping to establish a new field called cognitive psychology, the study of perception, memory, and thought processes. He and his colleagues in England were interested in some practical issues, such as how a pilot could land a plane safely when the radio instructions arriving through headphones are mixed with instructions being delivered simultaneously to other pilots landing other planes. That problem was eventually solved through advances in radio technology. In the psychology laboratory, however, the problem was studied using a technique called dichotic listening. One message was played to one ear and a different message in a different voice was played to the other ear concurrently. The task was to shadow, or repeat, everything presented to one ear. From time to time the tape was turned off and the subject was asked to repeat any information presented in the other ear, the ignored message. Subjects were stumped. They often could remember a few words from the end of the ignored message, but that was all. This led to a distinction between attended and unattended aspects of immediate memory. So, for example, you take in stimuli from all sensory modalities and some of the sensations outlast their stimuli. You can practically hear the last notes of a symphony for some seconds after they have been played. Following a flash of lightning on a dark night, your memory of the sight of twisting trees lingers on. Usually, though, there is such a wealth of stimulation that you cannot attend to all of it. A few sensations are selected for further processing, memorization, and understanding. They might be the actual words and ideas from a conversation, or a single canvas in an art museum.

Broadbent's theory was essentially that stimuli that are not relevant to one's current concerns are filtered out of attention and awareness. It is relatively easy to filter out nondesired stimuli on the basis of physical cues, as when one ignores a certain person's voice or picks only red berries, not green ones. It is more effortful, but still possible,

to filter out nondesired stimuli on the basis of their meanings, as in a room of chatter when one ignores all conversations that are not related to particular subject, or when one goes into the garage to pick up several tools that are needed for a particular carpentry job while leaving others behind. In the latter type of selection, any kind of mental preoccupation can easily result in a failure of selection. You may then tend to hear a loud but irrelevant conversation, or you may pick up the wrong tools.

At one point early in the history of cognitive psychology, it was proposed that people actually take all stimuli into the mind for further processing, and are limited only in the ability to respond to so much at once. That late-filter theory was based partly on a further dichotic listening study by Neville Moray in 1959, in which the subject's own name was presented in the ignored message. It was found that some people notice their name, suggesting that the message was not actually filtered out. This finding was often highlighted in textbooks for many years, albeit without further support. (Instead, the field seemed to focus on visual stimuli.) Much later studies in the 1990s did confirm that this registration of one's own name sometimes occurs.

A study by Andrew Conway and his colleagues in 2001 discussed Moray's result and suggested that this type of finding, noticing one's name in an ignored message, does not actually support the late-filter theory. Another explanation is that Broadbent was essentially correct in suggesting an early attentional filter, but that some people fail to maintain a strong task set so that the information actually attended does not reliably match the information that was supposed to be attended. Conway and colleagues tested subjects on a complex immediate-memory test called operation span, in which arithmetic problems were to be solved along with memorization of a word presented after each problem. All of the words were to be recalled after the last arithmetic problem was solved. The operation span test, like certain other complex working memory span tests, correlates with intelligence rather well, and it correlates with the ability to pay attention. For example, recent research by Michael Kane shows that individuals with low spans experience more involuntary mind-wandering when they are trying to pay attention than do high-span individuals. If dichotic listening really does involve selective filtering and subjects' attention sometimes wanders, it could wander off of the message that was supposed to be attended and onto the message that was supposed to be ignored, or it could be split between the two messages. That should happen more frequently among low-span individuals. Consistent with this interpretation, Conway and colleagues found that 65% of the subjects in the lowest quartile of operation span noticed their names, whereas only 20% of the subjects in the highest quartile noticed. So Moray's result does not strongly challenge Broadbent's filter theory after all. The results for noticing one's name are consistent with an early filter theory.

Johnston and Heinz, in 1978, used a selective listening task to show that people can pay attention either to sensory features, as in the early filter theory, or to conceptual features, as in the late-filter theory. However, it takes much more effort to pay attention according to the late-filter theory. In this experiment, there was a secondary task, reaction time when a light was presented. Reaction time to the light was about the same when the task was carried out alone or along with selective listening based on a sensory feature, voice quality. During selective listening to one of two messages about different topics spoken in the same voice, however, reaction time to the light was much slower. This indicated that it takes effort to pay attention to conceptual features, whereas paying attention to physical features is easy and natural as in the early filter theory.

A Schematic Model of the Information Processing System

Using the information that has been presented so far, you can begin to form a concept of how the brain's information processing system acts. One way to conceive of the system is shown in Figure 1, from Nelson Cowan's writings. Incoming stimuli all make contact with elements in long-term memory that become activated, and these activated elements of memory include sensory memory. However, not all activated elements enter the focus of one's attention, just as the entire room full of

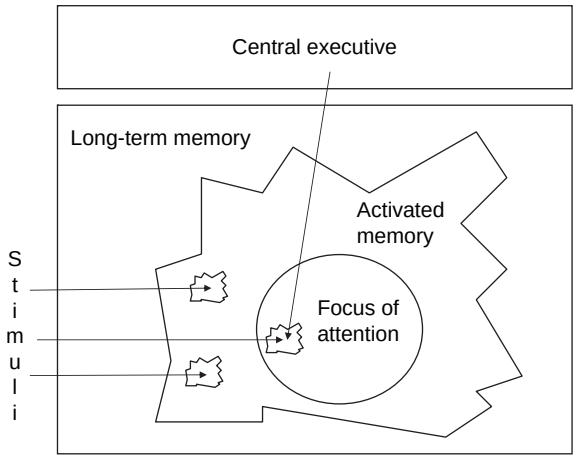


Figure 1 A concept of the immediate-memory system, which includes both sensory and conceptual features. Activated memory and the focus of attention are embedded processes that both contribute to immediate memory. For further support see Cowan's 1995 and 2005 books.

chatter mentioned above, although it can be heard, cannot all be comprehended. The Central Executive represents mental processes that you use to impose your own will in terms of what information to focus your attention on and what to do with that attended information, as when you decide to concentrate on understanding one speaker while ignoring other conversations. However, other factors help control the focus of your attention, such as abrupt noises, sudden changes in the lighting, and attractive displays on television. These can either help your central executive, as when a speaker uses dramatic changes in voice intonation to get across a point, or can hinder your central executive, as when you try to listen to a monotonous speaker during a thunder storm. The information that is attended receives further processing in any case, resulting in a collection of concepts that accompany the sensory images in one's conscious mind and mold one's interpretation of these images. (Both sensations and concepts can count as activated elements from long-term memory.) The tendency for the central executive to be able to control which sensory stimulus channels to attend and which to ignore is roughly equivalent to Broadbent's filter. In the illustration, paying attention on a sensory or perceptual basis (as in picking only the red berries) is like staying on one stimulus input arrow, whereas paying attention on a conceptual basis (as in finding which tools one needs) is like hopping from one arrow to another.

Not everything in immediate memory can be an activated set of representations from long-term memory, of course. New information is represented by a new set of links between already-existing features of some sort. For example, if you saw a blue apple, it would probably be the first link of the blue feature and the shape associated with apples. You might also receive a new name for it. These new links between a shape, a color, and a combination of phonemes would coexist in the focus of attention and would produce a new long-term memory representation.

One reason to use a schematic diagram like this one is that the brain is complicated. Any one of these components included in Figure 1 could correspond to a twisted tangle of tree-like neural structures distributed widely throughout the brain. The schematic model is at a higher level of analysis. By analogy, one can delineate the executive, legislative, and judicial branches of the US government when, in fact, the three branches of government are represented by complex and overlapping webs of people, buildings, and regulations. Nevertheless, there are known neural correlates of these elements of mind. The central executive processes rely a great deal on frontal lobe structures (behind the forehead). The parietal areas, more toward the upper middle of the head, are very much involved in the focus of attention; damage to those areas can result in patients who are unaware of one part of space (unilateral neglect) or are unaware that they have a serious paralysis or other disability (anosognosia). Recent brain research shows that parts of the parietal lobes are especially active when a challenging amount of information is being saved in immediate-memory tasks. The temporal lobe areas (behind the sides of the head) include the hippocampus and other neighboring areas that are especially important in retaining information and forming new long-term memories, and diverse areas represent different kinds of memories that become electrically active when the ideas that are represented are activated by stimulation or thought processes.

The characteristics of sensory and immediate memory that make them particularly interesting stem partly from their limits. The discussion of

what these kinds of memory can do is intricately related to what they cannot do. In the remainder of this article, two types of limits will be discussed. The first is time limits in activated memory and the second is capacity limits in the focus of attention.

Time Limits of Activated Memory

It is possible to demonstrate directly that there are limits in the activation of information from long-term memory, and this temporarily activated memory forms the basis of the human information processing scheme shown in Figure 1. These are time-related limits, although the exact limit seems to depend on the nature of the stimulation. Some concepts related to human awareness may be based on this temporarily activated memory. These will be discussed first, and then more direct tests of the time limits of activation will be discussed.

There are two concepts related to time limits and awareness that are of special interest, called the perceptual moment (also sometimes called the psychological moment) and the psychological present. The perceptual moment is the small amount of time in which two successive events seem simultaneous, which happens only when they are within about 200 ms of one another. As an example, if you move your finger back and forth as rapidly as you can, it will visually appear as if your finger is in two places at once (as well as being smeared in between these two places). Why does this occur? One way of explaining it is that the neural processes representing each visual viewpoint outlast the actual stimulus and are still ongoing when the next visual viewpoint arrives. This, in fact, is how movies are perceived. Another way of explaining it is to say that there is a visual sensory memory or afterimage of the stimulus that seems like a continuation of that stimulus after it is gone, and that the two afterimages are combined in our awareness into a single stimulus. If there were more time between the stimuli, one afterimage would fade before the next one was formed and the two viewpoints would not appear simultaneous. A great deal of research suggests that this neural explanation and this psychological explanation are fundamentally compatible and that both are apt. The perceptual moment depends on the type of stimuli that are presented but typically ends within a couple hundred milliseconds.

In contrast to the fleeting perceptual moment, the psychological present is a longer period of time of a second or so, during which a single event seems to take place. The rapid beats of a drum roll may be grouped together to form a single event, whereas slower beats of a drum may be perceived as separate events. Whereas, for the perceptual moment to be the same for two stimuli, ongoing perceptual processing must overlap, the psychological present has a looser criterion; the second stimulus need only be presented while the first is still vividly recalled. It is as if there are two phases of temporarily activated sensory features in memory: a vivid afterimage that seems as if the stimulus is containing, lasting for several hundred milliseconds, and then a vivid recollection of the sensory events, for a second or so. Like the latter, William James in the late 1800s used to think of primary memory (in this chapter, immediate memory) as the trailing edge of the conscious present.

Many studies could be used to demonstrate these concepts. We will focus on a couple of them that seem especially thought-provoking. In 1968, D.A. Allport carried out a study to distinguish between two varieties of the perceptual moment hypothesis. In the discrete moment hypothesis, the stimuli are accumulated into one moment until some fixed amount of time; then subsequent stimuli are accumulated into another moment until a similar fixed amount of time; and so on, just as each sheet of paper in an office could be stamped with a date and multiple sheets that arrived at different times would share the same date stamp. In the continuous moment hypothesis, though, events would be viewed as if through a sliding window (i.e., sliding in time). First the window might integrate events 1, 2, and 3; then the oldest event 1 would fade from the window, to be replaced by a new event 4; and so on.

To distinguish between these hypotheses, Allport developed a simple but ingenious test using an oscilloscopic display, as shown in Figure 2. (Actually, the display included 12 lines but the point can be made more simply using four lines as shown here.) The lines would be presented one

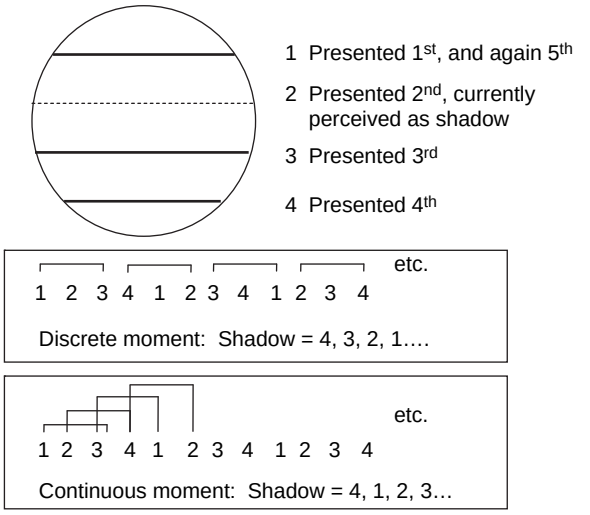


Figure 2 Allport's experiment to distinguish between discrete and continuous varieties of the perceptual moment. The top part of the figure illustrates a particular moment of perception in the situation in which Lines 1–4 were presented one at a time, in order, repeatedly at a fast rate. Line 2 happens to be perceived as a shadow at the moment depicted here, but the perceived shadow location moves as the display progresses. The proposed mechanism is illustrated according to the discrete moment hypothesis in the middle of the figure (with brackets showing which lines four consecutive moments would include), and according to the continuous moment hypothesis at the bottom of the figure (with brackets showing which lines four overlapping, temporally sliding moments would include).

at a time in the repeating order 1, 2, 3, 4, 1, 2, 3, 4. . ., and so on with no break. The subjects were to adjust the rate of line presentation until what was perceived was all but one of the lines at the same time. (If the rate were any faster, all of the lines would be perceived at once or, if the rate were any slower, fewer lines would be perceived at once.) With the rate so adjusted, the line that was not perceived at any one moment was called a shadow and the change over time in which line was seen as a shadow was called shadow movement. Then different predictions about shadow movement could be made for the two types of perceptual moment.

For the discrete moment, as shown in the middle of the figure, the assumption was that the succession of lines was parceled out into nonoverlapping moments of a certain duration. That process would result in shadow movement in the opposite direction from the actual movement of lines. If the first discrete moment were long enough to include Lines 1, 2, and 3, then 4 would be the shadow. The next discrete perceptual moment would open up just in time to capture Line 4, and also the next presentation of Lines 1 and 2, but then it would close before it could include the next presentation of Line 3, which would go in the next window instead. Thus, the shadow has moved from Line 4 to Line 3 and continues in that manner, in a direction opposite from that of the line presentations. In contrast, in the continuous moment hypothesis, shown in the bottom of the figure, the shadow movement would go in the same direction in which the lines were presented because the temporally sliding window always includes the most recently presented three lines and leaves out the least recently presented line. The evidence from the experiment unequivocally supported the continuous moment hypothesis, in which our events are grouped together according to a sliding window of perception. This sliding window can be viewed as sensory memory, in which the least recent line always was the first to fade from view.

Albert Bregman has carried out a line of research on a phenomenon that can be seen as demonstrating the power of the psychological present, called auditory stream segregation. Suppose you hear a simple series of two tones in alternation: high–low–high–low in pitch, and so on. How will the series sound? That depends on various factors, but one of these is the time between tones, as shown in Figure 3. When the series is presented slowly, the mind links together adjacent tones into a single, coherent perceptual stream going up and down. When the series is presented more quickly, it is the tones of the same pitch that seem to be grouped together, into a high-pitched stream and a second, low-pitched stream. In fact, it is difficult to hear exactly when the high and low pitches were presented relative to one another; there is no clear perceptual organization across streams. Tones in each stream share the grouping principle of similarity. When the series is considerably slowed down, however, what may be different is that successive tones of the same pitch are no longer within the same psychological present, and so they cannot be grouped together. In terms of memory, it may be that the vividness of one tone has faded by the time

a similar tone is presented, so that they no longer seem to belong together in time.

Time-dependent activated memory does not rest primarily on notions of the perceptual moment and the psychological present, but has repeatedly been examined more directly through a wide variety of experimental procedures. One of these, by George Sperling, was published in 1960. He carried out a number of experiments. A typical one is illustrated in Figure 4.

The first panel of the figure shows an array of 12 letters that was presented very briefly, with different letters on every trial. In one kind of trial (not the kind shown in the figure), the task was simply to recall all 12 letters by writing them down. This was called the whole report procedure, and it showed that practiced adults could recall only about 4 of the 12 letters in their correct locations. Sperling realized, though, that he did not know the basis of this limit. It could be that more items than this were held in an activated form of memory but that they disappeared from memory before they could all be written down. What Sperling did to address this problem was to develop the partial report procedure, shown in Figure 4. Subjects learned that they should respond in a manner dictated by a tone presented after the array. If the tone was high, the task was to recall the top row; medium, the middle row; or low, the bottom row. Because the tone was not presented until after the array, when the array was presented the subject had to retain all three rows. However, after the tone was presented, the task became simply to recall one row, which did not exceed the number of items that subjects could recall in the whole report situation. The result would indicate how much information about the cued row was held in memory at the time that the cue arrived. Given that the other rows had to be held at the same time, the total amount of information available in memory when the tone cue arrived could be calculated by multiplying by 3 the average number of items recalled in partial report.

The outcome is shown in Figure 5. It turned out that the result depended on the amount of time between the array and the tone cue. If the tone was presented very soon after the array, almost all items in the array were still available for recall. However, if the tone cue was delayed for a quarter second (250 ms), it was of almost no value. The number of items available for recall was then no greater than the whole-report limit of about four items. This suggested that all of the items in the array were held in memory at first, but that they faded from memory within about 250 ms. Other experiments showed that the cue had to indicate a physical feature, such as the row of the array to

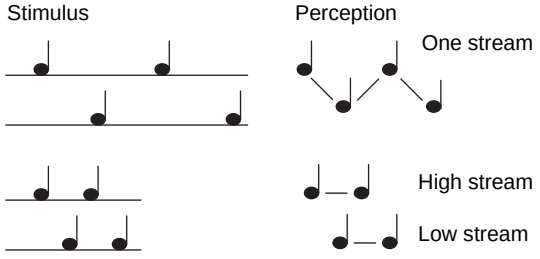


Figure 3 The phenomenon of auditory stream segregation, as in Bregman's work.

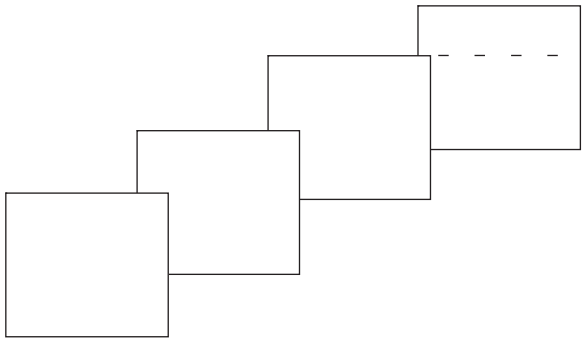


Figure 4 Sperling's partial report procedure.

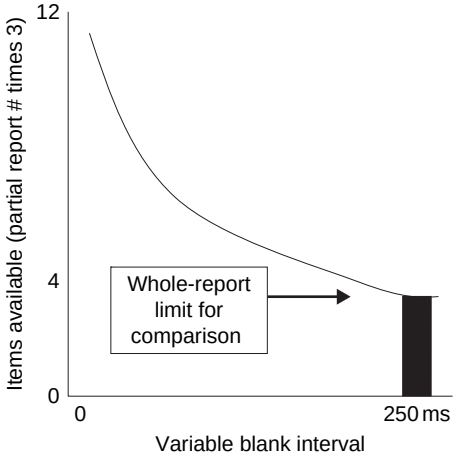


Figure 5 Results of Sperling's whole and partial report procedures.

recall, and not a more abstract, conceptual feature, such as whether to recall letters or digits from an array that included both types of symbols. This suggests an explanation that is entirely consistent with Donald Broadbent's explanation of the dichotic listening results. A large amount of sensory information is held for a brief time, but only a small subset of that sensory information can be shifted to the focus of attention before it is lost. It can be selected for attention and further processing on the basis of physical cues (such as voice quality or direction, or written letter color or location) more easily than it can be selected on the basis of conceptual cues (such as the topic of speech, or the category of the written symbols).

One possible discrepancy between auditory and visual sensory memory in these studies is that the duration of auditory sensory memory appeared to be several seconds, whereas the visual sensory memory only appeared to last a fraction of a second. In a review that Cowan published in 1988 and discussed further in a 1995 book, an alternative interpretation of the literature on sensory memory was suggested. It appeared that, all information considered, there are two separate phases of sensory memory: a vivid mental afterimage experienced while perceptual processing of the items continued for about 250 ms, and a longer-lasting, vivid sensory recollection for several seconds. There has not been a great deal of discussion of this issue (i.e., whether there are sensory-modality-specific memory durations or, instead, two phases of sensory memory in all modalities), but it remains one on which researchers do not seem to agree. The two phases of sensory memory discussed in the 1995 book, if they do exist, would appear to underlie the perceptual moment and the psychological present, respectively, as we have discussed.

These experiments leave open the question of whether nonsensory, conceptual information that is activated also has a limited time period of persistence. This is not an easy issue to examine, because sensory information ordinarily may have to be drawn into the focus of attention before conceptual features can be perceived. Then it becomes possible to rehearse the information by repeating it covertly (mentally and silently) or simply by continuing to attend to it. One can prevent this rehearsal by presenting distracting tasks, but these can cause interference as well. In 1973, Michael Watkins and his colleagues presented lists of words followed by tones that either could be ignored or, in a different type of trial, had to be identified by button press, and found that memory was lost over time – especially during the first 3 s of a 20-s period – only when the tone information had to be identified. On the other hand, in 2004, Stephan Lewandowsky and his colleagues took a different approach and found that memory for letters was not lost over time when the words had to be recalled either slow or fast, regardless of whether subjects had to recite a word over and over to prevent covert verbal rehearsal. Some details of both studies, and of other studies in this literature, seem to be such that the issue of the loss of conceptual activated memory features over time remains unresolved.

Even if a loss of activated memory over time is observed, it has to be asked whether it is the amount of time itself that is important, or the way in which that time is perceived relative to recent events. This is clear in a study of brain function based on electrical signals at the scalp, or event-related potentials, by István Winker and his colleagues in 2001. They presented series of tones that remained identical for six tones, with the seventh tone sometimes shorter than the others. Even though the subjects are busy reading during this kind of procedure, so that the tones are ignored, subjects' brains still respond to tone changes if the unusual tone (the deviant) is presented shortly after the other tones (the standards). This mismatch negativity response is said to occur because the brain saves a sensory memory of the standard tones and compares each new tone to that standard representation. Figure 6 shows three different stimulus arrangements and indicates the results. When all tones in the series were a half second apart, all subjects' brains responded to the change in tone duration. However, when the standard tones were one half second apart and the deviant did not occur until 7 s after the last standard, some subjects' brains still responded, whereas others' brains did not. For the subjects whose brains did not respond, one might think that a memory of the standards was lost by 7 s. However, another interpretation is that their brains no longer considered those standard tones to serve as a fair, current

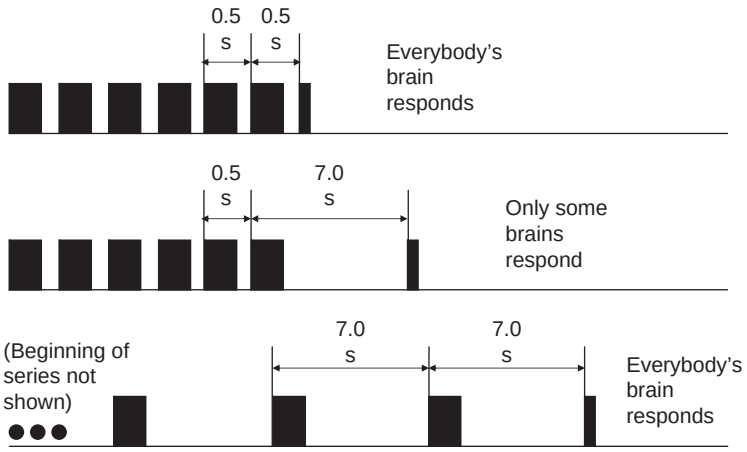


Figure 6 The procedure that Winkler and colleagues used to examine why tone memory declines over time.

comparison. After all, the standards in each series formed a tight group in time and the deviant was not part of that group. Verifying that interpretation, all subjects' brains responded when all tones were 7 s apart, as in the third row of the figure.

To summarize this discussion of temporarily activated memory, it can include sensory or conceptual features from long-term memory that have been activated by recent events or thoughts. The conceptual features are much more likely to be activated only after information is drawn into the focus of attention, but then the focus of attention might shift elsewhere and leave behind activated conceptual features that are unattended. For sensory and conceptual activated features, it is clear that there is a time limit to activation but, after all these years of research, it is not yet clear what kind of time limit that is. All of the currently available information might be compatible with the notion that the critical time interval may be derived by the brain relative to the pace of events. At least, that may be the case for the longer form of activated memory that we have linked to the psychological present. It is usually within the range of a few seconds but it is hard to pin down exactly.

Within that possibly flexible duration of the psychological present, there appears to be a shorter duration of about 250 ms that may be determined by the time it takes for perceptual processing of a single stimulus to be completed, which therefore may be a more precisely fixed duration. It corresponds to a vivid sensory afterimage in the brain and to the perceptual moment.

Capacity Limits of the Focus of Attention

One of the functions of the focus of attention is to overcome time limits. Suppose you want to check the oven in 3 min. Without an external timer, you have no choice but to invest a considerable amount of effort attending to the oven so that your mind does not wander on to other things. The fundamental reason why this is effortful is that there are competing stimuli and competing thoughts, and not all of them can be attended at once. That is, there is a capacity limit to the focus of attention.

One dramatic result of the capacity limit is what Dan Simons and others have called inattentional blindness. You can demonstrate it by playing a trick on a friend in the office. Wait until they turn around for a moment and then remove a small item from plain sight on their desk, such as a stapler. If you do this discretely chances are that, when your friend returns attention to your direction a moment later, he or she will not notice the missing object. Provided that the item was not recently attended, despite being well within the perceptual field of view, its disappearance will not attract notice. Simons and his colleagues demonstrated inattentional blindness dramatically by having one experimenter stop a pedestrian to ask directions. In a staged event, two other confederates of the experimenter holding a door in a horizontal position walked between the experimenter and the hapless pedestrian who was asked directions, obscuring their view of each other. Secretly, on the side of the

door across from the pedestrian, the experimenter grabbed the door and was replaced by one of the confederates who had been holding the door. When the two people with the door moved along out of the way, the confederate acted as if he was the original experimenter. In many cases, the pedestrian did not notice that there had been a change in the person to whom he was giving directions before the interruption, and continued to give directions. He had been attending to his directions, and not primarily to the exact features of the person receiving those instructions.

Sperling's partial report procedure shown in Figure 4 could be considered an inattentional blindness procedure because attention is insufficient to encode into the focus of attention all of the items from the array held in visual sensory memory. In 1997, Steve Luck and Ed Vogel simplified that procedure so that it would not require that multiple items be recalled, in order to get the most sensitive measure of what items are in working memory. These might be thought of as the items encoded into the focus of attention. A version of that procedure used recently by Jeffrey Rouder and his colleagues does an excellent job of estimating items conceptually held in working memory. The procedure is illustrated in Figure 7. An array of small, colored squares is briefly presented. (Each pattern in the figure represents a different color.) An interval of about 1 s then elapses, giving the subject plenty of time to encode the array into attention. Then a masking display is presented with a set of multicolored squares in the same location as the colored squares in the previous array. This serves to eliminate any remaining sensory memory. What remains are any conceptual aspects of working memory that the subject has retained even though the sensory memory features have been overwritten by the mask. The masking display also serves to remind the subject just where the squares were on the screen. This is followed by a single colored square, and the task is to indicate whether this square is the same color as the square in that location within the original array. If it is not, it is a color that was not found in the array. The probe remains until an answer is given.

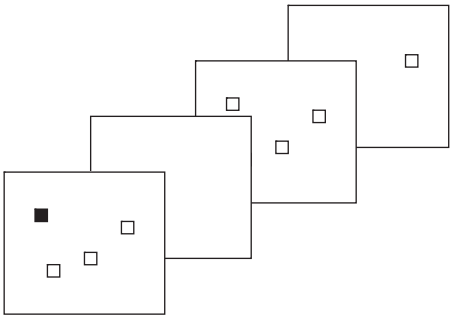


Figure 7 A version of the visual working memory procedure by Steven Luck and Edward Vogel.

This procedure produced results that conformed very well to a simple mathematical model to calculate capacity. The model included an assumption that on a certain percentage of trials, the subject was not paying attention at some point. That happened on 12% of the trials according to the model. Each subject was said to have a certain capacity k and, when he or she paid attention, k items were retained in working memory. When the array item in the same location as the probe was in working memory, the subject was assumed to know whether there was a color change or not. If the item was not in working memory, either because of inattention or because capacity was exceeded, the subject was assumed to guess. In different trial blocks, different proportions of trials involved color changes to manipulate the guessing rate. The estimate of k stayed fixed across guessing rates and across different numbers of items in the array, as it should according to the mathematical model. The mean k was 3.35 items, consistent with previous estimates of working memory capacity such as those reviewed in Cowan's 2005 book.

Scott Saults and Cowan recently showed that this kind of working memory examined by Rouder and his colleagues is not limited to visual information. They set up four loudspeakers surrounding the subject and played four different spoken digits at the same time, each in a different voice, while also presenting an array of four or eight colored squares on the computer screen. The subject was responsible sometimes for the visual information only, sometimes for the auditory information only, and sometimes for both modalities at once. Provided that a postperceptual mask was presented to eliminate sensory memory in both modalities, the visual and auditory information traded off. In particular, subjects remembered about 3.5 visual items when they did not have to retain auditory items also, and

they remembered about 3.5 items total when they had to remember the auditory along with the visual. The verbal spoken items displaced some of the nonverbal visual items in working memory, which seems to be evidence that the information being stored in a capacity-limited manner is conceptual and abstract rather than specific to a sensory modality.

The memory in these procedures we have been discussing seems less than what was discussed in a famous article by George Miller, who, in 1956, suggested that there is a "magical number seven plus or minus two," describing how many items people can recall. This estimate is roughly true, but it does not seem to reflect the most basic type of immediate memory. Instead, it seems to reflect a process whereby the items in working memory are grouped together to make larger items. (That grouping process was in fact one of Miller's main points.) When one reads the telephone number 356–4129, one tries to form a group of three items followed by a group of four. When the groups are formed, each group may take up only one slot in working memory. It is these smaller groups that may reflect the basic capacity limit, with the larger limit occurring as several smaller groups are entered into working memory together. Anders Ericsson and his colleagues have made the point that, with enough practice, at least some individuals manage to raise their memory span from seven items to the astonishing level of eighty or more items, but only for the type of material being practiced (such as digits). They do it by learning to form large chunks based on knowledge they already have, such as memorized athletic records. More precisely, chunks of up to about four items are formed and then these chunks are combined to form even larger, higher-level chunks.

The capacity of working memory increases in childhood and then decreases again in old age, as Cowan, Moshe Naveh-Benjamin, and their colleagues have demonstrated recently. Also, as Randall Engle and his colleagues have emphasized, tests of working memory capacity do a good job of discriminating between young adults with more versus less aptitude on complex cognitive tasks and intelligence tests. Yet, we do not understand what it is that distinguishes someone with a high span from someone with a lower span. One theory is that individuals with a high span simply have more slots in working memory than do lower-span individuals. However, an alternative theory is that individuals with a high span are able to pay attention better, and therefore are more likely to fill their working memory capacity with relevant items as opposed to thoughts that are irrelevant to the task that has been assigned.

Jim Gold and his colleagues found evidence for the slots theory in distinguishing schizophrenic individuals from normal controls. In one experiment, for example, they presented arrays with colored items of two shapes and usually tested subjects on one of these shapes. Occasionally, they would test subjects on the other shape. In this way they could tell how well individuals were filtering out the usually irrelevant shape (by examining the advantage in performance on the usually tested shape over the other shape) and they could tell how much individuals encoded into working memory (by examining the total of capacity k for the two shapes added together). Surprisingly, the schizophrenic individuals filtered just fine but were below normal on total capacity. In contrast, Vogel and his colleagues have done similar testing using event-related brain potentials and have argued that high- and low-span normal individuals differ primarily in the ability to filter out the irrelevant information. We know of areas in the brain that specialize in holding working memory information (certain areas in the parietal lobes) and areas in the brain that specialize in filtering out irrelevant information (the basal ganglia). It stands to reason, then, that the available slots in working memory and the ability to fill these slots appropriately and efficiently are two important, related, but nonidentical aspects of individual and group differences.

Conclusion

Immediate memory, or its close cousin working memory, is critical for human thought and awareness of the environment. It is, as William James said, the trailing edge of the conscious present. The conceptualization of this memory as shown in Figure 1 has proved to be useful in understanding the mind. Activated features of memory, within which sensory memory usually predominates, are collected by our brains, apparently for all stimuli, without the need to pay attention. Even though the

information that is activated remains so only briefly, it provides a rich backdrop from which information can be drawn selectively into the focus of attention for further processing in a timely manner. Conceptual features that come out of this selective processing can be added back into the pool of activated memory. There are some key questions that remain unanswered regarding the principles that produce the limits in activated memory and the focus of attention, but the scientific endeavor is briskly marching toward some tentative answers.

Acknowledgments

This work was supported by NIH Grant R01 HD-21338.

See also: Visual Experience and Immediate Memory.

Suggested Readings

Allport DA (1968) Phenomenal simultaneity and the perceptual moment hypothesis. British Journal of Psychology 59: 395–406.

Bregman AS (1990) Auditory Scene Analysis: The Perceptual Organization of Sound. Cambrdige, MA: MIT Press.

Broadbent DE (1958) Perception and Communication. New York: Pergamon Press.

Cowan N (2005) Working Memory Capacity. Hove, East Sussex, UK: Psychology Press.

Engle RW, Tuholski SW, Laughlin JE, and Conway ARA (1999) Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. Journal of Experimental Psychology: General 128: 309–331.

Ericsson KA, Delaney PF, Weaver G, and Mahadevan R (2004) Uncovering the structure of a memorist's superior ''basic'' memory capacity. Cognitive Psychology 49: 191–237.

Gold JM, Fuller RL, Robinson BM, McMahon RP, Braun EL, and Luck SJ (2006) Intact attentional control of working memory encoding in schizophrenia. Journal of Abnormal Psychology 115: 658–673.

Johnston WA and Heinz SP (1978) Flexibility and capacity demands of attention. Journal of Experimental Psychology: General 107: 420–435.

Kane MJ, Brown LH, McVay JC, Silvia PJ, Myin-Germeys I, and Kwapil TR (2007) For whom the mind wanders, and when: An experience-sampling study of working memory and executive control in daily life. Psychological Science 18: 614–621.

Lewandowsky S, Duncan M, and Brown GDA (2004) Time does not cause forgetting in short-term serial recall. Psychonomic Bulletin & Review 11: 771–790.

Luck SJ and Vogel EK (1997) The capacity of visual working memory for features and conjunctions. Nature 390: 279–281.

McCollough AW and Vogel EK (2008) Your inner spam filter. Scientific American Mind, June/July: 32–35.

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63: 81–97.

Simons DJ and Levin DT (1998) Failure to detect changes to people during a real-world interaction. Psychonomic Bulletin & Review 5: 644–649.

Winkler I, Schröger E, and Cowan N (2001) The role of large-scale memory organization in the mismatch negativity event-related brain potential. Journal of Cognitive Neuroscience 13: 59–71.

Biographical Sketch

Nelson Cowan (PhD, 1980, University of Wisconsin) is a Curators' Professor at the University of Missouri. He has been interested in the brain and mental processes every since he was in high school, and this interest developed further in college at the University of Michigan. His research, funded by the National Institutes of Health since 1984, focuses on working memory, selective attention,

and the childhood development of these processes. He is the author of *Attention and Memory: An Integrated Framework* (1995, Oxford University Press) and *Working Memory Capacity* (2005, Psychology Press). He has served as an associate editor of three journals, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *Quarterly Journal of Experimental Psychology*, and *European Journal of Cognitive Psychology*. As of September 2008, he is on the governing board of the Psychonomic Society and is the president of the experimental psychology division of the American Psychological Association. Nelson is married and has three grown children; his wife, Jean Ispa, is a professor of human development and family studies.

instance, selected as a mate or preferred over others in cooperative alliances.

The problem with the CST is that the theory does not specify how exactly behavior is affected by REM sleep and dreaming. The theory does not define what the external behavioral cues are, supposedly modulated by preceding REM sleep and dreaming, which the other group members can monitor and evaluate. Unlike typical costly features that play a role in mate selection (as the peacock's tail), the costly features of REM sleep and dreaming might not be as such directly observable. However, some directly observable characteristics must be highly and reliably correlated with REM sleep and dreaming so that by preferring these features, the costly features of REM sleep and dreaming would consequently be selected for, albeit indirectly. Furthermore, we should be able to verify that other group members actually find these behavioral cues desirable and that the cues affect favorably the selection of the individual as mate or cooperative party.

The psychological problem solving and creativity function

Ernest Hartmann, who has suggested a psychological healing function for dreams, has also proposed that the capacity of dreaming to form new connections within the neural networks of the brain might have served two functions useful for our ancestors. Hartmann's first evolutionary argument concerns the idea that dreaming, as a process making new connections, has a kind of psychological problem solving or psychotherapeutic function, especially in handling traumatic experiences:

> Dreams after trauma may appear to represent a rare situation and fortunately for many of is this is so. However…. one hundred thousand years or so ago, when the human brain was gradually developing to its present form, our lives were considerably more traumatic; the after-effects of trauma may well have been an everyday reality and the resolving of trauma a constant necessity.… (Hartmann, 1998: 158).

Hartmann's second evolutionary proposal is that the making of broader and wider associations during dreaming might have helped in bringing material together in new ways and this would have been useful to our ancestors in their waking lives. Dreams might have provided new innovations in exploiting resources and solving problems related to everyday life. This particular argument emphasizes the creative and problem-solving nature of dreams.

> . . .In other words, the functions of dreaming. . . may have been especially important for us at earlier times in our species' development. . .Only our dreams gave us a chance to do this – to make broader and wider connections, to integrate trauma or other new material, and also to bring material together in new ways that occasionally might have been useful to us in our waking lives. (Hartmann, 1998: 209)

In sum, Hartmann suggests, at least indirectly, that the evolutionary origin of the function(s) of dreaming is in the ancestral environment where life was dangerous, trauma resolution was required on a daily basis, and creative new ideas could have provided valuable selective advantage. Individuals who were able to regain emotional balance and well-being after trauma were better off than those who did not, and therefore the psychotherapeutic function of dreaming was selected for and became a universal feature of the human mind. Similarly, individuals with creative problem-solving dreams were better adapted to their environment and consequently left more offspring than individuals not having extra help from their dreams.

We have already dealt with the problems related to the psychological healing theories. In addition, empirical evidence on the creative and problem-solving nature of dreams is mostly negative. New and useful solutions to waking life problems are extremely rarely depicted in dreams. Thus, whatever the function of dreaming is, it does not seem to be the finding of creative, new, and useful solutions to problems faced in waking life. Therefore, these two suggestions for the evolutionary function of dreaming are not supported by the currently available empirical evidence.

Simulation functions of dreams

Three different simulation function ideas have been suggested: that dreams are similar to play behaviors in mammals, dreams are simulations of social interactions, and that dreams are specialized in the simulation of threatening events. All these views are based on the claim that the dream experience is

functionally constructed to resemble waking experiences, and therefore shows clear design features for a world simulation function.

Dreaming as play

There are indeed several similarities between dreaming and play behavior, but also a number of dissimilarities. Both seem to be limited, in their clearest forms, to mammals only. Both can simulate reality and rehearse different types of situations and interactions in a safe context. Both may exaggerate, transform, and display enormous variation of behaviors that are originally related to other contexts outside play. Both are energetically costly, biologically programmed behaviors that should therefore be in some way useful.

Unfortunately the adaptive functions of play are not entirely clear; thus, it is not possible to explain the functions of dreaming by saying that they are similar to the functions of play behavior. Most likely play has multiple functions, some of which might be closely related to the threat simulation function of dreams. There are forms of mammalian play that seem to be rehearsals of hunting behavior, aggressive encounters, or predator avoidance. Thus, playing and dreaming might have complementary functions in the rehearsal of behaviors: dreaming is perceptually more realistic than play, whereas play is motorically more realistic, involving actual execution of motor programs, muscular movements, and physical exhaustion.

Dreaming as social simulation

Another currently popular suggestion is that dreaming simulates human social interactions and rehearses social perception and social skills. Several slightly different versions of the 'social simulation hypothesis' have been proposed. The 'social mapping hypothesis' suggests that dreaming allows simulation of self, location, and awareness of others, including awareness of their internal mental states. Dreaming is thus suggested to have rehearsed the perceptual and emotional features required in successful social mapping in human evolutionary history, eventually leading to the emergence of self-awareness. Relatedly, David Kahn and J. Allan Hobson emphasize that awareness of what others are thinking and feeling is a robust aspect of human consciousness, and this aspect is maintained during dreaming despite the changes in chemistry and activation patterns of the brain during sleep. Thus, even though Kahn and Hobson do not explicitly express it, they imply that awareness of the minds of others ('theory of mind') during dreaming might have contributed to the ability to anticipate the intentions of others while awake. Dreaming about the intentions of others could prepare us for social encounters when awake. This idea leads us to another version of the social simulation hypothesis, suggested by several different researchers.

As many of the selection pressures faced by ancestral humans were posed by complex human social life, modeling human relationships, and interpersonal bonds, for example, family politics, attachment, love affairs, and status battles, might have had adaptive value. Interacting with other members of the group was an important selection pressure in the ancestral environment, and simulation of skills such as how to find a mate, build coalitions, and avoid conflict would have been useful. In dreams, it is possible to practice dealing with complex social situations, and because those most adept in their social environment were likely to have the best access to resources in their social group, simulation of social situations would have been selected for. Furthermore, strong family and group cohesion would have enabled organized defenses against predators and other enemies and enhanced survival and health of group members.

The social simulation hypothesis is to some extent compatible with what we know about the form and content of dreams. About half of the human characters in our dreams are persons familiar to us, and appearance, behavior exhibited by the character, and feelings evoked by the character in the dreamer are regularly used in the identification of the person. More than 80% of known dream characters evoke some kind of emotional response in the dreamer, most often affection or joy. A significant amount of time in dreams is spent wondering what other dream characters are thinking or planning. Thus, our dreams often represent human characters and give plenty of space for opportunities to practice social interactions. Nevertheless, aggression is a more common type of social interaction than friendliness, while sexual interactions are relatively rare in dreams. Thus, we get

less practice in forming positive social bonds, such as making friends and allies, than in dealing with negatively toned social interactions. Even less time in dreams is devoted to mate selection and practicing how to form romantic relationships.

Although the social simulation hypothesis is consistent with the fact that other people and multiple social interactions are frequently present in our dreams, the hypothesis has some problems as an evolutionary psychological account of the function of dreaming. First, we get a lot of practice in (nonthreatening) social interactions during our waking lives, and this practice does not have high costs. Thus, it remains unclear why it would be advantageous to practice or simulate something like that further in our dreams. Second, there is a lack of studies on the detailed nature of the social interactions in dreams, and the ones conducted reveal the often aggressive nature of social encounters. Do we, in fact, interact with other dream characters in reasonable ways that might be considered useful simulations of or rehearsals for real-life social interactions? To back up an evolutionary hypothesis, a detailed description of the type of dream content and the conditions under which it occurs is required, as well as a cost–benefit analysis that should show why the dream simulation is likely to be useful for us (or was likely to be useful for our ancestors). It remains open whether the social simulation hypothesis will receive support from the more detailed analyses of dream interactions and cost–benefit considerations.

Dreaming as threat simulation

The third version of simulation function theories is the threat simulation theory (TST), proposed by Antti Revonsuo. According to him, to study the biological function of dreams we are required to make a systematic, detailed analysis of the content of dreams across a wide range of large data samples: in the 'normal' population, in cross-cultural samples, and in various special populations, especially hunter-gatherers, children, frequent nightmare sufferers, and traumatized individuals. If in this analysis some dream content characteristics tend to pop out here and there, again and again, those features probably are traces of the original biological function of dreams.

The TST is based on currently available evidence of the systematically recurring dream content characteristics. As mentioned above, the major statistically significant features of dreams are biased toward representing negative elements. Negative emotions and aggression are prominent dream content characteristics, the universally most often reported dream theme is the dreamer being chased or pursued, and the most frequent themes of recurrent dreams and nightmares consist of the dreamer being chased or attacked. The TST interprets this evidence by suggesting that dream consciousness evolved as an off-line model of the world that is specialized in the simulation of various threatening events encountered in the ancestral environment. In the ancestral environment, a threat simulation system that selected memory traces representing life-threatening experiences from long-term memory and constructed frequent threat simulations based on them could have provided our ancestors with a selective advantage in practicing threat recognition and avoidance skills. During dreaming, threat coping skills could have been maintained and rehearsed without the risks of hazardous consequences that accompany threats in real situations. Due to its beneficial effects in enhancing survival and reproductive success, the threat simulation mechanism was selected for, thus propagating its own existence in the ancestral environment.

The predictions of TST have been empirically supported in several studies. Threatening events are frequent in normative dreams, the most frequent type of threat are aggressive encounters (varying in severity from verbal nonphysical aggression to escape and pursuit situations, and to direct physical aggression), the dream self is most often the target of the threatening event, approximately half of the threats pose a severe threat to the dream self, the dream self reacts to the threats in a relevant and adequate manner, and the source of the threat simulation is most often traced back to the personal experiences of the dreamer or to media exposure. Recurrent dreams and nightmares include more severe threat simulations; especially the dangerousness of events is exaggerated in these special dreams. Further, if we are exposed to traumatic events in our lives, our threat simulation system becomes highly activated. The more severe the

trauma, the better the dream recall, the more threatening events dreams contain, and the more life threatening and severe the threats in dreams.

In sum, the dream content studies on TST have mostly supported the predictions of the theory, although it is too early to draw any definitive conclusions about the accuracy of TST. The typical counterarguments against the theory include the following: (1) perhaps the high amount of threat-related content is only the result of selective memory for emotionally charged content; (2) perhaps the bizarre and disorganized nature of dream content does not allow realistic simulation of reality or real threats; (3) posttraumatic stress disorder (PTSD) and frequent nightmares are dysfunctional and disturb sleep, therefore they cannot be regarded as good or functional for the individual who suffers from them; (4) dreaming simulates so many other things too that surely threat simulation cannot be the (only) function of dreams; and (5) the TST is in principle untestable, and thus cannot be falsified or verified.

To begin to answer this critique, even though emotionally charged dream content is easier to recall than mundane content, the same applies for our everyday memories. We tend to forget the ordinary events in our lives, and only remember those that had an emotional impact. When the frequency of threatening events experienced during a specific time period in the waking life was compared with threats simulated in dreams during the same period, it was clear that the dream world contains threats much more frequently than the waking life does.

Second, practically all dreams are well-organized simulations of a world including the self, other characters, objects, and a setting or an environment where the dream takes place. Bizarreness disrupts some parts or features of this otherwise coherent organized world, but although dreams include bizarre elements, bizarreness appears to be a relatively small deviation in the otherwise coherently organized dream experience.

Third, it is true that many severely traumatized individuals suffer from sleep disturbances due to terrifying nightmares. But there are also reasons to believe that ancestral humans did not suffer from the effects of PTSD to the same extent as some individuals in the present environment. The threats in the ancestral environment were frequent, and the ancestral humans were most likely adapted to higher levels of stress and trauma from early childhood on than most contemporary humans. Moreover, in the ancestral setting the threats were often related to everyday activities and thus predictable.

Fourth, although some researchers are willing to accept the idea that dreams are simulations of significant selection pressures, they disagree on what types of events are simulated in dreams. Rehearsal of skills such as adjustment to novelty, social interactions, interpersonal understanding, motor functions, and spatial learning has been suggested. Thus far, however, none of these suggestions have undergone a similar evolutionary cost–benefit analysis as the TST. The suggested alternative evolutionary functions involve little costs if practiced during waking hours, while real threats often result in fatal consequences. Thus, fitness benefits for simulating threats during dreaming are higher than for simulating situations that yield no costs if practiced during wakefulness.

Finally, the main concern faced by the TST is its empirical testability. This concern is not unique only to TST, but to all evolutionary psychological theories that infer cognitive mechanisms from the selection pressures operating in the evolutionary environment of the species. We will never be able to acquire data that would tell us what our ancestors dreamed about thousands and thousands of years ago. But the more we learn about ancestral life and threats in that environment, the more specific hypotheses we can draw concerning what types of events should be simulated in dreams, and thus, indirectly test whether contemporary dream content is compatible with ancestral selection pressures.

To sum up, the TST takes into account the selection pressures most likely present in the human ancestral environment. It proposes a plausible explanation for how dreaming of negative and threatening events might have provided a slight advantage to our ancestors in maintaining and enhancing their threat recognition and avoidance skills. By referring to a single threat simulation mechanism, it furthermore manages to explain a wide variety of dream content data that already exist in the research literature.

Summary and Conclusions

Dreaming can be defined as a subjective experience occurring during sleep and taking the form of an organized, temporally progressing world simulation. Even though dream content is subjective and highly individual, some dream themes, such as being chased or attacked, falling, drowning, or flying, seem to be universal among humans. Detailed dream content studies also reveal many common underlying elements dreams are composed of. Regardless, we still do not know why people dream or if dreams serve any function. Numerous theories have been proposed to explain why we experience dreams, and some of them possess more explanatory power than others. Nevertheless, none of the theories posed are thus far accepted by all researchers in the field of dream science.

See also: Sleep: Implications for Theories of Dreaming and Consciousness.

Suggested Readings

Barrett D (ed.) (1996) Trauma and Dreams. Cambridge, MA: Harvard University Press.
Barrett D and McNamara P (2007) The New Science of Dreaming. Westport, CN: Praeger.
Domhoff WG (1996) Finding Meaning in Dreams: A Quantitative Approach. New York: Plenum Press.
Domhoff WG (2002) The Scientific Study of Dreams: Neural Networks, Cognitive Development, and Content Analysis. Washington, DC: American Psychological Association (APA).
Farthing WG (1992) The Psychology of Consciousness. New York: Prentice Hall.
Garfield P (2001) The Universal Dream Key: The 12 Most Common Dream Themes Around the World. New York: HarperCollins.
Hall CS and Van de Castle R (1966) The Content Analysis of Dreams. New York: Appleton-Century-Crofts.
Hartmann E (1998) Dreams and Nightmares: The New Theory on the Origin and Meaning of Dreams. New York: Plenum Press.
Hobson JA, Pace-Schott EF, and Stickgold R (2000) Dreaming and the brain: Toward a cognitive neuroscience of conscious states. Behavioral Brain Sciences 23(6): 793–842.
Moffit A, Kramer M, and Hoffman R (1993) The Functions of Dreaming. New York: State University Press.
Revonsuo A (2000) The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. Behavioral and Brain Sciences 23(6): 877–901.
Revonsuo A (2006) Inner Presence: Consciousness as a Biological Phenomenon. Cambridge, MA: MIT Press.
Schredl M and Hofmann F (2003) Continuity between waking activities and dream activities. Consciousness and Cognition 12: 298–308.
Strauch I and Meier B (1996) In Search of Dreams. Results of Experimental Dream Research. New York: SUNY Press.
Valli K and Revonsuo A (in press) The threat simulation theory in the light of recent empirical evidence – A review. The American Journal of Psychology.

Biographical Sketch

Katja Valli has a PhD in psychology. She is currently a researcher at the Centre for Cognitive Neuroscience, University of Turku, Finland, and a visiting lecturer at the University of Skövde, Sweden. She is a member of the Consciousness Research Group, led by Professor Antti Revonsuo. Valli has conducted research on dreaming since late 1990s, focusing on the biological function of dreaming. Her other areas of expertise include evolutionary psychology, sleep laboratory research, and sleep-related altered states of consciousness.

William James on the Mind and Its Fringes

B J Baars, The Neurosciences Institute, San Diego, CA, USA

Glossary

Automaticity, habit – The tendency of practiced, predictable events to fade from consciousness after repetition, including voluntary actions, concepts, mental images, and attentional routines. Automatic processes tend to be specialized, take up little conscious capacity, and may resist voluntary control.

Behaviorism – An influential physicalistic philosophy of psychology, some forms of which deny the existence or functional role of consciousness.

Conscious experiences, focal – Experiences like sensory perception, visual imagery, inner speech, visualizable action plans, etc., which tend to have sensory features like color, texture, taste, object identity, and the like.

Conscious experiences, fringe – Experiences that lack specific, sensory qualities, like the tip-of-the-tongue state (the intention to seek a missing word), feelings of knowing, familiarity, and plausibility, intuitive judgments, specific intentions to act, expectations, relational terms in grammar, logic and reasoning, abstract meanings, emotional connotations, and numerous other conscious or quasi-conscious events that can be reported with high accuracy but low sensory specificity.

Conscious experiences, operational definition of – While there is no agreed-upon theoretical definition of conscious experiences, in actual practice scientists have used ''accurate verbal report'' for several centuries with excellent reliability. This is a useful operational definition of a large range of conscious experiences, both sensory and endogenous (as in the case of conscious inner speech, visual imagery, and episodic recall). All of sensory psychophysics is based on this straightforward method, going back to Newton's discovery of the subjective color spectrum, which is highly reliable between subjects.

Hypnosis – A state of surprisingly high compliance with external suggestions, especially after a perceived induction procedure, which may be arbitrary or symbolic. Hypnotic suggests can influence sensory perception, the normal sense of voluntary control, emotions, imagery, pain perception, analgesia, and their known brain correlates. A high percentage of the normal population is considered to be ''highly hypnotizable'' as measured by standardized tasks. Such subjects seem to treat the hypnotist's suggestions as highly credible, although they do not act in violation of their ordinary social norms. Hypnosis may involve an absorbed state with minimal self-examination.

Ideomotor theory – In William James and others, the notion that conscious goals are inherently impulsive, and tend to be carried out by default unless they are inhibited by other conscious thoughts or intentions.

Introspective reports – Reports about conscious experiences, which can range from highly reliable ones (such as psychophysical reports) to unreliable ones (such as mental images evoked by abstract concepts).

Introspectionism – A controversial term attributed by behavioristic historians to nineteenth century researchers on the topic of consciousness. Introspection was explicitly disavowed as a useful method by the most productive experimental psychologist of the nineteenth century, Wilhelm Wundt.Wundt criticized

James' empirical side. Philosophers have done a better job in appreciating William James than have scientists, in general, though he was at least as much an empiricist as a metaphysician.

Fringe-conscious experiences provide an excellent example of the empirical James. Intuitively we tend to think of conscious experiences as clear, percept-like, reportable events that stand out well as figure from ground. Those are the most commonly studied cases. But a large part of our mental life is occupied with 'fringe' events, which are experienced as fuzzy or vague, but which have properties suggesting that something very precise is going on. They include feelings of knowing, of familiarity, of beauty and goodness, of mismatch, incongruity, or surprise. As James points out, the fringe also includes a great variety of judgments, expectations, intentions, abstractions, intuitions, and logical or grammatical relations (like 'if,' 'or,' and 'but'). Further, we seem to have accurate 'feelings of knowing' about potential conscious contents that are readily available, though they are not immediately conscious – our ability to retrieve words, our moods, potential actions, knowledge about others and ourselves, semantic memories, and much more.

Feelings of knowing have now been studied experimentally in considerable detail, and the evidence indicates that (1) they are often accurate; (2) they enjoy high confidence ratings; but (3) they do not involve detailed, structured experiences – unlike the sight of a coffee cup, where we can talk about shape, color, shading, texture, figure-ground contrast, clear temporal boundaries, and much else.

In addition, we now know of brain regions that seem to be activated by some fringe experiences. For example, the 'sense of mental effort' appears to evoke BOLD (fMRI) activity in the anterior cingulate and dorsolateral prefrontal cortex. These regions are known to be involved in expressed goals, and to be triggered by goal conflicts and barriers. These functional properties relate closely to the 'sense of mental effort.'

The most famous example of a fringe experience is the tip-of-the-tongue state – the tantalizing feeling of searching for an elusive word. At first it seems like a curiosity, but then begins to reveal deeper and deeper implications. James writes,

> Suppose we try to recall a forgotten name. The state of our consciousness is peculiar. There is a gap therein; but no mere gap. It is a gap that is intensely active. A sort of a wraith of the name is in it, beckoning us in a given direction, making us at moments tingle with the sense of our closeness, and then letting us sink back without the longed-for term. If wrong names are proposed to us, this singularly definite gap acts immediately so as to negate them. They do not fit into its mold.

The tip-of-the-tongue state is a delayed intention to find a missing word, a mental state that lacks qualities like color, sound, or taste; it has no clear boundaries in space and time, and no contrast between figure and ground. All expectations and intentions seem to be like this. To show the power of such states we need only interrupt some dense flow of predictable experience, for example, a printed _____ like this one. Spontaneously we want to fill in a word that fits. We can see the same effect by interrupting a joke just before the punch line; clever musical composers continuously play with our expectations about songs. Expectations and intentional states like this govern all our activities. They are not images or perceptions. Yet such colorless mental events compete against other sensory events for access to our conscious mental sphere.

James thought that 'fringe' states comprise perhaps a third of our mental life. Some of us now believe that they shape all of our conscious experience, without exception. All human thought and action appear to be driven by expectations. The tip-of-the-tongue state provides a good case to study, because it draws out a colorless expectation over many seconds. We now have the first brain imaging studies of the tip-of-the-tongue state. It shows, as we might expect, that the state activates frontal cortex much more than the sensory regions of the back of the brain. That is presumably why it lacks sensory qualities.

'The fringe' is therefore a fine example of the way in which nineteenth century science went far beyond common sense in studying consciousness. Probably the most famous fringe event is the 'tip-of-the-tongue' experience, which has proven to be a rich and productive domain of experimental study. But the range of fringe or 'vague' conscious phenomena, as James described it, is far broader than is generally recognized. It has only barely been touched in contemporary science.

James Rejected the Unconscious

Toward the end of the nineteenth century scientific thinkers like Pierre Janet and Sigmund Freud began to infer unconscious processes quite freely, based on hypnotic suggestion, conversion hysteria, slips of the tongue, self-serving forgetfulness, and the like. Freud's ideas achieved unparalleled influence, so that the art and literature of the twentieth century is incomprehensible without them.

Unlike Freud, William James fiercely resisted the psychological unconscious. In a remarkable section called 'Refutation of alleged proofs of unconscious thoughts,' he considers the possibility of unconscious intelligence. In his characteristically fair-minded way, he provided ten basic arguments 'pro,' followed by arguments 'con.' The ten arguments pro are still some of the best we have.

The first real clash comes in the famous chapter on 'Habit,' where James considers how it is that, in learning a new skill like riding a bicycle, the conscious details of pedaling, steering, and balancing soon turn into the unconscious routines of expert cycling. We know today that the brain contains numerous unconscious networks that analyze and control such things as balance, eye movements, visual space, and the muscular control. Their existence can be inferred from vast amounts of psychological evidence, and today, with neuroimaging, we can actually see them at work in the brain. Unconscious habits seem to involve less cortical activity, and more subcortical mechanisms like the basal ganglia and cerebellum.

James stated the case for the notion that habits are unconscious:

> . . .we do what originally required a chain of deliberately conscious perceptions and sensations. As the (habitual) actions still keep their intelligent character, intelligence must still preside over their execution. But since our consciousness seems all the while elsewhere engaged, such intelligence must consist of unconscious perceptions, inferences, and volitions.

But he could not tolerate unconscious intelligence. He wrote that,

> Reply: There is more than one alternate explanation. . . One is that the perceptions and volitions in habitual actions may be performed consciously, only so quickly and inattentively that no memory remains. Another is that the consciousness of these actions exists, but is split-off from the rest of the consciousness of the hemispheres. . . .

Habits may therefore reflect fast, hard-to-remember, or split-off conscious contents. Rapid conscious 'flashes' may in fact exist, and there may indeed be dissociated conscious contents, as we know from studies of hypnotic dissociation. So James' counterargument is by no means silly, but few scientists today rule out a major role for complex, unconscious intelligence.

Take another of James' arguments about the unconscious, drawn from biological instincts, like nest-building in birds.

> Instincts, as pursuit of ends by appropriate means, are manifestations of intelligence; but as the ends are not foreseen (consciously), the intelligence must be unconscious.

But again,

> Reply: . . .all the phenomena of instinct are actions of the nervous system, mechanically discharged by stimuli to the senses.

Unconscious processes are 'merely physical.' Crucially, James often resorts to a mind–body argument to rule out unconscious intelligence. In an age of computers, we no longer share James' intuitions that all intelligent processes must be conscious. But he was too much a child of his times to accept the shocking consequences of unconscious thought.

James and his contemporaries just could not imagine a high degree of unconscious intelligence. In their century, reason, purpose, and intelligence were believed to be the exclusive possession of consciousness. 'Unconscious intelligence' seemed a bizarre violation of common sense, as Helmholtz found out in the 1860s, when he suggested that the brain may come to unconscious conclusions about the visual world when direct information is missing. For example, each eye has a blind spot, about the size of a quarter at arm's length. There are simply no light receptor cells in the blind spot. But we almost never see it, because we 'fill in' the gap, based on surrounding colors and textures. This idea is perfectly plausible and is generally believed today. In the nineteenth century it was heresy, leading to furious protests until Helmholtz

felt compelled to withdraw the offending words 'unconscious conclusion' (unbewuesste Schluss). The existence of unconscious processes in the visual brain was not fully accepted until the 1970s! Today many scientists believe that the brain makes innumerable inferences every second.

Thus in the nineteenth century unconscious events had to be attributed to the physical brain. But the brain was held to be incapable of logic; it was still a mechanical servant of lived experience. Ever since Aristotle, logic and rationality were believed to be the unique preserve of consciousness, the seat of reason. Sigmund Freud, who made the unconscious interesting to the public, never really believed that it could reason. Freud's unconscious is bereft of logic or consistency, always in romantic turmoil. It is the irrational 'cauldron of seething excitations,' the dark dungeon of contending passions. The idea of an intelligent unconscious began to make sense only very recently, driven by the new tangible reality of the logic-crunching computer.

When James tried to understand how unconscious habits could form from conscious origins he faced a forbidding paradox. On one side, accepting a brain basis for consciousness would go against his lifelong commitment to free will; on the other, he could not abandon his understanding of brain science. His solution was amazingly awkward: Novel actions had to originate in the nonphysical land of consciousness. As they became habits and faded from consciousness, they would somehow cross the great mind–body divide into the physical domain of the brain. Wrote James,

> An acquired habit, from the physiological point of view, is nothing but a new pathway of discharge formed in the brain, by which certain incoming currents ever after tend to escape. . . .the philosophy of habit is thus a chapter in physics rather than psychology. (italics added)

James was not happy with this awkward dualism; he twisted and turned through all the alternatives, but could not escape contradiction, hemmed in by his own incompatible assumptions.

James and the Mind–Body Problem

This was the empirical James, who summed up the most important discoveries of the nineteenth century. But there is another James, the metaphysician. This is the person who tried for a lifetime to solve the unending problem of free will. The metaphysical James came to the fore again in the last decades of his life. Unfortunately, James' metaphysics undermined his own scientific writing, and may have destroyed his high reputation among psychologists, who were trying to create a stable academic profession.

On metaphysics, James admitted that his thinking was not purely rational. We know from his history of suicidal despair and the need to believe in free will that there is a great undercurrent of emotion in his thinking. James never denied that. In an essay called *The Will to Believe*, he defended a belief in God even in the face of a lack of evidence.

> We have the right to believe at our own risk any hypothesis that is live enough to tempt our will. . . . So if I accept the religious hypothesis because doing so makes me more happy than I would otherwise be, then I am rationally justified in my decision.

And in *Is Life Worth Living?* he wrote,

> These, then, are my last words to you: Be not afraid of life. Believe that life is worth living, and your belief will help to create the fact.

Saving Free Will

Human beings are all mind-brain philosophers, whether we know it or not. Are you freely responsible for your actions? If you say so, you are claiming free-will mentalism. If you do not believe in free will, but think that all human experience is only a fictional gloss on the firing of neurons, physicalism is your game. And if, like most of humanity, you find yourself switching between mind and body explanations in everyday life, you are adopting dualism.

In any moment of the day we can slip subtly between two very different ways of thinking about ourselves. We appeal to a physical vocabulary to explain the effect of aspirin on headaches, but we switch to mind words whenever we want to claim credit or to assert our freedom from external control. Did I break my glasses? No, a book just fell on them. (Physical) But do I work hard to provide for my family? You bet, and I expect a little credit for it. (Mental) Children learn early on to excuse their

actions as uncontrollable accidents when they might be blamed for the results, but to take personal credit when they do something praiseworthy; this childhood pattern hardly changes for many of us until the end of life. When we get a little more sophisticated we learn to import physical causation into psychological events: I did the wrong thing, yes, but it was because of sleepiness, distraction, something came over me. This is of course the key defense in the courts when a defendant claims extenuating circumstances – Prozac, the failures of society, a history of abuse, or the Twinkie defense, a murderous rage said to result from eating too much sugar.

The language of the law is the language of free will, personal responsibility and just deserts. But the language of science is the language of simple physical causation. From this perspective a murderer has no more responsibility for killing his victim than a billiard ball has for a missed shot.

The Mind–Body Vortex Tends to Swallow All Else

The mind–body problem is still today the dominant obsession in the philosophy of mind. It asks how the physical world could possibly be reflected in our private experience; how our subjectively free intentions could emerge in physical action; and how all this could relate to the physical substrate of experience, the brain. Whenever scientists make significant advances – as in two centuries of findings about color perception – philosophers routinely tell us it is not good enough; we still do not know about real consciousness, which is now redefined to exclude the new discoveries. It looks an awful lot like a city-slicker trick that psychologists, simple country folk, fall for with astonishing regularity.

Twentieth century science made a great commitment to physicalism. The most extreme versions of physicalism deny private experience completely, aiming to explain all things exclusively by public observables – neurons, or stimuli and responses, or molecules. Behaviorism is a psychological version of this philosophy, as B.F. Skinner often said. So is the neural reductionism that is widely held by neurobiologists today. Francis Crick's hypothesis is that our experience of each precious moment is fully explainable in terms of neurons in the brain. While this is a long-standing scientific hypothesis, it has an undeniable philosophical agenda.

However, mentalism is alive. The physics Nobelist Roger Penrose claims that consciousness can only be understood by way of quantum mechanics. Penrose argues for a modern mentalism, that reality is to be found in the 'quantum mind.' Physics seems to show a division between quantum phenomena and the visible world of objects, but quantum explanations are thought to be more fundamental. Penrose defines consciousness in terms of the direct apprehension of mathematical truths – exactly Plato's idea 24 centuries ago. The realm of consciousness – the quantum level – underlies visible reality. Those views make no contact with psychology or the brain, but they are sincerely held by some very intelligent scientists.

Each classical position on mind and body seems plausible at certain times and perverse at others. Each is seductive, and each seems to lead to paradox. Intuitively we all swing back and forth between the three classical positions, sometimes in a single sentence. Taking a 'physical' aspirin for a 'mental' headache is intellectually perplexing; being a physicalist and yet taking personal credit for one's own achievements – as if they were freely chosen – is equally inconsistent. Dualism avoids these contradictions at the price of its own unanswered puzzles: How could a mind relate to its brain? How do conscious intentions turn into the physical actions of the muscles? And how do physical sensations end up as conscious experiences?

The mind–body puzzle is not some artifact of Western thought. Each classical position appears early in Asia as well as Europe, starting in India and China and later in Japan and South East Asia. In the West, mentalism was first stated in writing by Plato in fourth century BCE. Athens. A few centuries earlier, it was articulated with great power by Gautama Buddha and the Vedanta philosophers in India, and by the early Taoists in China. All mystical philosophies are mentalistic – they claim, like Roger Penrose, that a transcendent reality underlies our everyday world. Asian philosophies acknowledge the physical world, but suggest that it results from an imperfect realization of one's own consciousness; at the bottom, reality is

mental. Another strand in Indian philosophy is called dvaita, or dualism, from the same root as 'dual.' Dualism, physicalism, and mentalism can be found in many parts of the world.

A lively philosophical cottage industry survives today on the mind–body problem. Philosophical thinking about consciousness is almost exclusively concerned with it, though thousands of other questions can be asked. Every novel has something to say about the varieties of human experience, but we dance around only one philosophical mulberry bush, and the dance never seems to change.

So far, no one has found a settled solution. After more than two millennia of written debate on the subject, arguments are as persistent as ever. Arthur Schopenhauer called it the 'World Knot' – unsolvable but also unavoidable. It is interesting that Schopenhauer's ideas were shaped by the Vedanta scriptures written more than 2000 years before, which had just been translated into Western languages in his time. Wilhelm Wundt, often called the founder of Western experimental psychology, was very much influenced by Schopenhauer, so that we can trace a direct line from the mind–body philosophers of the ancient Indian world to the beginnings of Western scientific psychology. The seductiveness of the World Knot is difficult to overstate. The *Encyclopedia of Philosophy* concludes that

> The mind–body problem remains a source of acute discomfort for philosophers. . . .It may well be that the relation between mind and body is an ultimate, unique, and unanalyzable one. If so, philosophical wisdom consists in . . . accepting it as the anomaly it is.

Science Usually Evades Unresolved Philosophical Puzzles

Great philosophical controversies always arise with major scientific changes. When Copernicus and Newton argued that the sun could keep the earth in orbit, philosophers attacked them for proposing an obviously absurd idea: action at a distance. There were no giant rubber bands connecting the sun and the earth, or the earth and the moon. They just happened to stay connected because of an invisible thing called 'gravity.' But gravity could not be seen or touched. It was an imaginary theoretical idea. Newton had no answer, and we still do not have one today. His response was to be purely pragmatic, saying '*non fingo hypothesi*' – I do not speculate – which was no answer at all.

Charles Darwin's evolutionary theory also evoked attacks from philosophical vitalists, like Henri Bergson, who argued for an invisible 'vital essence' in all living things. That may not seem a powerful argument in the age of biotechnology, but it convinced generations of philosophers that something was wrong with biological science. Biologists essentially ignored those attacks and went on studying the genetics of peas and fruit flies. Over many decades that paid off, and the philosophical arguments faded away.

Successful science does not wait for all philosophical questions to be solved. It is very pragmatic. One of its practical moves is to sidestep questions that cannot be resolved, and simply find a straightforward way to gather evidence. After 25 centuries of debate about mind versus body there is little doubt that scientists should not try to solve the metaphysical problem first – that effort has a long record of failure. They should simply collect evidence about human consciousness and try to understand it. We need to follow our empirical noses unburdened by metaphysical baggage. That is happening today. But William James could not make that choice. He was too deeply committed to his own need to believe in free will, which protected him from suicidal despair.

James Torn

Mind–body debates were one front in the long-running battle between science and religion in the nineteenth century. It took place in heated family arguments between parents and children around Victorian dinner tables, about ancient religious faiths and the new faith in Science and Progress. William James found himself squarely in the middle of this battle. He was a physician after all, trained in the physicalistic medicine of the nineteenth century; he was hired to teach brain anatomy at Harvard. Like other physicians of the time – Freud, Helmholtz, and Charcot – he learned to explain the mind's evolutions in terms of brain processes first of all.

But James was also a child of the transcendental tradition of Emerson and Thoreau, raised in a

family where religious and philosophical debate flourished. Henry James Sr. was an enthusiastic Swedenborgian mystic and close friend of Ralph Waldo Emerson. Henry Jr. became one of the foremost psychological novelists. Alice, their brilliant sister, spent a life of suffering and illness that was thought to have a psychological element. It is not surprising that William was torn between physical and mind-centered science; the perfect person, in fact, to symbolize psychology at the end of the nineteenth century.

James' Quest for Personal Meaning

James was a man blessed, and at times cursed, with an extravagance of talent. As his student wrote, "Brilliant, high-strung, dynamic, vivacious, resilient, unexpected, unconventional, picturesque – these are some of the terms that at once recur in recalling James." It was a widely shared judgment. He was that rare thinker who was also admired as a human being. From his Swedenborgian father, he first learned about metaphysics and religion; with his brother Henry, the novelist, and his sister Alice, he talked for years about everything that three brilliant young people of the nineteenth century could imagine. The family traveled to Europe and the children learned to speak French and German. William trained as a medical doctor and painter, studied with some of the foremost scholars in Europe, accompanied the naturalist Luis Agassiz on an expedition to the Amazon; and finally found a profession teaching at Harvard. In psychology, James' greatest achievement was *The Principles of Psychology* of 1890. It is still a bottomless well of ideas. A few years later he reduced it to a handier *Briefer Psychology*, which became the standard introductory text in America for the next 30 years.

James Sometimes Wandered Away from Standard Science

In the very next chapter, William James jumps off the edge of science as we know it. He begins with a warning:

> The reader who found himself swamped with too much metaphysics in the last chapter will have a still worse time of it in this one, which is exclusively metaphysical.

It is an ominous sign, for here James is forced to adopt panpsychism, the idea that all matter must have some rudimentary consciousness. In a section titled 'Evolutionary psychology demands a mind-dust' we read,

> If evolution is to work smoothly, consciousness in some shape must have been present at the very origin of things. . . .we find that the more clear-sighted evolutionary philosophers are beginning to posit it there. Each atom of the nebula, they suppose, must have had an aboriginal atom of consciousness linked with it; and, just as the material atoms have formed bodies and brains by massing themselves together, so the mental atoms. . . have fused into those larger consciousnesses which we know in ourselves and suppose to exist in our fellow-animals.

The trouble is that one could say the same thing about living matter – just fill in 'life' for 'consciousness.' Arguments for a life essence were common in James' day and well into the twentieth century, before biochemistry came of age. Almost no one believes them today. If we think of life as some indivisible essence it may seem right that it would exist to a tiny degree in every atom and molecule, but with a better understanding of carbon molecules this aura of reasonableness simply fades away. In just the same way, if we think of consciousness as some unanalyzable 'essence' we can make the beguiling leap of logic that everything must be conscious to some degree. But the more we learn about what neurons are doing to make conscious experience possible, the harder it is to believe in panpsychism. Consciousness is first of all a major biological adaptation.

Panpsychism is not testable today, and remains extrascientific. Science thrives on testable questions. By mixing good science with hotly debated metaphysics, James cast doubt on the very foundations of the new psychology.

By the end of the chapter James has rediscovered the soul.

> Many readers have certainly been saying to themselves for the last few pages: 'Why on earth doesn't the poor man say the soul and be done with it?'. . . all the arguments (made here) are also arguments for (the soul). . .. I confess, therefore, that to posit a soul influenced in some mysterious way by the brain-states and responding to them by conscious affections of its own, seems to me the line of least logical resistance. . .

And with a flourish he concluded:

> nature in her unfathomable designs has mixed us of clay and flame, of brain and mind... the two things hang indubitably together and determine each other's being, but how or why, no mortal may ever know.

Summary and Conclusions

William James summarized an extraordinary century of discovery in the 1400 pages of his *Principles* of 1890. Almost all his empirical phenomena are still well-validated today. But he became ensnared in mind–body issues, which were not testable or solvable in his time. Because he was deeply involved in nineteenth century debates about free will and personal meaning, it was not possible for James to fully separate his role as an empiricist from his other identity as a metaphysician. As the most famous advocate of philosophical pragmatism, James is still well known in philosophy today. But in the sciences, his contributions were lost during the behaviorist era, and are poorly understood even today, because of retrospective misunderstandings of his many empirical observations about conscious functions. The role of fringe consciousness is one example of a first-rate empirical discovery that has been well-validated by modern research, but which is still not widely understood as an entire separate category of mental life. Thus we still have a great deal to learn from James' *Principles* of 1890.

See also: Neuroscience of Volition and Action; Philosophical Accounts of Self-Awareness and Introspection; Self: Personal Identity; Self: The Unity of Self, Self-Consistency.

Suggested Readings

Baars BJ (1988) A Cognitive Theory of Consciousness. New York, NY: Cambridge University Press.

Bargh JA and Morsella E (2008) The unconscious mind. Psychological Science 3(1): 73–79.

Cohen MS, Kosslyn SM, Breiten HC, et al. (1996) Changes in cortical activity during mental rotation: A mapping study using functional MRI. Brain 119: 89–100.

Kikyo H, Kenichi O, and Miyashita Y (2002) Neural correlates for feeling-of-knowing: An fMRI parametric analysis. Neuron 36: 177–186.

Maril A, Wagner AD, and Schacter Daniel L (2001) On the tip of the tongue: An event-related fMRI study of semantic retrieval failure and cognitive conflict. Neuron 31: 653–660.

Wagner J, Stephan T, Kalla R, et al. (2008) Mind the bend: Cerebral activations associated with mental imagery of walking along a curved path. Experimental Brain Research 191: 247–255.

Biographical Sketch

Bernard J. Baars is former senior research fellow in theoretical neurobiology at the Neurosciences Institute in San Diego (www.nsi.edu). His PhD is in cognitive psychology from UCLA. He is interested in human language, the brain basis of consciousness, volition, and a variety of related topics including the history of scientific studies of consciousness, and neuroethics. Baars pioneered a cognitive theory of consciousness called Global Workspace Theory, which is widely cited in scientific and philosophical sources. Together with William P. Banks, Baars has edited the journal *Consciousness and Cognition* for more than

fifteen years (from Academic Press/Elsevier). With Nicole M. Gage, Baars has written an introductory text for cognitive neuroscience, called Cognition, Brain and Consciousness: An Introduction to Cognitive Neuroscience. (Baars and Gage, Eds. San Diego,: Elsevier/Academic Press, 2007). Baars was founding president of the Association for the Scientific Study of Consciousness and has an ongoing research collaboration for large-scale cognitive modeling with professor Stan Franklin (University of Memphis, Institute for Intelligent Systems).

Consciousness: An Introduction to Cognitive Neuroscience. (Baars and Gage, Eds. San Diego, Calif.: Elsevier/Academic Press, 2007). Baars was founding President of the Association for the Scientific Study of Consciousness and has an ongoing research collaboration for large-scale cognitive modeling with professor Stan Franklin (University of Memphis, Institute for Intelligent Systems).

Subject Index

Notes

Cross-reference terms in italics are general cross-references, or refer to subentry terms within the main entry (the main entry is not repeated to save space). Readers are also advised to refer to the end of each article for additional cross-references - not all of these cross-references have been included in the index cross-references.

The index is arranged in set-out style with a maximum of three levels of heading. Major discussion of a subject is indicated by bold page numbers. Page numbers suffixed by T and F refer to Tables and Figures respectively. *vs.* indicates a comparison.

This index is in letter-by-letter order, whereby hyphens and spaces within index headings are ignored in the alphabetization. For example, acid rain is alphabetized after acidity, not after acid(s). Prefixes and terms in parentheses are excluded from the initial alphabetization.

Where index subentries and sub-subentries pertaining to a subject have the same page number, they have been listed to indicate the comprehensiveness of the text.

A

D

E

F

G

H