Higher-order theories of consciousness

Article *in* Scholarpedia · January 2008

DOI: 10.4249/scholarpedia.4407 · Source: DBLP

2 authors:

David Rosenthal
The Graduate Center, CUNY
106 PUBLICATIONS **4,579** CITATIONS
SEE PROFILE

Josh Weisberg
University of Houston
25 PUBLICATIONS **342** CITATIONS
SEE PROFILE

Higher-Order Theories of Consciousness

An Anthology

Edited by Rocco J. Gennaro

Advances in Consciousness Research

Higher-Order Theories of Consciousness

Advances in Consciousness Research

Advances in Consciousness Research provides a forum for scholars from different scientific disciplines and fields of knowledge who study consciousness in its multifaceted aspects. Thus the Series will include (but not be limited to) the various areas of cognitive science, including cognitive psychology, linguistics, brain science and philosophy. The orientation of the Series is toward developing new interdisciplinary and integrative approaches for the investigation, description and theory of consciousness, as well as the practical consequences of this research for the individual and society.

Series A: Theory and Method. Contributions to the development of theory and method in the study of consciousness.

Volume 56

Higher-Order Theories of Consciousness: An Anthology
Edited by Rocco J. Gennaro

Higher-Order Theories of Consciousness

An Anthology

Edited by

Rocco J. Gennaro
Indiana State University

John Benjamins Publishing Company
Amsterdam/Philadelphia

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Higher-order theories of consciousness : an anthology / edited by Rocco J. Gennaro.
p. cm. (Advances in Consciousness Research, ISSN 1381–589X ; v. 56)
Includes bibliographical references and indexes.
1. Consciousness. 2. Self-consciousness. 3. Thought and thinking. 4. Phenomenological psychology. I. Gennaro, Rocco J. II. Series.

B105 C477H54 2004
126-dc22 2004041066
ISBN 90 272 5191 6 (Eur.) / 1 58811 495 3 (US) (Hb; alk. paper)
ISBN 90 272 5192 4 (Eur.) / 1 58811 496 1 (US) (Pb; alk. paper)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of contents

Contributors

Alex Byrne
Department of Linguistics
and Philosophy
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
abyrne@mit.edu

Peter Carruthers
Department of Philosophy
University of Maryland
College Park, MD 20742, USA
pcarruth@umd.edu

Zoltán Dienes
Department of Psychology
University of Sussex
Brighton BN1 9QG, UK
dienes@biols.susx.ac.uk

Donelson E. Dulany
Department of Psychology
University of Illinois, Urbana-Champaign
603 East Daniel Street
Champaign, IL 61820, USA
ddulany@s.psych.uiuc.edu

Rocco J. Gennaro
Department of Philosophy
Root Hall A-138D
Indiana State University
Terre Haute, IN 47809, USA
rocco@indstate.edu

Valerie Gray Hardcastle
Department of Philosophy
Virginia Polytechnic Institute
Blacksburg, VA 24061, USA
valerie@vt.edu

Christopher S. Hill
Department of Philosophy
Brown University
Providence, RI 02912, USA
Christopher_Hill@brown.edu

Robert W. Lurz
Department of Philosophy
Brooklyn College
2900 Bedford Avenue
Brooklyn, NY 11210, USA
rlurz@brooklyn.cuny.edu

William G. Lycan
Department of Philosophy
University of North Carolina
at Chapel Hill
Chapel Hill, NC 27599, USA
ujanel@isis.unc.edu

Josef Perner
Department of Psychology
University of Salzburg
Hellbrunnerstrasse 34
A-5020 Salzburg, AUSTRIA
josef.perner@sbg.ac.at

William S. Robinson
Department of Philosophy
and Religious Studies
402 Catt Hall
Iowa State University
Ames, IA 50011, USA
wsrob@iastate.edu

Edmund T. Rolls
Department of Experimental Psychology
University of Oxford
South Parks Road
Oxford OX1 3UD, UK
edmund.rolls@psy.ox.ac.uk

David M. Rosenthal
Department of Philosophy
and Cognitive Science
Graduate Center, City University
of New York
365 5th Avenue
New York, NY 10016, USA
dro@ruccs.rutgers.edu

William Seager
Division of Humanities/Philosophy
University of Toronto at Scarborough
Scarborough, Ontario
CANADA M1C 1A4
seager@scar.utoronto.ca

Robert Van Gulick.
Department of Philosophy
541 Hall of Languages
Syracuse University
Syracuse, NY 13244, USA
rnvangul@syr.edu

Acknowledgments and dedication

I would like to thank Maxim Stamenov and Bertie Kaal for so readily accepting the idea of an anthology devoted entirely to higher-order theories of consciousness. I appreciate their encouragement of and dedication to the production of this volume. Their helpful advice and prompt responses to my inquiries enabled this book to appear in a timely fashion and prevented unnecessary delays. It is perhaps most important for me, as editor, to acknowledge the work and cooperation of the contributors to this volume. An anthology is, in the end, only as good as the essays it contains. Authors frequently sent drafts of their papers to each other and commented on other papers. Contributors were also very responsive to my comments. I thank all authors for their hard work and wonderful contributions.

Finally, I would like to acknowledge the fact that some of my work on this anthology was supported by an Indiana State University Summer 2003 Research Grant. I would also like to thank David Peter of Indiana State University for some technical assistance in producing the final electronic copy of the manuscript.

This book is dedicated to my son, Joseph Rocco Gennaro, who is 11 months old at the present time. My wife, Deidra, and I are truly fortunate to have him in our lives. He is, in my view, already having quite a number of higher-order thoughts, although they are, no doubt, very primitive ones at this early stage in his development!

November 2003
R.J.G.
Indiana State University

Chapter 1

Higher-order theories of consciousness

An overview

Rocco J. Gennaro

1. General introduction and terminology

Explaining the nature of consciousness is one of the more important and perplexing areas in philosophy. What is consciousness? How is the conscious mind related to the body? Can consciousness be explained in terms of brain activity? How can we explain the sensation of the smelling of a rose or a conscious visual experience? Given that philosophy is interdisciplinary by its very nature, the problem of consciousness is also explored in such related fields as psychology and neuroscience. One question that should be answered by any viable theory of consciousness is: What makes a mental state a conscious mental state? That is, what transforms a nonconscious mental state into a conscious one? There is a long tradition that has attempted to understand consciousness in terms of some kind of higher-order awareness. For example, John Locke (1689/1975) once said that "consciousness is the perception of what passes in a man's own mind." This intuition and attempt to explain consciousness has been recently revived by a number of philosophers (Rosenthal 1986, 1990, 1993a, 2000, 2004; Armstrong 1968, 1981; Lycan 1996, 2001). In general, the idea is that what makes a mental state conscious is that it is the object of some kind of higher-order representation (HOR). A mental state M becomes conscious when there is a HOR of M. A HOR is a "meta-psychological" state, i.e. a mental state directed at another mental state. So, for example, my desire to write a good introduction becomes conscious when I become "aware" of the desire. Intuitively, it seems that conscious states, as opposed to nonconscious ones, are the mental states that I am "aware of" in some sense. Any theory which attempts to explain consciousness in terms of higher-order states is known as a higher-order (HO) theory of consciousness. It is best initially to use the more neutral term 'repre-

sentation' because there are a number of different kinds of higher-order theory, depending upon how one characterizes the HOR-state in question. Moreover, to be clear, the sense of 'conscious state' that I have in mind is the same as Nagel's (1974) sense, i.e. there is "something it is like to be in that state" from the subjective or first-person point of view. When I am, for example, having a conscious visual experience, there is something it "seems" or "feels" like from my first-person subjective perspective.

The key to HO theories is their hierarchical or iterative structure, and so they have been called "double-tiered" theories (Güzeldere 1995). HO theories are also attractive to some philosophically inclined psychologists and neuroscientists partly because they suggest a very natural realization in the brain structure of humans and other animals (see e.g. Rolls 1999; Weiskrantz 2000). At the risk of oversimplification, if we think of the brain as developing layers upon layers corresponding to increasing sophistication in mental ability, then the idea is that mental states corresponding to various "higher" areas of the brain (e.g. cortex) are directed at various "lower" states rendering them conscious. It is important to note, however, that HO theories do not attempt to reduce consciousness *directly* to neurophysiological states. Unlike some other theories of consciousness (Crick & Koch 1990; Crick 1994), they are not reductionist in the sense that they do not attempt to explain consciousness directly in physicalistic (e.g. neurophysiological) terms. Instead, HO theories attempt to explain consciousness in *mentalistic* terms, that is, by reference to such notions as 'thoughts' and 'awareness.' Thus, conscious mental states arise when two *non*conscious mental states are related in a certain specific way; namely, that one of them (the HOR) is directed at the other (M). HO theorists are normally of the belief that such mental states are identical with brain states, but they tend to treat this matter as a further second step reduction for empirical science. Moreover, it is important to keep in mind that all HO theorists are united in the belief that their approach can better explain consciousness than any purely first-order representational (FOR) theory, such as those presented by Dretske (1995) and Tye (1995).

Two terminological distinctions should also be made at this point (see e.g. Rosenthal 1990, 1993a). First, a distinction is often made between *creature* consciousness and *state* consciousness. The former recognizes that we often speak of whole organisms as conscious or even as simply "awake." The latter recognizes that we also speak of individual mental states as conscious. Explaining state consciousness is the primary focus for most researchers, though there are also, no doubt, some interesting connections between state and creature consciousness. Second, some authors also utilize a distinction between intransitive

and transitive consciousness. We sometimes use the word 'conscious' as in our being "conscious of" something. This is the transitive use. On the other hand, we also have the "x is conscious" locution and, due to the lack of a direct object, this is called the intransitive use. Thus, analyzing state consciousness in terms of this distinction leads to the idea that a mental state is intransitively conscious just in case one is transitively conscious of it. We must be careful to guard against any suggestion of circularity here (see Section 3.1 below); it is important to keep in mind that the HO (transitively conscious) state is not normally itself intransitively conscious. There is the potential for confusion, however, so some HO theorists prefer to speak of the HO (nonconscious) "awareness of" the lower-order state instead of invoking the transitive use of 'consciousness.'

2. Different kinds of HO theory

There are various kinds of HO theory depending on how one understands the HOR. The most common division is between higher-order *thought* (HOT) theories and higher-order *perception* (HOP) or higher-order *experience* (HOE) theories. HOT theorists, such as David M. Rosenthal, think it is better to understand the HOR as a thought of some kind. HOTs are here treated more like *cognitive* states involving some kind of conceptual component. The latter argue that the HOR is closer to a *perceptual* or *experiential* state of some kind (e.g. Lycan 1996) which does not require the kind of conceptual content invoked by HOT theorists. Due to Kant (1781/1965), HOP theory is also sometimes referred to as "inner sense theory" as a way of emphasizing its sensory or perceptual aspect. Although HOT and HOP theorists agree on the need for a HOR theory of consciousness, they are sometimes concerned to argue for the superiority of their respective positions (such as in Rosenthal, this volume; Lycan, this volume). Some philosophers, however, have argued that the HOP theory ultimately reduces to the HOT theory (Güzeldere 1995). Others have argued that the difference between these theories is perhaps not as important or as clear as some think it is (Gennaro 1996; Van Gulick 2000).

Finally, Peter Carruthers (2000) has recently proposed that it is better to think of HOTs as *dispositional* states instead of the standard view that the HOTs are *actual*, though he also understands his dispositional HOT theory to be a form of HOP theory (Carruthers, this volume). His overall basic idea, however, is that the conscious status of an experience is due to its *availability* to higher-order thought. A key idea is that "conscious experience occurs when perceptual contents are fed into a special short-term buffer memory store, whose

function is to make those contents available to cause HOTs about themselves." (Carruthers 2000: 228). So some first-order perceptual contents are available to a higher-order "theory of mind mechanism," which transforms those representational contents into conscious contents. Thus, no actual HOT occurs. Instead, some perceptual outputs acquire a dual intentional content; for example, a conscious experience of red not only has a first-order (analog) content of 'red,' but also has the higher-order content 'seems red' or 'experience of red.' Carruthers also makes interesting use of so-called "consumer semantics," such as teleosemantics (Millikan 1984) and inferential role semantics (e.g. Peacocke 1992) in order to fill out his theory of phenomenal consciousness. The basic idea is that the content of a mental state depends, in part, on the powers of the organisms which "consume" that state, e.g. the kinds of inferences which the organism can make when it is in that state. Daniel Dennett (1991) is sometimes credited with an earlier, though somewhat different, version of a dispositional account (see Carruthers 2000: Chapter 10). Carruthers' dispositional theory is criticized, in this volume, by Gennaro and Rosenthal.

One other source of disagreement within HO theories concerns the issue of whether or not the HO state should be understood as *extrinsic* to (i.e. entirely distinct from) its target mental state. This is the view defended by David Rosenthal. On the other hand, several authors have recently challenged this assumption. For example, following in the tradition of Brentano (1874/1973), Gennaro (1996: 21–30) has argued that, when one has a first-order conscious state, the HOT is better viewed as *intrinsic* to the target state, so that we have a complex conscious state with parts. Gennaro calls this the "wide intrinsicality view" (WIV) and he has also recently argued that Jean-Paul Sartre's theory of consciousness can be understood in this way (Gennaro 2002). Robert Van Gulick (2000, this volume) has also explored the alternative that the HO state is part of the "global" conscious state. He calls such states "HOGS" (= higher-order global states) within which the lower-order state is "recruited" and becomes conscious. Both Gennaro and Van Gulick have suggested that conscious states can be understood materialistically as global states of the brain, and it would be better to treat the first-order state as part of the larger complex brain state. This idea was also briefly explored by Thomas Metzinger who focused on the fact that consciousness "is something that unifies or *synthesizes* experience." (Metzinger 1995: 454; cf. Gennaro 1996: Chapter 3). Metzinger calls this the process of "higher-order binding" and thus uses the acronym 'HOB.' A point of emphasis in all of these alternatives is on the concept of global meta-representation within a complex brain state.[1]

3. Some important issues

There are a number of reoccurring themes in the literature on HO theories, some more important and controversial than others. I cannot hope to do all of them justice in a short introduction, but it is worth at least briefly mentioning the following six:[2]

3.1 Circularity and regress

A common initial objection to HOR theories is that they are circular and lead to an infinite regress. For example, it might seem that the HOT theory results in circularity, i.e. by defining consciousness in terms of HOTs. It also might seem that an infinite regress results because a conscious mental state must be accompanied by a HOT, which, in turn, must be accompanied by another HOT *ad infinitum*. However, the standard reply from the HOT theorist is that when a conscious mental state is a first-order world-directed state the higher-order thought (HOT) is *not* itself conscious; otherwise, circularity and an infinite regress would follow. Moreover, when the HOT is itself conscious, there is a yet higher-order (or third-order) thought directed at the second-order state. In this case, we have *introspection* which involves a conscious HOT directed at an inner mental state. When one introspects, one's attention is directed back into one's mind. For example, what makes my desire to write a good introduction a conscious *first-order* desire is that there is a (nonconscious) HOT directed at the desire. In such a case, my conscious focus is directed at the introduction and so I am not consciously aware of having the HOT from the first-person point of view. When I introspect that desire, however, I then have a *conscious* HOT (accompanied by a yet higher, third-order, HOT) directed at the desire itself (on this point see Rosenthal 1986: 337–338). So instead of threatening the HOT theory, this issue actually brings out an important subtlety of any viable HOR theory. Although most opponents of HO theories are inclined to accept the standard reply to this objection, some are still not so easily satisfied (Rowlands 2001).

3.2 Animal and infant consciousness

Perhaps the most common objection to HO (especially HOT) theories is that various animals are not likely to have to the conceptual sophistication required for HOTs, and so that would render animal (and infant) consciousness very unlikely (Dretske 1995; see also Seager, this volume). Are cats and dogs capable

of having such complex higher-order thoughts to the effect that "I am in mental state M."? This is normally treated as a problem for HOT theory because the vast majority of us believe that animals and infants are clearly capable of having conscious states. Although most who bring forth this objection are not HO theorists, one notable HO theorist actually embraces the conclusion that animals and infants do not have phenomenal consciousness (Carruthers 1989). Gennaro (1993, 1996) has replied to Carruthers on this point; for example, it is argued that the HOTs need not be as sophisticated as it might initially appear and that there is evolutionary and comparative neurophysiological evidence supporting the conclusion that animals have conscious mental states. It seems fair to say that most HO theorists do not wish to accept the absence of animal or infant consciousness as a consequence of holding the theory. And even the mere perception that HOT theories of consciousness can lead to this consequence clearly causes many to shy away from any form of HO theory. The debate continues, however, in Carruthers (2000) and Gennaro (this volume). This issue also brings out a host of related important questions: Must all conscious mental states have conceptual content? How should concepts be understood and defined in the context of HOT theory? What exactly is the difference between perception and thought? How "rich" in content are our conscious states? Are there degrees of self-consciousness?

3.3 The problem of the rock

Another objection to HO theories has been referred to as the "problem of the rock" (Stubenberg 1998) and the "generality problem" (Van Gulick 2000, this volume), but it is perhaps originally due to Alvin Goldman (Goldman 1993). When I have a thought about a rock, it is certainly not true that the rock thereby becomes conscious. So why should I suppose that when I think about a lower-order mental state (M), it becomes conscious? Indeed, why should being the intentional object of a meta-state confer consciousness on M? This seems puzzling to many and the objection forces HO theorists to explain how adding the HO state changes a nonconscious state to a conscious one since having a similar state directed at outer objects does not render them conscious. There have been, however, a number of responses to this kind of objection (Rosenthal 1990/1997; Lycan 1996; Van Gulick 2000; Gennaro, unpublished). A common theme is to remind the objector that there is a principled difference between the objects of the HO states in question. Rocks and the like are not mental states in the first place, and so HO theorists are first and foremost trying to explain how

a *mental state* becomes conscious. Nonetheless, the prospect of a damaging *reductio* of HO theories looms if this objection cannot be met properly.

3.4 The hard problem of phenomenal consciousness

The above objection leads somewhat naturally to the following problem: In the spirit of David Chalmers' (1995; cf. Shear 1997) discussion of what he calls the "hard problem of consciousness," it might be asked just how exactly any HO theory really *explains* the subjective aspect of conscious experience (Stubenberg 1998; Siewert 1998). How or why does a mental state come to have a first-person qualitative "what it is like" aspect by virtue of the presence of a HOR directed at it? I think it is fair to say that some HO theorists have been slow to address this problem, though a number of overlapping responses have emerged. Some have argued that this objection simply misconstrues the main and more modest purpose of (at least, their) HO theories. The claim here is that HO theories are theories of consciousness only in the sense that they are attempting to explain what differentiates conscious from nonconscious states, i.e. in terms of a higher-order awareness of some kind. A full account of 'qualitative properties' or 'sensory qualities' (which can themselves be nonconscious) can be found elsewhere in their work, but is independent of their theory of consciousness (Rosenthal 1991; Lycan 1996, 2001). Thus, a full explanation of phenomenal consciousness does require more than a HO theory, but that is no objection to HO theories as such. Another response is that proponents of the hard problem unjustly raise the bar as to what would count as a viable explanation of consciousness so that any such reductivist attempt would inevitably fall short (Carruthers 2000). Part of the problem also, then, is a lack of clarity about and disagreement on what would even count as an explanation of consciousness (Van Gulick 1995). Anyone familiar with the literature knows that there are also significant terminological difficulties in the use of various crucial terms (see Byrne, this volume), which also sometimes inhibits progress and mutual understanding on this matter.[3]

3.5 Misrepresentation

A further important objection to HO approaches is the question of how such theories can explain cases where the HO state might misrepresent the lower-order (LO) mental state (Byrne 1997; Neander 1998; Levine 2001). After all, if we are dealing with a representational relation between two states, it seems possible for misrepresentation or malfunction to occur. If it does, then what

explanation can be offered by the HO theorist? If my LO state registers a red percept and my HO state registers a thought about something green due, say, to some neural misfiring, then what happens? It seems that problems loom for any answer from a HO theorist and the cause of the problem has to do with the very nature of the HO theorist's belief that there is a representational relation between the LO and HO states. For example, if the HO theorist takes the option that the resulting conscious experience is reddish, then it seems that the HO state plays no role in determining the qualitative character of the experience. This is an objection that must be taken seriously and it forces HO theorists to be clearer about just how to view the relationship between the LO and HO states. A reply to this objection is offered by Gennaro (in this volume).

3.6 The causal and/or inferential relation between the lower and higher-order states

A final pair of related issues concerns other aspects of the relationship between the LO and HO states. First, it is agreed upon by all HO theorists that the HO state must become aware of the LO state *non-inferentially*. We might say, then, that the HOR must be caused non-inferentially by the LO state in order to make the LO state conscious. The point of this condition is mainly to rule out alleged counterexamples to HO theory, such as cases where I become aware of my nonconscious desire to kill my boss because I have consciously inferred it from a session with a psychiatrist or where my envy becomes conscious after making inferences about my own behavior. The characteristic *feel* of such a conscious desire or envy may be absent in these cases, but, since awareness of them arose via conscious inference, the HO theorist attempts to account for them by adding this non-inferential condition. Of course, some still ask: why should it matter so much how I become aware of the LO state?

Second, and perhaps more controversial, is whether or not to understand the LO state as the only or primary "cause" of the HO state, in some important sense of that term (see e.g. Rosenthal 1993b; Gennaro 1996: 73–75). HO (especially HOP) theorists sometimes speak that way and there are advantages to such a view; for example, it would be a natural way to explain how the HO state gets directed at or refers to the appropriate LO state. However, there are also reasons to shy away from such a straightforward causal connection and to adopt the weaker notion of "accompaniment" or "co-occurrence" between states. In addition to avoiding the more general vexed problems of causality and reference within the context of HO theory, mere accompaniment would seem sufficient to explain state consciousness (especially according to HOT theo-

rists). In any case, HOT theorists, such as Rosenthal, tend to hold that the LO state is somehow implicated in causing the HOT, but there must be other factors that figure in causing it as well. Additionally, it is possible that there is instead a reliable tracking condition (in the brain) which obtains between the LO state and the HOT directed at it. For example, the LO state and its HOT might have a common cause such that whenever the former is produced the latter is also (typically and reliably) caused as well. Needless to say, however, not everyone is satisfied with the HO theorist's treatment of these issues (Francescotti 1995).[4]

4. The essays

This book is divided into two general parts. **Part I** contains essays by authors who have defended some form of HO theory. Although they are often concerned to discredit one or more of the *other HO theories*, it must be kept in mind that they are strongly united in their agreement in the superiority of HO theories over, for example, various first-order (FO) accounts of consciousness (Dretske 1995; Tye 1995). Part I begins with a chapter entitled "Varieties of Higher-Order Theory" (Chapter 2) by David M. Rosenthal who argues for his preferred version of "extrinsic HOT theory" by systematically reviewing and critiquing alternative HO theories, including HOP and dispositional HOT theory. In the following chapter entitled "Higher-Order Thoughts, Animal Consciousness, and Misrepresentation: A Reply to Carruthers and Levine" (Chapter 3), Rocco J. Gennaro defends HOT theory at length against two of the key problems explained in the previous section: animal consciousness and misrepresentation. In "Higher-Order Global States (HOGS): An Alternative Higher-Order Model of Consciousness" (Chapter 4), Robert Van Gulick further develops his HOGS version of HO theory. For example, Van Gulick attempts to explain how HOGS theory can avoid some of the difficulties with standard HOP and HOT theory, and why his theory should indeed be understood as a kind of *higher-order* theory instead of a FO theory. In Chapter 5, William G. Lycan argues for "The Superiority of HOP to HOT." In addition to responding to critics, he argues that HOP theory is preferable on a number of grounds, such as its ability to account for so-called "recognitional concepts" and the voluntary control we find in consciousness. Peter Carruthers' cleverly titled "HOP over FOR, HOT Theory" (Chapter 6) argues for what he calls "the dispositionalist HOT version of HOP theory." Carruthers argues that FOR and actualist HOT theories cannot give an adequate account of purely recognitional con-

cepts of experience; nor can they properly explain the distinction between conscious and nonconscious perceptual states, especially when we examine such phenomena as blindsight and visual agnosia. In Chapter 7, Edmund T. Rolls presents a somewhat different form of HOT theory which he calls "A Higher-Order Syntactic Thought (HOST) Theory of Consciousness" primarily against the background of a theory of emotion. Citing a wealth of empirical literature, he argues, for example, that it is best to take an information processing and brain design approach to consciousness such that the HOTs in question are directed at semantically based thoughts and that HOTs have important evolutionary adaptive value. Finally, in "Assumptions of a Subjective Measure of Consciousness: Three Mappings" (Chapter 8), Zoltán Dienes and Josef Perner use the HOT theory as a tool by which they analyze the appropriate use of various subjective measures of conscious awareness, such as the so-called "zero-correlation criterion." They focus on the confidence-accuracy relationship and then use the zero-correlation criterion in both subliminal perception and implicit learning.

Part II includes papers by authors who do not subscribe to any form of HO theory. Thus, their papers not only criticize one or more HO theory, but often also urge us to look for an altogether different solution to the problem of consciousness. Part II begins with an essay entitled "What Phenomenal Consciousness is Like" by Alex Byrne (Chapter 9). He first does us the enormous favor of helping us to navigate through the terminological jungle that we find in the literature, especially regarding various attempts by HOR theorists to distinguish between "experiences" and "conscious experience." Byrne then argues that HOR theories are mistaken primarily on the basis of rebutting those attempts. Robert Lurz argues that the assumed dichotomy between FOR and HOR theories of consciousness is a false one. In "Either FOR or HOR: A False Dichotomy" (Chapter 10), Lurz rejects both of these alternatives, and then argues for a third option he calls "same-order representationalism" (SOR) whereby what makes a mental state (M) conscious is that one is aware of the *intentional content* of M. William Seager takes "A Cold Look at HOT Theory" (Chapter 11) by focusing on whether or not animals can attribute mental states to other animals. Citing much important empirical literature, he argues that since most animals cannot do so, they therefore cannot also self-attribute mental states and, hence, on HOT theories, they cannot be conscious beings. As we saw earlier (Section 3.2), this is a conclusion which most HO theorists wish to avoid. In Chapter 12, Valerie Gray Hardcastle pulls no punches against HO theory, and particularly against David Rosenthal's arguments for HOT theory. As the title suggests, her "HOT Theories of Con-

sciousness: More Sad Tales of Philosophical Intuitions Gone Astray," primarily aims at exposing weaknesses underlying the entire HO approach, such as the "intuition" that we are "aware" that we are in conscious states, that HOT theory is empirically adequate, and the preference for reducing consciousness in mentalistic terms. In "A Few Thoughts Too Many?" (Chapter 13), William S. Robinson uses the fact that we often engage in subvocal speech to argue that, amongst various explanations, a simpler view not including any HORs should be preferred on grounds of parsimony. Significant doubts are thus raised about the claim that HORs are required for consciousness. Donelson E. Dulany, in "Higher Order Representation in a Mentalistic Metatheory" (Chapter 14), critiques various HO theories and offers an alternative "mentalistic" metatheory of consciousness. Dulany also challenges the current orthodoxy that there are unconscious mental states, such as unconscious perceptions, which form a key basis for any HO theory. Finally, in "Ouch! An Essay on Pain" (Chapter 15), Christopher S. Hill presents a theory of pain such that awareness of pain is akin to such paradigmatic forms of perceptual awareness as vision and hearing. Hill ultimately considers his theory in relation to HOP theory, but, despite some apparent similarities, argues that they are importantly different and for the superiority of the former over the latter.

Notes

1. Thomas Natsoulas also has a continuing series of papers defending intrinsic theory, beginning with Natsoulas (1996). See also Kriegel (2003) and many of the essays in Kriegel and Williford (forthcoming).

2. Virtually all of them are addressed to some extent in Byrne (1997).

3. For another HO theorist's attempt to address the hard problem, see Gennaro (unpublished).

4. For more on this issue, see the exchange between Natsoulas (1993) and Rosenthal (1993b).

References

Armstrong, D. (1968). *A materialist theory of the mind*. New York: Humanities.
Armstrong, D. (1981). *The nature of mind and other essays*. Ithaca, NY: Cornell.
Brentano, F. (1874/1973). *Psychology from an empirical standpoint*. New York: Humanities.
Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies, 86*, 103–129.

Carruthers, P. (1989). Brute experience. *Journal of Philosophy, 86*, 258–269.
Carruthers, P. (2000). *Phenomenal consciousness* (New York: Cambridge University Press).
Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*, 200–219.
Crick, F. (1994). *The astonishing hypothesis: The scientific search for the soul*. New York: Scribner.
Crick, F. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2*, 263–275.
Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown, and Co.
Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.
Francescotti, R. (1995). Higher-order thoughts and conscious experience. *Philosophical Psychology, 8*, 239–254.
Gennaro, R. (1993). Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers, 22*, 51–69.
Gennaro, R. (1996). *Consciousness and self-consciousness: A defense of the higher-order thought theory of consciousness*. Amsterdam & Philadelphia: John Benjamins.
Gennaro, R. (2002). Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy, 32*, 293–330.
Gennaro, R. (unpublished). The HOT theory of consciousness: between a rock and a hard place?
Goldman, A. (1993). Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition, 2*, 364–382.
Güzeldere, G. (1995). Is consciousness the perception of what passes in one's own mind? In T. Metzinger (Ed.), *Conscious experience* (pp. 335–357). Schöningh: Imprint Academic.
Kant, I. (1781/1965). *Critique of pure reason*. Norman Kemp Smith (trans.). New York: St. Martin's Press.
Kriegel, U. (2003). Consciousness as intransitive self-consciousness: two views and an argument. *Canadian Journal of Philosophy, 33*, 103–132.
Kriegel, U. & K. Williford (Eds.) (forthcoming). *Consciousness and self-reference*. Cambridge, MA: MIT Press.
Levine, J. (2001). *Purple haze*. Cambridge, MA: MIT Press.
Locke, J. (1689/1975). *An essay concerning human understanding*. P. Nidditch (Ed.). Oxford: Clarendon.
Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.
Lycan, W. G. (2001). A simple argument for a higher-order representation theory of consciousness. *Analysis, 61*, 3–4.
Metzinger, T. (1995). Faster than thought: holism, homogeneity and temporal coding. In T. Metzinger (Ed.), *Conscious experience* (pp. 425–461). Schöningh: Imprint Academic.
Millikan, R. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435–450.
Natsoulas, T. (1993). What is wrong with appendage theory of consciousness. *Philosophical Psychology, 6*, 137–154.
Natsoulas, T. (1996). The case for intrinsic theory I. An introduction. *The Journal of Mind and Behavior, 17*, 267–286.

Neander, K. (1998). The division of phenomenal labor: a problem for representational theories of consciousness. In James Tomberlin (Ed.), *Language, mind, and ontology* (pp. 411–434). Oxford: Blackwell.

Peacocke, C. (1992). *A study of concepts*. Cambridge, MA: MIT Press.

Rolls, E. T. (1999). *The brain and emotion*. New York: Oxford University Press.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359.

Rosenthal, D. (1990). A theory of consciousness. Report No. 40 on MIND and BRAIN, Perspectives in Theoretical Psychology and the Philosophy of Mind (ZiF), University of Bielefeld. A version of this paper is reprinted in (1997) N. Block, O. Flanagan, and G Güzeldere (Eds.), *The nature of consciousness: philosophical debates* (pp. 729–753). Cambridge, MA: MIT Press.

Rosenthal, D. (1991). The independence of consciousness and sensory quality. In E. Villenueva (Ed.), *Consciousness* (pp. 15–36). Atascadero, CA: Ridgeview.

Rosenthal, D. (1993a). State consciousness and transitive consciousness. *Consciousness and Cognition, 2*, 355–363.

Rosenthal, D. (1993b). Higher-order thoughts and the appendage theory of consciousness. *Philosophical Psychology, 6*, 155–166.

Rosenthal, D. (2000). Introspection and self-interpretation. *Philosophical Topics, 28*, 201–233.

Rosenthal, D. (2004). *Consciousness and mind*. New York: Oxford University Press.

Rowlands, M. (2001). Consciousness and higher-order thoughts. *Mind and Language, 16*, 290–310.

Seager, W. (1999). *Theories of consciousness*. New York and London: Routledge.

Shear, J. (Ed.). 1997). *Explaining consciousness*. Cambridge, MA: MIT Press.

Siewert, C. (1998). *The significance of consciousness*. Princeton, NJ: Princeton University Press.

Stubenberg, L. (1998). *Consciousness and qualia*. Amsterdam: John Benjamins.

Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: MIT Press.

Van Gulick, R. (1995). What would count as explaining consciousness? In T. Metzinger (Ed.), *Conscious experience* (pp. 61–79). Schöningh: Imprint Academic.

Van Gulick, R. (2000). Inward and upward: reflection, introspection, and self-awareness. *Philosophical Topics, 28*, 275–305.

Weiskrantz, L. (2000). *Consciousness lost and found*. Oxford: Oxford University Press.

Part I

Defenders of higher-order theories

Chapter 2

Varieties of higher-order theory

David M. Rosenthal

1. Introductory

A touchstone of much modern theorizing about the mind is the idea, still tacitly accepted by many, that a state's being mental implies that it's conscious. This view is epitomized in the dictum, put forth by theorists as otherwise divergent as Thomas Nagel (1979: 174) and Daniel Dennett (1991: 132), that the appearance and reality of mental states coincide.

Traditionally this claim was cast not in terms of a state's being conscious, but in terms of one's being conscious of the state. Thus Descartes writes that "no thought can exist in us *of which* we are not conscious at the very moment it exists in us."[1] These remarks echo Aristotle's claim that, since thoughts take on the forms of their objects, thoughts actually have themselves as objects, from which he concludes that, whenever we perceive or think, we perceive that we do.[2]

Today we typically cast things not in terms of what states we're conscious *of*, but what states are conscious, *tout court*. But doubtless this is meant to capture the same phenomenon.[3] If an individual is in a mental state but is in no way whatever conscious of that state, we would not intuitively count it as a conscious state. So a state's being conscious consists of one's being conscious of it in some suitable way; perhaps, for example, we must be immediately conscious of it. It is this equivalence of a state's being conscious with one's being conscious of it in some suitable way that points toward a higher-order theory of what it is for a mental state to be conscious. We can explain a state's being conscious in terms of a higher-order state's being directed on that state because a state's being conscious consists in one's being conscious of it.

It is occasionally held that explaining a state's being conscious in terms of one's being conscious of that state is circular, since it explains consciousness in terms of consciousness (e.g., Goldman 1993: 366). But this is a mistake. We

understand what it is to be conscious of something independently of what it is for a state to be conscious. We are conscious of things when we sense them, perceive them, or have thoughts about them as being present to us. Circularity would threaten only if these kinds of state could not themselves occur without being conscious states. Not only does it beg the question against such a theory to assume that; we have ample independent reason, both from everyday experience and experimental findings, to hold that such states do occur without being conscious. To keep things clear, we can refer to being conscious *of* something as *transitive* consciousness and to a mental state's being conscious as *state* consciousness (Rosenthal 1990/1997).

There are, however, a variety of ways in which one might seek to characterize the higher-order awareness of our conscious states. We are aware of things by sensing them, perceiving them, and having thoughts about them; which of these is operative when we are aware of our conscious states? And is the higher-order awareness distinct from the target state one is aware of, or might that higher-order awareness be part of that state itself? Might a state's being conscious consist, moreover, simply in a disposition for a higher-order state to occur, rather than in the actual occurrence of that higher-order state? Finally, if being in a conscious state is a matter of one's being aware of oneself as being in the relevant state, might conscious states consist simply in that higher-order awareness, without any first-order target at all? It is these questions that I address in what follows.

2. Higher-order sensing

Sensing things is one way to be conscious, or aware, of those things.[4] So one possibility, is that we sense our conscious states. This appeal to inner sense[5] has considerable intuitive force. One reason such a model is inviting has to do with the intuitive immediacy that Descartes urged characterizes the way we're conscious of our conscious states. We seem to be immediately aware of the things we sense. Nothing seems subjectively to intervene between the things we sense and our sensing of them; so the things we sense seem always to be immediately present to us. Higher-order sensing readily explains these appearances.

In addition, state consciousness appears to be among the most basic mental phenomena; so sensing, which is doubtless the earliest mental function, both phylogenetically and ontogenetically, seems well-suited to explain it. Presumably any creature with conscious states also has the capacity to sense, and it might seem intuitively that creatures with conscious states need have no higher

mental capacity. And, even in creatures like ourselves, with higher mental functions, state consciousness seems intuitively closely tied to sensory functioning. It may even seem as though sensing is the only way we are transitively conscious of things. If so, inner sense is the only option for a higher-order theory.

Sensing always involves mental qualities, in virtue of which a creature senses various perceptible properties, and the qualities that occur when sensing is conscious present a formidable challenge for any theory of consciousness. When qualitative states are conscious, there is something qualitative that it's like for the creature to be in those states. The states seem subjectively to "light up." A theory of consciousness must give some account of what such lighting up amounts to as well as some credible story about how it arises.

Higher-order sensing may seem especially appealing in this connection. If the higher-order states in virtue of which qualitative states are conscious are themselves qualitative, perhaps that will help explain what it is for conscious qualitative states to light up and how they do. A qualitative state's lighting up would consist in one's being conscious of that state by way of a higher-order qualitative state.

But despite these apparent advantages, such a theory faces insuperable difficulties. It seemed that being aware of qualitative states by way of higher-order states that themselves exhibit mental qualities would explain the first-order states' lighting up. But it's unclear how such higher-order mental qualities might help, and even that there could be any such higher-order qualities. Higher-order qualities can't help explain the lighting up unless they are themselves conscious. But, if a state's being conscious consists in its being sensed, the higher-order qualities would be conscious only if there were, in turn, third-order sensations that sensed those second-order sensations. The threat of regress looms. Moreover, the only mental qualities we're ever conscious of are those of first-order conscious states. We're never conscious of distinct, higher-order qualities, even when we become aware of being conscious of the first-order states, as we do when we focus introspectively on those states.

An even greater difficulty arises about what those higher-order qualities might be. Perhaps they're the same in each case as the first-order qualities, so that conscious seeing, for example, occurs when we visually sense our first-order visual states. But that cannot be. Mental qualities are keyed to the properties they enable a creature to perceive. The range of qualities that characterize visual sensations, for example, resemble and differ in ways that make it possible to discriminate physical objects in respect of their perceptible color and spatial properties. These mental qualities plainly aren't the same as the perceptible properties to which they enable visual access. Whatever the property an

object has in virtue of which we call it red, that property cannot be the same as any property of the visual sensations by means of which we sense objects as being red.

So, if we do actually sense our visual sensations, it must be that the higher-order mental qualities in virtue of which we do so are distinct from the first-order mental qualities that enable visual access to colored physical objects. But it's wholly unclear what qualities these might be. Nor is it clear how unknown qualities could help explain what it is for conscious qualitative states to light up. Higher-order sensing cannot make good on its promise to do justice to the intuitive immediacy and basic mental character of consciousness.

3. Higher-order perceiving

Sensing has no conceptual content. But perceiving not only has qualitative character, as does sensing, but has conceptual content as well; so it's convenient to think of perceiving as conceptualized sensing. Just as its conceptual content distinguishes perceiving from mere sensing, so its qualitative character distinguishes it from mere thinking. No mental quality need occur if I merely think that there is a red object in front of me, whereas perceiving a red object involves some distinctive mental quality.

So the difficulty about higher-order qualities also undermines the hypothesis that we are aware of our conscious states by perceiving them, rather than just sensing them. Perception always exhibits mental qualities, and there are no suitable higher-order qualities for higher-order perception to exhibit. So there are no higher-order perceptions, properly so called.

Still, the absence of higher-order qualities may seem to be more decisive against higher-order sensing than against higher-order perceiving. Qualitative character is the only distinctively mental way we have to characterize sensing; so without higher-order qualities there is no higher-order sensing. Perceiving, because it exhibits both conceptual content and qualitative character, has mental properties apart from its mental quality. So it's open to argue that, despite the absence of higher-order qualities, perceiving still provides the best model of the higher-order awareness we have of our conscious states. Thus William Lycan writes that a theory can appeal to higher-order perceiving without claiming that such higher-order awareness "is like external perception in every single respect" (1996: 28; cf. Lycan, this volume: §5).

Since higher-order awareness of conscious states cannot resemble perceiving in respect of mental quality, it must be in respect of its other mental prop-

erties, which are intentional. So to sustain higher-order perceiving, we must show that the higher-order awareness is, in ways that matter, more like perceiving than like thinking. Lycan advances a number of respects in which he believes this is so.[6] For one thing, he urges, we cannot have a thought about something without already being aware of it; so we would have to be perceptually aware of our mental states before having thoughts about them. But we have thoughts about many things we have never perceived. Since mental states are arguably just the kinds of things we cannot perceive, they are doubtless examples of this.

Lycan also argues that first-person appearances support a perceptual model. The phenomenology of being in conscious states, he urges, represents those states as present to us, and we seem able to attend to those states in the way we can attend to things we perceive. But we attend to objects of thought no less than those of perception. And thoughts can also represent things as present to us. Perceiving does always represent things as being to some degree present, whereas thoughts often do not; so higher-order awareness resembles only some cases of thinking, whereas it resembles all cases of perceiving. But that doesn't show that it's more like perceiving than like the relevant cases of thinking.

Lycan urges that we have considerable voluntary control over which perceptual states in our sensory fields are conscious, and that such voluntary control is more characteristic of perceptual awareness than of the awareness that comes from thinking about things. One might doubt that we have all that much voluntary control over our higher-order awareness. But, whatever the case about that, we can also direct and focus our thought processes, perhaps even more readily than we can our perceiving.

Lycan argues that our higher-order awareness monitors what states we are in, as perceiving does with various physical properties of things; it's less obvious, by contrast, that we could monitor things by having thoughts about them. Perceptual monitoring relies on sense organs, or other suitable mechanisms in the case of enteroceptive and proprioceptive perceiving, to discern how things are. If higher-order awareness were like perceiving, it too would require some such mechanism, which presumably would be a matter of cortical connections. And such connections could subserve such monitoring equally well if, independently of this issue, such higher-order awareness were more like thought than like perception.

A perceptual model, Lycan maintains, can capture the way our awareness of our mental states comes in degrees, which a model based on thoughts can't. Lycan's examples suggest that he actually has in mind degrees of attention; we attend more to some states than to others. But thoughts vary in how focused

and attentive they are no less than perceptions. Lycan claims that higher-order awareness is like perceiving in that we regard it as a reliable source of information. But that holds at least as well for the thoughts we have about small numbers, simple shapes, logical connections, and the everyday behavior of commonsense objects.

Any higher-order theory must explain how it is that we can become aware of exquisitely fine-grained differences among our qualitative states. This looks problematic for a higher-order theory based on thoughts, since we arguably don't have concepts corresponding to all the different mental qualities we can be conscious of. Lycan urges that higher-order awareness resembles perceiving in this respect, since perceiving is dedicated to discerning fine differences in perceptible properties.

But thinking distinguishes among properties in at least as fine grained a way as does perceiving. And, though we plainly don't have distinct concepts for all our conscious mental qualities, that doesn't show that we lack the conceptual resources needed to capture all those qualitative variations, since we can readily capture them using comparative concepts. We describe shades of red, for example, as more or less yellowish, lighter or darker, more or less saturated, and the like. And we do this with mental qualities no less than with the perceptible properties those qualities enable us to perceive. Our conceptual resources are, after all, sufficient to make us aware of all the fine-grained variations among our conscious mental qualities.

There is compelling experimental evidence that we actually are aware of our mental qualities in respect of such comparative aspects. We're aware of far more fine-grained differences among mental qualities when they occur together than when they occur one at a time (see Raffman 1995). It might be argued that this tells against any higher-order theory. Why wouldn't the higher-order states in virtue of which we're conscious of mental qualities be independent of whether qualitative states occur together or not? And if they are, we would be conscious of qualitative states with the same fineness of grain whichever way they occur.

But there is no basis for assuming that we are conscious of our mental qualities in the same way whatever qualities accompany them. And this result provides evidence that we aren't, but rather are conscious of mental qualities at least partly in comparison with one another, as exemplifying particular mental qualities to a greater or less degree than some accompanying quality. And this in turn suggests that our higher-order awareness of mental qualities are thoughts, since thoughts are more versatile than perceptions in characterizing things comparatively.

Lycan urges that we have purely recognitional concepts (see Loar 1997), which apply to our sensations not by way of ties with other concepts, but solely in virtue of one's ability to recognize what type of sensation one has. And he urges that a perceptual model will better accommodate such concepts than a model based on thoughts.

But it's unlikely that any concepts for qualitative states are purely recognitional in this way. Rather, our concepts for mental qualities connect in crucial ways with our concepts for the physical properties that those mental qualities enable us to perceive; indeed, our concepts for mental qualities very likely derive from our concepts for perceptible properties. We don't have individual concepts for each type of mental quality, nor for each corresponding perceptible property. Still, we can conceptually single out every mental quality by its location within a suitable quality space, just as we can for the corresponding perceptible property. We distinguish properties of both types as being more or less similar to and different from other properties in their quality space. So our concepts for mental qualities are intimately tied both to concepts for other mental qualities and to concepts for perceptible properties (Rosenthal 1999a, 1999b, forthcoming: Ch. 7).

Lycan's arguments rely on our folk-psychological conceptions of the relevant mental functioning. But there is also a folk-psychological consideration that suggests that thoughts are a better model for higher-order awareness than perception. The things perceiving gives us access to are physical objects or states of affairs; we perceive red objects, the growling of our stomachs, damage to our bodies, and the positions of our limbs, and we do so by way of characteristic sensations. But it is folk-psychologically odd also to speak of perceiving those sensations. No such oddness attends the idea that we have thoughts about those sensations and feelings; we can have thoughts about anything.

Folk psychology aside, the perceptual model has caused considerable theoretical mischief. Hume's famous problem about the self is due to his never "perceiv[ing] any thing but the perceptions" (1978/1939: 634); he assumes that any awareness of the self would be perceptual (Rosenthal 2004). And some, like John Searle (1992: 97), deny we are ever conscious of our mental states at all because they assume such awareness would have to be perceptual.

On balance, then, neither folk psychology nor empirical findings sustain higher-order perceiving, and the initial problem about higher-order qualities remains. The perceptual model, it seems, does not withstand scrutiny.

4. Dispositional higher-order thoughts

Sensing and perceiving are not, however, the only ways we are conscious of things. We are also conscious of things by having thoughts about them as being present. If I think, independently of any sensory input, of an object as being present, that's a way of being conscious of it. The requirement that the thought represent the object as being present echoes the ordinary case of perceiving, but no perceptual input is needed.

We can conclude, then, that our higher-order awareness of conscious states is in some way a matter of having thoughts about those states. Still, there are different versions of theories that appeal to such higher-order thoughts (HOTs).

Theorists who otherwise have little use for higher-order theories sometimes acknowledge that HOTs may be suitable to explain introspective consciousness (e.g., Block 1995: 235). We are introspectively conscious of a state when we are not simply aware of that state, but aware of it in a deliberate, attentively focused way. Ordinary, nonintrospective consciousness, by contrast, occurs when one is aware of being in the state but not in this deliberate, attentively focused way. Because we consciously focus on introspected mental states, introspection involves actually being aware that we are conscious of those states.

So it's reasonable to explain being introspectively conscious of a state not as simply having a HOT about that state, but as having a HOT about it that is itself conscious. And if a state isn't conscious, introspectively or otherwise, there is no HOT. That suggests that a state's being conscious in the everyday, nonintrospective way results from something in between these two.

The natural conclusion is that a mental state's being nonintrospectively conscious consists in its being accompanied by a HOT when that HOT is not, itself, a conscious thought. When no HOT occurs, the target state isn't conscious; when there's a conscious HOT, the target is introspectively conscious. So when a HOT occurs that isn't conscious the target is conscious but not introspectively so.

A HOT model seems especially appealing for introspective consciousness because HOTs are then conscious; so we are aware of their occurrence. Since we aren't conscious of HOTs except when we introspect, the model is subjectively less inviting for nonintrospective consciousness; if HOTs aren't conscious, it seems subjectively that none occur. So it might seem that we must explain nonintrospective consciousness by appeal to something other than HOTs, or indeed higher-order awareness of any sort.

But there is another alternative. States that are conscious but not introspectively so fall in some way between introspectively conscious states and states that aren't conscious at all. So we need something that falls between having a conscious HOT and having no relevant HOT. Still, we needn't countenance nonconscious HOTs for the intermediate case. Perhaps a state is nonintrospectively conscious if, rather than being accompanied by some HOT, it is simply disposed to be accompanied by a conscious HOT. This dispositional alternative makes room for the temptation to avoid positing HOTs that aren't conscious. And it fits with the pretheoretic idea that a state's being conscious is at least partly the dispositional properties of its being available for introspective access.[7]

Peter Carruthers (2000) has defended such a dispositional HOT theory, though not by appeal to the foregoing line of argument. He explicitly recognizes the possibility of nonconscious HOTs, but argues that the number of actual HOTs needed to capture all the phenomenological detail and subtle variation in our conscious experience at any moment would be prohibitively large. It is implausible, he claims, "that so much of our cognition should be occupied with formulating and processing the vast array of higher-order thoughts necessary to render our experience conscious at each moment of our waking lives" (2000: 221).

Carruthers recognizes that one could respond to this "objection from cognitive overload" by urging, with Dennett (1991: Ch. 11) that our conscious experience is not as rich in detail as it seems. But it's not clear that this reply helps. Dennett credibly argues that our sense of a rich, detailed phenomenological scene is illusory, since phenomenology suggests far more detail than we can actually discriminate. But the consciousness of our experience is a matter not of what we can discriminate, but of how our qualitative experience seems to us. And, as Dennett concedes, it does seems to be richly detailed.

Still, it's not easy to know how to assess the objection from cognitive overload. Cortical resources are far greater than needed for most, perhaps all, of our cognitive processing. So sufficient cortical resources are doubtless available to accommodate HOTs for all our conscious states. The consciousness of experience seems effortless, whereas cognitive processing requires some conscious effort and attention; so we would not have guessed that experience's being conscious calls for substantial cognitive resources. But we should be wary of letting such commonsense considerations affect our appraisal of a theory.

Carruthers urges that occurrent HOTs would make us conscious of all the fine-grained distinctions among our visual perceptions only if we had "a concept for each just-discriminable shade of colour" (2000: 222). But, as we saw

earlier, we are conscious of such discriminations largely in comparative terms; so we needn't have distinct concepts for each shade.

Like Lycan, Carruthers holds that our higher-order concepts for qualitative states are purely recognitional, and he sees this as a difficulty for a theory based on occurrent HOTs. Just as our perceptual awareness of physical green guides the application of our recognitional concept of that color, so our higher-order recognitional concepts, he claims, must also apply by way of some independent, nonconceptual awareness of the thing recognized. But a theory on which we are conscious of our qualitative states solely by way of occurrent HOTs allows no room for an independent, nonconceptual awareness that could guide the application of the recognitional concepts that figure in those HOTs. So the application of these concepts would be groundless on such a theory, and would not parallel that of our first-order recognitional concepts (this volume: § 3).

But such a parallel would be surprising if the relevant concepts were indeed purely recognitional. Some mental occurrence must guide the recognizing of things, and the only mental occurrence that might do that for the recognizing of physical colors is our sensations or perceptions of those colors. But sensations of green could guide not only our recognition of physical green, but also our recognition of the corresponding mental quality of green, which characterizes such sensations; one needn't in addition have some independent awareness of that sensation. Carruthers claims that reflection reveals that our recognition of sensations relies on some independent awareness of those sensations. But reflection is the unusual situation in which we have some third-order, introspective awareness, and in that special case our second-order awareness is indeed independent of our introspecting.

It's in any case unlikely, as noted earlier, that the concepts that figure in occurrent HOTs are purely recognitional. Rather, our concepts single out mental qualities by locating them within the quality space distinctive of the relevant sensory modality. And, since this mimics the way our concepts for perceptible physical properties locate those properties within a corresponding quality space, it captures the parallel that actually does obtain between the higher- and first-order concepts.

Carruthers favors a recognitional model of our higher-order awareness because he thinks it helps avoid difficulties about conscious qualitative states. But the appeal of a recognitional model doubtless also reflects the continuing hold of the traditional thesis that mind is transparent to consciousness. Since recognizing is factive, being aware of something by recognizing it can't go wrong. So, if our consciousness of our mental states is recognitional, that consciousness will automatically get things right. But consciousness does not always represent

our mental states accurately. Consciousness seems infallible because it never shows itself to be mistaken and it's tempting to think that there's no other way to know what mental states one is in. But consciousness is not the only way to determine what mental state one is in, and there is sometimes compelling independent evidence that goes against what consciousness tells us.

Carruthers argues that there could have been little evolutionary pressure for the routine generation of HOTs for all the thoughts and experiences in daily practical activity. True; but our daily experience is seldom if ever conscious in respect of all the detail relevant to such practical activity. Armstrong's (1980: 60) familiar example of a person driving long distance on psychological automatic pilot helps here. It is unlikely that the person's relevant visual states literally fail to be conscious; rather, they simply aren't conscious in respect of the rich detail needed for the task of driving. Occurrent HOTs would fail, in this case, to represent that rich detail. This often happens with activities that require concentrated effort; perceptual states that figure in an activity are seldom conscious in respect of all the detail relevant to carrying out the activity.

Indeed, it is in any case puzzling what evolutionary pressure there could have been for mental states to be conscious, whatever the explanation of their being conscious. Evolutionary pressure on mental functioning operates only by way of interactions that such functioning has with behavior. And mental functioning interacts with behavior solely in virtue of its intentional and qualitative properties. If a mental state's being conscious does consist in its being accompanied by a higher-order state, that higher-order state would contribute to the overall causal role, but this contribution would very likely be minimal in comparison with that of the first-order state. So there could be little adaptive advantage in states' becoming conscious.[8]

In any case, evolution often occurs without clear adaptive advantage; there are side effects whenever a feature is selected. Doubtless adaptive pressures favored high cortical mass, and that may have resulted in creatures with far more cortical resources than needed for everyday functioning. And such excess cortical area might then have come to subserve states that represent other cortical states. So creatures could well have come to have the actual HOTs needed for their mental states to be conscious, even though their being conscious is not itself especially adaptive.

Edmund T. Rolls (this volume: §2) has made a compelling case that HOTs might, in some kinds of case, make a pivotal contribution to overall adaptive value. Plans for action sometimes involve a sequence of first-order hypothetical thoughts, and it may well sometimes be useful to review these hypotheticals to evaluate and, if necessary, correct them. Such evaluating and correcting,

Rolls argues, would likely involve thinking about those first-order hypothetical thoughts, which would require HOTs that can reveal the syntactic connections among those thoughts. And these HOTs would make one conscious of the first-order thoughts.

This proposal reflects the pretheoretic idea that consciousness is somehow important for the critical evaluation of rational thinking. But the need to evaluate one's thinking critically is likely to produce evolutionary pressure for HOTs only if there is no easier way to get satisfactory results. And revising plans of action can result instead simply from having additional first-order thoughts about the relevant situations and the connections that hold among them. That mechanism for revision is especially plausible for our distant, less cerebral ancestors, on whom such evolutionary pressures could have had an effect. Nor does the revising of plans require identifying difficulties in one's first-order thinking by having HOTs about that thinking; even nonconscious conflicts often cause mental tension that in turn elicits compensatory adjustments in our thoughts and plans.

Carruthers's arguments against actual HOTs to one side, it's not obvious that dispositions for HOTs can help. The principle advantage of higher-order theories is that they explain how it is that we're transitively conscious of our conscious states. But being disposed to have a thought about something doesn't make one conscious of that thing, but only potentially conscious of it. A state's being conscious in the everyday, nonintrospective way does dispose one to be introspectively conscious of that state. But a higher-order theory must also explain ordinary, nonintrospective consciousness. Nor does a dispositional theory have any advantage in explaining why a nonintrospectively conscious state disposes one to introspect that state; we can expect that an occurrent HOT that isn't conscious disposes one to become conscious of that HOT.

It's also not initially obvious how dispositions could help with the difficulty about cognitive overload. Dispositions are themselves states, no less than actual HOTs; so dispositions to have HOTs should be no less cognitively demanding than actual HOTs. Again, it may be only the absence of conscious cognitive effort that makes dispositions seem intuitively preferable.

Carruthers's answer to both worries, about becoming conscious of conscious states and avoiding cognitive overload, appeals to the particular theory of intentional content he endorses. A state's intentional content, he urges, is in part a matter of what other mental states and what behavior that state is disposed to cause. This causal potential takes a special form when a state is suitably connected to a psychological subsystem capable of discriminating among types of states, which Carruthers thinks of as the mind-reading system.

A state then has the potential to cause in that mind-reading system a HOT about the state itself. And, by itself, that causal potential gives the state a certain higher-order content, in addition to its ordinary content in virtue of which it represents features of nonmental reality. Conscious experiential and intentional states thus have dual content; an experience of red, for example, will also have the content *seems red* or *is an experience of red*. And it will "have these [higher-order] contents categorically, by virtue of the *powers* of the HOT consumer system, in advance of any HOT being tokened" (2000: 242). This higher-order content comes automatically with the availability of the states to the mind-reading system.

Cognitive overload is thus avoided, Carruthers urges, since generating this higher-order content makes no cognitive demands beyond those incurred in the states' simply having first-order contents and their being connected to the mind-reading system. And these higher-order contents explain how it is that we're aware of our conscious states, as Carruthers puts it, how states "can come to acquire the properties of *subjectivity* and *what-it's-likeness* distinctive of phenomenal consciousness" (2000: 242; Carruthers's emphasis).

One might prefer a theory of consciousness not to be hostage to such a heavy load of other controversial views. But that aside, there are difficulties for this account. Each first-order state has some higher-order content in virtue of the state's availability to the mind-reading system, and that higher-order content results in one's being conscious of the state, so that there's something it's like for one to be in the state. But, if it's sufficient for a state to be conscious simply that it have some particular first-order content and that it be available to the mind-reading system, it's unclear how that state, or any state of its type, could ever occur without being conscious. It cannot be its first-order content is sometimes different, since we type states by way of their content. And, in any case, every suitably connected first-order state will have such higher-order content. Nor can it be that the mind-reading system sometimes shuts down, since other states remain conscious.

So it must be that individual states are sometimes available to the mind-reading system and sometimes not. But it's implausible that such shifts in availability would occur, since the first-order states themselves presumably occur in systems with fairly stable connections to the mind-reading system. And making such shifts would doubtless be at least as cognitively demanding as simply having occurrent HOTs. A theory cast in terms of occurrent HOTs, by contrast, has no difficulty with a state's shifting between being conscious and not; occurrent HOTs presumably come and go, just as other occurrent thoughts do.

Intentional content, on Carruthers's view, is a matter of a state's causal potential in respect of other mental states and behavior. The specifically higher-order content in virtue of which a state is conscious consists in that state's having the causal potential to affect something external to itself, namely, the mind-reading system.

But Carruthers acknowledges that conscious first-order states "have these [higher-order] contents categorically, ... in advance of any HOT being tokened." It is a state's having this higher-order content which explains there being something it's like for one to be in that state, and thus that state's being conscious. So it is the actual occurrence of a state with higher-order content, not merely a disposition to have a higher-order state, that explains a state's being conscious. As with most theories of content, it's a dispositional matter according to Carruthers that these states have their higher-order content, but it's an occurrent state with higher-order content that explains consciousness. It's the theory of content that's dispositional, not the theory of consciousness.[9]

Conscious first-order states have, Carruthers holds, higher-order content "in advance of any HOT being tokened." What, then, would an occurrent HOT add? Since first-order states are already nonintrospectively conscious without any occurrent HOT, the only possibility is that the an occurrent HOT would result in the first-order state's becoming introspectively conscious. But when we introspect, we are aware of our being conscious of the introspected state; so the occurrent HOT is itself also conscious. Despite Carruthers's recognition that HOTs need not be conscious, the only role he leaves open for occurrent HOTs is the role conscious HOTs have in introspection.

5. Intrinsic higher-order thoughts

If a state is conscious in virtue of some higher-order awareness of that state, it's natural to assume that this higher-order awareness consists in the occurrence of a distinct state with suitable higher-order content. Absent some reason to the contrary, distinct mental functions call for distinct states. And distinct higher-order states also explain how states shift between being conscious and not, which would be puzzling if the higher-order were built into the first-order target itself.

Nonetheless, a number of theorists have urged, following Franz Brentano, that the higher-order content in virtue of which we are conscious of our conscious states is intrinsic to those states (Gennaro, this volume, 1996; Kriegel 2003; Natsoulas 1999). As Brentano put it, all mental acts "apprehend [them-

selves] incidentally" (1973/1874:128). Every conscious state, on this view, is about itself as well as some nonmental reality.

Carruthers's view actually seems to be a version of such an intrinsic theory. The higher-order content in virtue of which a state is conscious, on his view, belongs categorically to the very state that is conscious. Though the higher-order content is a matter of the state's having connections with the mind-reading system, that content is a property of the target state one is conscious of, not a distinct higher-order state.

When a state is conscious, we are conscious of that state. But, except for the special case of introspective consciousness, we are not also conscious of being conscious of the state; it seems subjectively that we are conscious of only one state. So, if one relied on consciousness to reveal mental functioning, one would conclude that, when a state is conscious, there aren't two states but only one. This consideration has led some, such as Brentano, Karen Neander (1998), and Joseph Levine (2001:105, 168), to reject a higher-order theory altogether. But it has led others, such as Gennaro (this volume) and Natsoulas (1999), to argue instead that the higher-order content in virtue of which a state is conscious is intrinsic to that very state.

This argument assumes that consciousness reveals everything about our mental functioning, or at least everything relevant to the issue at hand. But we know that this isn't so, since there are many mental states that aren't conscious. Consciousness does reveal the phenomenological data that a theory of consciousness must do justice to. But to save these phenomena, we need only explain why things appear to consciousness as they do; we need not also suppose that these appearances are always accurate. A theory of consciousness explains how things appear to consciousness.

When a state is conscious, one is conscious of that state, though not typically aware of being conscious of it. But that doesn't support a theory on which only a single state occurs, nor does it undermine a higher-order theory. Rather, it shows only that the higher-order awareness of the state one is conscious of is rarely conscious. We needn't suppose that there's only one state, but merely that there's only one conscious state.

There is a related objection to the view that the higher-order awareness is due to a distinct state, an objection that does not rely on how things appear to consciousness. If the higher-order awareness is distinct from its target, the two might occur independently. It's theoretically unproblematic if the target occurs without the higher-order awareness; then the target simply isn't conscious. But, if the states are distinct, the higher-order awareness might also occur without any target. And, even if there's a target, the higher-order awareness might fail

to represent it accurately. Since it might seem that there is no good answer to how things would be phenomenologically in these kinds of case, perhaps this possibility tells against a higher-order theory (Levine 2001: 108; Neander 1998: 420).

One might advance an intrinsic theory as a way to meet this difficulty (Gennaro, this volume; Kriegel 2003). If the higher-order awareness is actually part of its target, it plainly can't occur without the target. And perhaps it also cannot then misrepresent that target.

It's not obvious that absent or misrepresented targets actually lead to any difficulty, phenomenologically or otherwise (see §6). But an intrinsic theory would in any case have no advantage in meeting whatever difficulties do occur. The distinction between an absent target and a misrepresented target is in an important way arbitrary. Suppose my higher-order awareness is of a state with property *P*, but the target isn't *P*, but rather *Q*. We could say that the higher-order awareness misrepresents the target, but we could equally well say that it's an awareness of a state that doesn't occur. The more dramatic the misrepresentation, the greater the temptation to say the target is absent; but it's plainly open in any such case to say either. The two kinds of case, moreover, should occasion the same kinds of phenomenological perplexities, if any. A higher-order awareness of a *P* state without any *P* state would be subjectively the same whether or not a *Q* state occurs. The first-order state can contribute nothing to phenomenology apart from the way we're conscious of it.

On an intrinsic theory, we do have a nonarbitrary reason to say that the target is never absent; whatever state the higher-order awareness is part of counts as the target. But it's not on that account obvious why misrepresentation could not occur. Simply being intrinsic to the target does not guarantee that the higher-order awareness will be accurate. And, since misrepresentation would occasion the same difficulties, if any, that would occur for an absent target, an intrinsic theory has no advantage in this connection over a theory on which the higher-order awareness is distinct from its target.

On Carruthers's dispositional theory, our higher-order awareness is intrinsic to its target, since that target has higher-order content in virtue of its tie to the mind-reading system. That tie enables the mind-reading system to "generat[e] recognitional concepts of experience, riding piggy-back on the first-order contents of experience" (2000: 241). But that tie between target and mind-reading system cannot protect that higher-order content against error. However accurate the mind-reading system may be, the connection any particular state has to that system can go wrong, and with it the mind-reading system's capacity to recognize that state and hence the state's higher-order con-

tent. A dispositional theory has no advantage in ensuring accuracy of higher-order content.

We're conscious of our conscious states in a way that seems, subjectively, to be direct and unmediated. And one might urge that this supports an intrinsic theory, since nothing would mediate between the higher-order awareness and its target if that awareness is actually part of the target. But the datum we need to explain is not actual immediacy, but rather subjective immediacy. And these needn't go together, at least not unless we assume that mental functioning is transparent to consciousness. It could be that nothing actually mediates between our higher-order awareness and its target even though it subjectively seems as though something does.

More likely, something might mediate even though we are subjectively unaware of anything doing so. Typically we aren't conscious of any HOTs; it doesn't subjectively seem as though there are any. So, if we're conscious of our conscious states by way of of distinct HOTs that aren't conscious, it will seem as though nothing mediates between our consciousness of those states and the states themselves. And, when those distinct HOTs are conscious, it will still seem as though nothing mediates between them and their targets if we're conscious of no inference from which the HOTs seem to arise. The HOTs will then seem subjectively to be spontaneous and uncaused. A theory that appeals to distinct HOTs can thus do justice to phenomenological immediacy, whereas an intrinsic theory must somehow explain why the actual lack of mediation issues in the corresponding appearance. It is likely that an intrinsic theory will appeal here to the very same considerations as a theory cast in terms of distinct HOTs.

Not only do the advantages claimed for an intrinsic theory fail to hold up; such a theory faces several important challenges. For one thing, it must explain why, if the higher-order awareness is intrinsic to the target, that higher-order awareness is about the target only in respect of its other, first-order content.[10]

An intrinsic theory must explain what happens when a state goes from being nonintrospectively conscious to being introspectively conscious. Does the target state come to have third-order content in virtue of which one comes now to be aware of its second-order content? Or does one instead become aware of that second-order content by way of a distinct third-order state? Neither answer is theoretically satisfying. It would be surprising if one's becoming introspectively conscious of a state were a matter of that state's actually changing, by taking on new, third-order content. But, if a distinct third-order state is responsible for introspective consciousness, why not a distinct second-order state for ordinary, nonintrospective consciousness?

The need to explain introspection recalls the initial point that an intrinsic theory may have difficulty giving an informative explanation of how states shift between being conscious and not. On an intrinsic theory, a state's coming to be conscious consists in its actually changing its content, and its ceasing to be conscious consists in a loss of intrinsic content. It would be theoretically more satisfactory to assume that such shifts leave the state unchanged.

Underlying all these issues is a pressing need for an independent way of individuating states, which doesn't beg the question between intrinsic and extrinsic higher-order awareness. It might seem that individuating mental states is inevitably arbitrary, and so wouldn't preclude either position. But there is a compelling consideration that undermines individuating in a way that makes the states of higher-order awareness intrinsic to their targets.

Intentional states differ not only in content, but in mental attitude as well. One can hold toward any particular content a variety of distinct attitudes, such as believing, desiring, anticipating doubting, wondering, and many others. And we individuate intentional states so that no state exhibits more than one mental attitude.[11]

But intentional states don't make one conscious of the things they are about unless those states exhibit a suitable attitude toward that content. Wondering, doubting, and desiring about something do not make one conscious of that thing. For an intentional state to make one conscious of the thing it's about, it must involve an assertoric attitude towards its content.

Consider a conscious case of an intentional state, such as wondering or doubting, which exhibit some nonassertoric attitude. That state is conscious in virtue of some higher-order awareness of it. Since we have ruled out higher-order sensing and perceiving, only some higher-order intentional content will do, and one must hold an assertoric attitude toward that higher-order content. So, when the conscious state is an intentional state with a nonassertoric attitude, the higher-order awareness must be distinct from its target itself, since we individuate intentional states so that no single state has two distinct attitudes. Considerations of mental attitude decisively undermine an intrinsic higher-order theory cast in terms of HOTs.

Brentano's version of an intrinsic theory is evidently perceptual; his prime example of a conscious state is hearing, and he argues that a conscious act of hearing apprehends itself (1973/1874: 128). And hearing something does make one conscious of it. But, as we saw earlier, all perceiving requires some mental quality, and the mental qualities special to audition enable us to hear only sounds, not auditory states themselves. Perceptual modality tells against an in-

trinsic higher-order perception theory just as mental attitude tells against an intrinsic HOT theory.

6. Higher-order thoughts and the intentional stance

The advantage of a higher-order theory is that it explains how it is that we're conscious of our conscious states. We are conscious of those states in virtue of some higher-order awareness. That higher-order awareness, moreover, explains not simply that we are conscious of those states, but also the particular way we are conscious of them. As with any case of being aware of something, the representational character of a higher-order awareness determines just how one is conscious of the state that the awareness represents.

This has implications for cases in which targets are absent or inaccurately represented. If I consciously take something I see to be a cow when it's actually a horse, phenomenologically it's as though I consciously see a cow. Similarly, if I am conscious of myself as being in a *P* state, it's phenomenologically as though I'm in such a state whether or not I am. If I'm not in a *P* state, that will make a difference to my overall mental functioning, just as it may make a difference to my interactions with my environment if the thing I take to be a cow is actually a horse. But the phenomenology is determined solely by the way I'm aware of things, whether perceived physical objects or my own mental states.

Typically we see things accurately, and it's also likely that consciousness ordinarily represents correctly what mental states we are in. But misrepresentation of such states can happen, (see, e.g., Nisbett & Wilson 1977), and it is an advantage of a higher-order theory that it accommodates such occurrences.

That consciousness sometimes represents us as being in states we have independent reason to think we aren't in is an important theme of Dennett's (1991). When we see wallpaper decorated with repeating, identical objects or Warhol's repeating, photographic portraits of Marilyn Monroe, it seems subjectively that we see all the repeating tokens at once with equal resolution. But that cannot be; we see most of those tokens parafoveally, and parafoveal resolution is far below that of foveal vision. Here as elsewhere, we're aware of our representations in ways that diverge from what we know those representations must be (1991: 354; see Ch. 11 *passim*).

In stressing this kind of occurrence, Dennett's discussion is congenial to higher-order theories. But there is another respect in which it isn't. Suppose I see somebody I don't know, but my memory of a friend causes me consciously to misperceive the person I see as my friend. Dennett famously urges that there

is no fact of the matter about whether the memory contaminates my current perception before or after that perception becomes conscious. On a higher-order theory, by contrast, there is precise moment when the perception becomes conscious, even if we can't easily determine when that is; it becomes conscious just when the higher-order awareness occurs.

Dennett's denial of a precise moment for the contamination seems to conflict with his recognition that consciousness sometimes mischaracterizes the way we represent things. If the way we are aware of a representation mischaracterizes that representation, there must be two distinct states, the representation and our awareness of it. But then the occurrence of this distinct awareness of a representation should provide the exact moment at which the representation becomes conscious.

But there is no conflict between Dennett's denial of an exact moment of consciousness and his recognition that consciousness sometimes mischaracterizes our representations. Dennett sees the representations that consciousness mischaracterizes not as the commonsense states of folk psychology, but as subpersonal events of content fixation. Though we are conscious of ourselves as being in various folk-psychological states, the representational states that actually occur are not those folk-psychologically taxonomized states, but merely those subpersonal events of content fixation. It is such subpersonal events, not folk-psychological states, which consciousness mischaracterizes when we seem to see many Marilyns at once with equal resolution, and similarly for other such cases.

These subpersonal events "are precisely locatable in both space and time" (1991:113). But that doesn't, according to Dennett, enable us to resolve the puzzles about timing, since he holds that those subpersonal events do not correspond to the folk-psychologically taxonomized mental states we're conscious of ourselves as being in. There are, strictly speaking, no first-order folk-psychological states, only subpersonal events of content fixation and our taking of ourselves to be in folk-psychologically taxonomized states (1991:Ch. 10). When I take myself to perceive somebody, I am in effect conscious of myself as perceiving that person (Rosenthal 2000). But there is no folk-psychological perceiving whose contamination by memory occurs at some particular moment, only the the interplay of subpersonal events, which do not correspond in any straightforward way to the folk-psychological states we interpret ourselves as being in.

We can thus see Dennett's view as a kind of higher-order theory on which there are acts of higher-order interpretation, but not their apparent first-order, folk-psychological targets. Thus Dennett holds that there is no way that things

seem apart from the way that they seem to seem, no "category of the objectively subjective – the way things actually, objectively seem to you even if they don't seem to seem that way to you" (1991: 132). The only seeming we can sensibly speak of is higher-order seeming.

This eliminativism in respect of first-order folk-psychological targets fits comfortably with Dennett's well-known view that intentional states are simply real patterns of behavior discernible from the intentional stance (1981). But that view aside, there seems little reason to deny that suitably grouped events of content fixation often constitute folk-psychological states. And if so, not only do HOTs occur, but also the first-order states those HOTs represent us as being in.

7. Distinct higher-order thoughts

The foregoing considerations all point to a higher-order theory cast in terms of distinct, occurrent HOTs (Rosenthal 1986, 1990/1997, 2002, forthcoming). Such a theory has all the advantages of other higher-order theories, and avoids their shortcomings.

Many of the difficulties that affect higher-order theories come from trying to determine the nature of the higher-order awareness by appeal to folk-psychological considerations or to how things seem subjectively. But we are rarely conscious of that higher-order awareness; so subjective impressions and folk psychology can have little useful to tell us. We can best think of that awareness as being a theoretical posit, designed to explain the phenomenological appearances of our conscious mental lives. It is success in that explanatory task that will establish the existence of that higher-order awareness and tell us about its nature. By that criterion, distinct, occurrent HOTs do the best job.

That HOTs are theoretical posits doesn't mean that we are never aware of them in an intuitively direct way; indeed, that's just what happens when we introspect. But we learn in the first instance that HOTs occur by developing a suitable theory to explain the subjective appearances of consciousness.

When qualitative states are conscious, there is something qualitative that it's like for one to be in those states; when they aren't conscious, there is nothing it's like. A higher-order theory must explain how being aware of a state results in there being something qualitative that it's like for one to be in such states. Higher-order sensing or perceiving initially seemed best suited to do so, since the qualitative character of such higher-order states might explain why being conscious of a state results in its lighting up qualitatively.

Since there are no higher-order qualities, that strategy cannot work. But higher-order qualities couldn't help in any case. A higher-order quality would explain why there is something qualitative that it's like for one to be in a particular state only if the higher-order quality is conscious. But, since higher-order qualities would be conscious only when we introspect, they couldn't help with nonintrospective consciousness, which is our principle concern. And when we do introspect, we aren't conscious of any such higher-order qualities, but only of the qualities of the first-order states. Higher-order qualitative states cannot explain why there is something qualitative that it's like for one to be in conscious, first-order qualitative states.

But it may seem even less likely that HOTs could help. HOTs, like other thoughts, have no qualitative properties; their only mental properties are intentional. How, then, could being conscious of a state by having a HOT about it result in there being something qualitative that it's like for one to be in that state?

It is important to be clear about just what would count as an answer to this question. One might hold that a answer would be satisfactory only it resulted in its tracing a purely rational connection between the occurrence of a HOT and there being something qualitative that it's like for one to be in the target state. Thus Levine argues that we can explain why water boils at 212°F at sea level in ways that make it inconceivable that it wouldn't, and we should demand no less for a satisfactory answer to the question about conscious qualitative character.

The explanation of water's boiling at 212°F involves, as Levine notes, a "rich elaboration of" chemical theory (2001: 79). But, as the history of chemistry testifies, such theory is hardly a purely rational matter, having to do only with what's conceivable. It's inconceivable, relative to chemical theory, that water would fail to boil at 212°F, but chemical theory is a matter of how things are, not just what we can conceive. Chemical theory is not only surprising and unexpected, but something we can conceive to be false.

If we had a parallel theory for qualitative states, we could doubtless explain why such states phenomenologically light up under certain conditions. And relative to that theory, it would seem inconceivable that it could have been otherwise. But can cannot demand that such a theory should antecedently seem rational, or a matter simply of what we can and cannot conceive. Like chemistry, that theory will come from examining how things actually are.

Once we see that such a theory could not be a matter simply of what's conceivable, we can develop, in advance of such a theory, a good theoretical hunch about how conscious qualitative character can be explained. And there is reason to think that HOTs actually do make a difference to what it's like

qualitatively for one to be in particular states, and hence that they may well be responsible for there being something qualitative that it's like.

Somebody who hasn't learned to distinguish the gustatory sensations of wines or the sounds produced by oboes and clarinets may well be conscious of these distinct mental qualities as though they were the same. And it sometimes happens that these mental qualities come then to seem distinct to such a person by that person's learning words for the distinct mental qualities. This is typically in a context in which one consciously concentrates on those mental qualities, so the new words are used for the mental qualities, not the physical stimuli, as Carruthers (2000: 240) and others have claimed. As one learns the new words, a subjective difference emerges between mental qualities that had previously been subjectively indistinguishable.

What might explain this change? Learning new words makes a difference only if one learns their meanings, which means learning to deploy the corresponding concepts in thoughts. And, since the concepts apply to mental qualities, the thoughts in which those concepts figure will be about mental states that exhibit those qualities; they will be HOTs. So learning how to have HOTs about states as exhibiting distinct mental qualities results somehow in the subjective emergence of a conscious qualitative difference between them; HOTs do make a difference to what it's like qualitatively for us to be in various mental states. And, if having new HOTs can make a difference in what it's like qualitatively for one to be in a particular state, having a HOT can presumably make the difference between there being nothing qualitative that it's like for one and there being something that it's like for one.

Carruthers claims that this account might be right about the causal antecedents of there being something it's like for one, but not about what constitutes there being something it's like (2000: 240). But it's unclear what kind of constituting Carruthers has in mind. The theory doesn't attempt a conceptual analysis of what it's like, any more than Carruthers's does. But the theory does argue that there is something qualitative that it's like for one to be in a state whenever one has a suitable HOT that one is in that state and the HOT characterizes the state in qualitative terms. If so, having such a HOT does constitute there being something qualitative that it's like for one to be in the state. The theory claims not that HOTs cause qualitative consciousness to occur, but that the having of suitable HOTs is what it is for qualitative states to be conscious.

Carruthers maintains that identifying the having of a suitable HOT with there being something it's like for one to be in the target state means giving up on a reductive explanation of qualitative consciousness (this volume: §4). But whatever one's views on reductive explanation, this is a theoretical identifica-

tion, supported both by data and theoretical considerations, and it thus carries substantial explanatory force.

It might be thought that new conscious qualities emerge not because we come to be conscious of existing qualities in new ways, but by what psychologists call perceptual learning. Learning new words might actually result in new mental qualities coming to occur, rather than our simply becoming conscious of mental qualities that were already there. But this is unlikely, since, unlike perceptual learning, the effect is dramatic and rapid.

If mental qualities were automatically conscious, new conscious qualities would imply new qualities, and perceptual learning would be the only possibility. But mental qualities need not be conscious. What it is to be a particular mental quality is a matter of its position in the quality space of the relevant modality. And we can construct such quality spaces by reference to the perceptual discriminations a creature can make (Rosenthal 1999a, 1999b), independent of whether the relevant qualitative states are conscious.

Levine objects that we "have a more determinate and substantive conception of [mental qualities] . . . , a conception that is not exhausted, or adequately captured by the formal description of a location in a similarity space" (2001: 107). Perhaps so; but, as with other commonsense phenomena, our conception of the properties of our qualitative states may well not be true to the actual nature of these properties.

HOTs explain how we're conscious of our conscious states. When those states are qualitative, HOTs characterize their qualitative properties in terms of how much they resemble and differ from other qualities in the relevant modality. So a particular quality often seems different to consciousness when it occurs with other qualities, since we can then be conscious of it in comparison with others. And, because HOTs result in our being conscious of ourselves as being in various qualitative states, the occurrence of a HOT about a qualitative state will be subjectively the same whether or not that state occurs.

By the same token, the subjective appearances will be the same whether the target causes the HOT or the HOT arises independently. Doubtless targets are often causally implicated in the occurrence of HOTs, though other factors must also figure, since otherwise the target would always cause a HOT and hence always be conscious. But what matters to the phenomenological appearances is simply what state the HOT makes one conscious of oneself as being in.

Since HOTs are distinct from the states they are about, it's possible for a HOT to occur without its target. And, since having a HOT results in one's being conscious of oneself as being in a particular state, a HOT's accompanying its target is subjectively indistinguishable from a HOT's occurring in the absence

of that target. When a target doesn't occur, one isn't in the state one is conscious of oneself as being in, but that's not a problem. All that matters to one's being in a conscious state is what it's like for one, and what it's like for one in that case is that one is in the state in question. Being in a conscious state is not being in that state and being conscious of being in it, but simply being conscious of oneself as being in the state.

Language is likely not necessary for thinking, and the conceptual resources for HOTs are not too demanding for infants and nonhuman animals to have them. A HOT has the content that one is in a particular state, a state we would characterize as mental, though the HOT need not do so. And the reference HOTs make to a self need involve nothing more than a distinction between oneself and everything else (Rosenthal 2004).

A higher-order theory is the only type of theory that can explain how it is that we're conscious of our conscious states. It seems that a higher-order theory based on distinct, occurrent HOTs can best accomplish that task.

Notes

1. 1964–1975: VII, 246 (*Fourth Replies*). Also: "the word 'thought' applies to all that exists in us in such a way that we are immediately *conscious of* it" (1964–1975: VII, 160 [Geometrical Exposition of *Second Replies*]). (Translations and emphasis mine.)

2. For the first claim, see 1984: II, *Metaphysics* XII, 7, 1072b20–23; for the second, see 1984: II, *Nicomachean Ethics* IX, 9, 1170a31–2; 1984: I, *de Anima* III, 2, 425b12–20.

3. The late 19th century shift to describing mental states as being conscious or not is very likely due to the growing recognition at that time that many mental states fail to be conscious. We try to describe things in a way that draws some contrast. Saying that we're conscious of all mental states draws a clear contrast, since we're not conscious of everything, whereas saying that all mental states are conscious leaves the intended contrast obscure, since we don't describe other comparable things as conscious.

4. I'll use 'conscious' and 'aware' and their cognates as equivalent.

5. The term derives from Kant (1998/1787: 174 [A22/B37]); Locke had earlier written of "internal Sense" (1975: 105 [II, i, 4]).

6. I appeal here to Lycan (this volume: §6), which he generously made available in advance of publication.

7. The relevant accessibility here is specifically introspective, as against the more general informational availability that underlies Dennett's contention that "[c]onsciousness is cerebral celebrity" (1993: 929). In that way it also differs from Block's notion of access consciousness, on which a state's content is "poised to be used as a premise in reasoning, . . . [and] for [the] *rational* control of action and . . . speech" (1995: 231).

8. So it's not a problem for higher-order theories, as Dretske (1995: 117) urges, if they can't explain such adaptive advantage.

9. As noted earlier, Carruthers claims that, because the concepts of mental qualities that figure in HOTs are purely recognitional, there must be an independent, nonconceptual awareness of the qualities such HOTs are about. His view provides for such an independent awareness, since it holds that the higher-order content that conscious qualitative states have in virtue of their connections with the mind-reading system is experiential content. Such content is "analog relative to a certain conceptual repertoire," since it "admit[s] of significantly *more* variations than there are concepts to classify them" (2000: 134; Carruthers's emphasis); so Carruthers holds that it isn't conceptual. Such experiential contents also allow for there to be distinct higher-order contents for each mental quality, which Carruthers believes is needed.

Carruthers denies that this is an inner-sense view, since he regards such analog content as a species of intentional content (2000: 232). But since analog content is experiential content, the view does appeal to higher-order perception (cf. Carruthers, this volume, esp. §3). And it in any case doubtful that experiential content is wholly nonconceptual.

I am grateful to Robert Lutzker for pressing me on a number of the issues discussed in this section.

10. Brentano actually assumed that the higher-order awareness was about the entire state, so that the higher-order content was also conscious (1973/1874: 129), though rarely attended to. But that content is actually seldom conscious.

11. As Brentano himself in effect observed, in noting that we individuate intentional states by reference to the mental act being performed (1973/1874: 127). Plainly two distinct attitudes would issue in two distinct mental acts.

References

Aristotle (1984). *The complete works of Aristotle*, ed. Jonathan Barnes. Princeton, Princeton University Press.

Armstrong, D. M. (1980). What is Consciousness? In David Armstrong, *The nature of mind*. St. Lucia, Queensland: University of Queensland Press, 55–67.

Block, Ned (1995). On a confusion about a function of consciousness. *The behavioral and brain sciences*, 18, 2 (June), 227–247.

Block, Ned, Owen Flanagan, and Güven Güzeldere (Eds.) (1997). *The nature of consciousness: Philosophical debates*. Cambridge, MA: MIT Press.

Brentano, Franz (1973/1874). *Psychology from an empirical standpoint*, Oskar Kraus (Ed.), English edition ed. Linda L. McAlister, tr. Antos C. Rancurello, D. B. Terrell, and Linda L. McAlister. London: Routledge & Kegan Paul.

Carruthers, Peter (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge: Cambridge University Press.

Carruthers, Peter (this volume). HOP over FOR, HOT theory.

Descartes, René (1964–76). *Oeuvres de Descartes*, Charles Adam and Paul Tannery (Eds.). Paris: J. Vrin.

Dennett, Daniel C. (1981). True believers: The intentional strategy and why it works. In A. F. Heath (ed.), *Scientific explanation*, ed. Oxford: Oxford University Press, 53–75; reprinted in Daniel C. Dennett, *The intentional stance* (13–35). Cambridge, Massachusetts: MIT Press/Bradford Books, 1987.

Dennett, Daniel C. (1991). *Consciousness explained*. Boston: Little, Brown and Company.

Dennett, Daniel C. (1993). The message is: There is no *me*dium. *Philosophy and phenomenological research* 53, 4 (December), 919–931.

Dretske, Fred (1995). *Naturalizing the mind*. MIT Press/Bradford Books.

Gennaro, Rocco J. (1996). *Consciousness and self-consciousness*. Amsterdam and Philadelphia: John Benjamins Publishing Company.

Gennaro, Rocco J. (this volume). Higher-order thoughts, animal consciousness, and misrepresentation: A reply to Carruthers and Levine.

Goldman, Alvin I. (1993). Consciousness, folk psychology, and cognitive science. *Consciousness and cognition*, 2, 4 (December), 364–382.

Hume, David (1978/1939). *A treatise of human nature*, L. A. Selby-Bigge (Ed.), rev. P. H. Nidditch. Oxford: Clarendon Press.

Kant, Immanuel (1998/1787). *Critique of pure reason*, Paul Guyer and Allen W. Wood (tr. and Eds.). Cambridge: Cambridge University Press.

Kriegel, Uriah (2003). Consciousness as intransitive self-consciousness: Two Views and an argument. *Canadian journal of philosophy* 33, 1 (March), 103–132.

Levine, Joseph (2001). *Purple haze: The puzzle of consciousness*. New York: Oxford University Press.

Loar, Brian (1997). Phenomenal states. In Block et al. (1997), 597–616

Locke, John (1975/1700). *An essay concerning human understanding*, Peter H. Nidditch (Ed.). Oxford: Oxford University Press.

Lycan, William G. (1996). *Consciousness and experience*. Cambridge, Massachusetts: MIT Press/Bradford Books.

Lycan, William G. (this volume). The superiority of HOP to HOT.

Nagel, Thomas (1979). What is it like to be a bat? *The philosophical review* 83, 4 (October 1974): 435–450; reprinted in Thomas Nagel, *Mortal questions* (165–179). Cambridge: Cambridge University Press.

Natsoulas, Thomas (1999). The case for intrinsic theory: IV. An argument from how conscious$_4$ mental-occurrence instances seem. *The journal of mind and behavior*, 20, 3 (Summer), 257–276.

Neander, Karen (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Philosophical perspectives* 12 (Language, mind, ontology), 411–434.

Nisbett, Richard E., and Timothy DeCamp Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review* 84, 3 (May), 231–259.

Raffman, Diana (1995). On the persistence of phenomenology. In Thomas Metzinger (Ed.), *Conscious experience* (293–308). Exeter, UK: Imprint Academic.

Rolls, Edmund T. (this volume). A higher order syntactic thought (HOST) theory of consciousness.

Rosenthal, David M. (1986). Two concepts of consciousness. *Philosophical studies* 49, 3 (May), 329–359.

Rosenthal, David M. (1990/1997). A theory of consciousness. In Block et al (1997), 729–853. First published as Report 40/1990, Center for Interdisciplinary Research (ZiF), University of Bielefeld.

Rosenthal, David M. (1999a). The colors and shapes of visual experiences. In Denis Fisette (ed.), *Consciousness and intentionality: models and modalities of attribution* (95–118). Dordrecht: Kluwer Academic Publishers.

Rosenthal, David M. (1999b). Sensory quality and the relocation story. *Philosophical topics*, 26, 1 and 2 (Spring and Fall), 321–350.

Rosenthal, David M. (2000). Content, interpretation, and consciousness. In Don Ross, Andrew Brook, and David L. Thompson (Eds.), *Dennett's philosophy: A comprehensive assessment* (287–308). Cambridge, Massachusetts: MIT Press/Bradford Books.

Rosenthal, David M. (2002). Explaining consciousness. In David J. Chalmers (ed.), *Philosophy of mind: Classical and contemporary readings* (406–421). New York: Oxford University Press, 2002.

Rosenthal, David M. (2004). Being conscious of ourselves. *The Monist* 87, 2 (April).

Rosenthal, David M. (forthcoming). *Consciousness and mind*. Oxford: Clarendon Press.

Searle, John R. (1992). *The rediscovery of the mind*. Cambridge, Massachusetts: MIT Press.

Chapter 3

Higher-order thoughts, animal consciousness, and misrepresentation

A reply to Carruthers and Levine

Rocco J. Gennaro

The higher-order thought (HOT) theory of consciousness has been defended most notably by David Rosenthal (1986, 1990, 1993, 2004, this volume) and also by myself (Gennaro 1993, 1996). It has also come under attack from several prominent authors over the past few years. In this chapter, I take the opportunity to defend the HOT theory against two of the most important objections raised in the recent literature. The first has to do with animal consciousness, and the second with the charge that HO theories cannot handle cases where the HO state misrepresents the lower-order state.

The standard HOT theory says that what makes a mental state conscious is the presence of an actual (i.e. occurrent) higher-order thought directed at the mental state. For a variety of reasons, I prefer the HOT theory to the higher-order perception (HOP) model, though I have argued that the difference is not as great or as important as has been portrayed in the literature (Gennaro 1996: 95–101; see also Van Gulick 2000). As will become clear in Section 2, I also hold a somewhat modified version of the Rosenthal's HOT theory.

1. Animal consciousness and the dispositionalist HOT theory (Carruthers)

1.1 Preliminaries

In his terrific book entitled *Phenomenal Consciousness: A Naturalistic Theory*, Peter Carruthers (2000) argues for an alternative "dispositionalist" version of the HOT theory of consciousness. I will critique Carruthers' dispositionalist

HOT (DHOT) theory as opposed to the more familiar "actualist" HOT model (AHOT), but my primary focus will be to reply to his treatment of animal consciousness. This section, then, can be thought of as a continuation of a previous exchange on this very topic (Carruthers 1989, 1992; Gennaro 1993, 1996; Carruthers 1999, 2000). I should first point out, however, that there is much that I agree with in Carruthers' book, such as his thoroughgoing "naturalist" approach to consciousness, his criticisms of so-called "mysterians" about consciousness, and his rationale for preferring a HO theory of consciousness to a first-order (FO) theory (e.g. Drestske 1995; Tye 1995). Nonetheless, I continue to disagree strongly with him on animal consciousness.

One of the most common sources of objection to the HOT theory comes from the concern that it rules out or, at least, renders unlikely the possibility of animal (and even infant) consciousness. Indeed, this objection is raised as a matter to routine in the literature (e.g. Dretske 1995; Seager, this volume), including various introductions to philosophy of mind (Kim 1996: 164–168). It must be taken very seriously, and so I have attempted to respond to it at length in previously published work (Gennaro 1993, 1996). For example, I have argued that the HOT need not be as sophisticated as it might seem to those who raise this objection. Since most of us believe that many animals have conscious mental states, a HOT theorist must explain how an animal can have the seemingly sophisticated HOTs necessary for conscious states. A simple general argument against the HOT theory might be put in the *modus tollens* form:

(1) If the HOT theory is true, then most animals do not have conscious experiences (because they are incapable of having the relevant HOTs).
(2) Most animals do have conscious experiences.
Therefore, (3) The HOT theory is false.

Most HOT theorists, such as myself, are concerned to show that premise 1 is false; that is, we try to show how the HOT theory does not rule out animal consciousness. Peter Carruthers, however, while himself a HOT theorist, rejects premise 2. So what is frequently viewed as an external criticism against the HOT theory is, in this case, an attack from within as far as I am concerned. Carruthers, in short, accepts the HOT theory *and*, without apology, the apparent consequence that animals do not have phenomenal consciousness. Put another way, Carruthers argues as follows:

(1) If the HOT theory is true, then most animals do not have conscious experiences.

(2) The HOT theory is true.
Therefore, (3) Most animals do not have conscious experiences.

While most HOT theorists would again reject premise 1, Carruthers believes that this argument is sound. I think it is important, however, to continue to defend the HOT theory against this objection regardless of its origin. We should reject the notion that a lack of animal consciousness follows from the HOT theory. In my view, doing otherwise can only weaken the theory and makes it very difficult to convince others of its truth.

1.2 Animal concepts: A reply

It is important first to be clear about Carruthers' somewhat idiosyncratic terminology. When he speaks of conscious mental states, he uses the term 'phenomenal' states and the Nagelian (1974) expression 'something it is like' to undergo such mental states. However, Carruthers also speaks of *nonconscious* 'feels' due to an ambiguity in the term 'feel':

> If animal experiences are not phenomenally conscious. . .then their states will lack *feel*. But if the pains of animals, too, lack feel, then doesn't that mean that animals don't feel them?. . .There is no real objection to HOR theory here. . .merely an ambiguity in the term 'feel'. . .The relational property of *feeling pain* can thus be understood in purely first-order and non-phenomenological terms, just as can the property of *seeing red*, [but]. . .we can (and should) deny that the pains which animals feel are phenomenally conscious ones. So we should deny that animal pains have subjective feels to them, or are *like anything* to undergo. . .In fact the idea of a feeling of pain which lacks *feel* is no more problematic than the idea of a percept of red which lacks *feel*. . . (2000: 200–201)

We therefore have the key term in the title of Carruthers' book; namely, 'phenomenal.' When he uses that term, he means the kind of conscious subjective feel that is familiar to each of us. He calls it "experiential subjectivity" as opposed to mere "worldly-subjectivity." Carruthers' then offers an account of phenomenal consciousness in terms of dispositional HOTs, as we shall see below. He also responds to my (1993, 1996) criticisms of his views on animal consciousness. He first presents the following summary statement of my position as follows: "[i]n order for [mental state] M to count as phenomenally conscious, one does not have to be capable of entertaining a thought about M *qua* M. It might be enough, [Gennaro] thinks, if one were capable of thinking

of M as *distinct from* some other state N." (2000: 195) Carruthers then offers the following reply:

> What would be required in order for a creature to think, of an experience of green, that it is distinct from a concurrent experience of red?. . .*something* must make it the case that the relevant *this* and *that* are colour experiences as opposed to just colours. What could this be? There would seem to be just two possibilities. [1] Either the *this* and *that* are picked out as experiences by virtue of the subject deploying. . .a concept of experience, or some narrower equivalent. . .On the other hand, [2] the subject's indexical thought about their experience might be grounded in a non-conceptual *discrimination* of that experience as such. . . (2000: 195)

Carruthers rejects both possibilities but neither reply is persuasive. He rejects possibility 1 mainly because "this first option just returns us to the view that HOTs (and so phenomenal consciousness) require possession of concepts which it would be implausible to ascribe to most species of animal." (2000: 195) But Carruthers has once again overestimated the sophistication of such concepts. He mentions concepts such as 'experience,' 'sensation,' and 'seeming red.' But why couldn't those animal HOTs simply contain concepts more like 'looking red' or 'seeing red'? Is it so implausible to ascribe *these* concepts to most animals? I think not. Animals need not have, for example, the concept of 'the *experience* of red' as opposed to just 'seeing or looking red.' "I am now seeing red" seems to be a perfectly good HOT.

Similarly, animals need not have HOTs containing the concept 'experience' in any sophisticated sense of the term, but why couldn't they have, say, the concept of 'feeling'? To use another example, perhaps animals do not have any sophisticated concept of 'desire,' but why not some grasp on the related notion of a 'yearning' for food. I fail to see why attributing such concepts to the lion chasing the deer would be so "implausible." Once again, perhaps most animals cannot have HOTs directed at pains *qua* pains, but why can't those HOTs contain the related indexical concepts 'this hurt,' or 'this unpleasant feeling'? Having such concepts will then also serve, in the animal's mind, to distinguish those conscious states from others and to re-identify those same types of mental states on different occasions. In addition, just as there are degrees of sophistication of mental concepts, so there are also degrees of "I-concepts" contained in the HOTs and much the same goes for an animal's ability to possess them. (Gennaro 1996: 78–84)

Carruthers then rejects possibility 2 mainly because "this second option would move us, in effect, to a *higher-order experience* (HOE) account of phe-

nomenal consciousness," (2000: 196) but he then defers any critical discussion of this alternative until the following chapter eight. I cannot fully address this topic here, but there is room for several brief replies: First, I have already argued (in Gennaro 1996: 95–101) that the difference between the HOT and HOE [= HOP] models is greatly exaggerated. Others have also questioned this traditional division of HO theories (Van Gulick 2000), and some have even argued that the HOP model ultimately reduces to a HOT model (Güzeldere 1995). Thus, Carruthers' criticism that my view might eventually "move us" to the HOP model is not as damaging as he seems to think. In other words, he has not really replied to my critique; instead, he has at best shifted the debate to the value of the HOT/HOP distinction itself.

Second, after deferring possibility 2 to chapter eight, Carruthers himself also questions the value of the HOP [= HOE] model over and above the HOT model: "The take-home message is: we would never have evolved higher-order experiences (HOEs) unless we already had higher-order thoughts (HOTs); and if we already had HOTs then we did not need HOEs. Upshot: if we are to defend any form of higher-order representation (HOR) theory, then it should be some sort of HOT theory. . ." (2000: 219) Whether or not Carruthers is right about this, it clearly does not address possibility 2 left over from his previous chapter. Moreover, he seems now to be dismissing the value of the HOP view in favor of the HOT model, without returning to his objection to my view. On the other hand, as Carruthers (this volume) makes even clearer, he understands his HOT theory to be a form of HOP theory, and so it is again difficult to see why any "move" in that direction would be so problematic for me.

Third, when Carruthers speaks of "non-conceptual discrimination" in, say, one's perceptual field, it seems to me that this is very misleading or, at least, rather ambiguous. It is thus open to a similar counter-reply to possibility 1. Such experiences may not include the ability to apply some concepts, but surely at least *some* other less sophisticated concepts are still required in order to have the experience itself. An infant or dog may not experience the VCR *as* a VCR, but they surely at least apply some concepts to that visual experience, e.g. black and rectangular. Similarly, an animal or infant may not be aware of its desire to eat *qua* concept of desire, but they can still be aware of that mental state in virtue of some other related concepts, e.g. 'yearning to eat something.' I believe that conscious experience always involves the application of some concepts, though I do recognize that this has become a hotly contested claim. I believe that the idea of non-conceptual experience is, at the least, a somewhat misleading and ambiguous notion. (For more on the literature on nonconceptual content, see Gunther 2003.)

Fourth, in my 1993 and 1996 replies to Carruthers, I was careful not to rely solely on the conceptual considerations he cites. I also put forth behavioral, evolutionary, and comparative brain structure evidence for the conclusion that most animals are conscious. For example, I explained how many lower animals even have some kind of cortex, not to mention the fact that they share with us many other "lower" brain structures known to be associated with conscious states in us. While Carruthers is very knowledgeable about brain science and discusses evolution elsewhere in his book, his failure to do so in the context of this disagreement is significant because the *cumulative* effect of such strong inductive evidence in favor of animal consciousness is lost. It is very puzzling why a thoroughgoing naturalist would not take such collective evidence to outweigh any considerations which may or may not follow from the HOT theory.

Finally, as Lurz (2002: 12–13) explains, it seems unlikely that the belief that many animals have conscious states can be so easily explained away, as Carruthers tries to do, merely as some anthropomorphic process of "imaginatively projecting" what it is for us to have certain mental states onto various animals. For one thing, it certainly does not describe what goes on in me when I attribute mental states to animals. I agree with Lurz that my main initial reason for believing that animals have conscious mental states has more to do with the fact that their behavior is best explained and predicted by attributing such folk psychological notions to them. In addition, as I mentioned above, such a conclusion is only further supported upon examination of the scientific brain evidence.

1.3 Brain structure and evolution

With regard to brain structure, Carruthers is mainly concerned to discredit the AHOT theory with what he calls the "objection from cognitive overload." (2000: 221). Indeed, this is Carruthers' main objection to the AHOT theory:

> . . .a major problem with the actualist version of HOT theory. . .[is] the implausibility of supposing that so much of our cognition should be occupied with formulating and processing the vast array of higher-order thoughts necessary to render our experience conscious at each moment of our waking lives. . . (2000: 221)

But, first, it is never made clear why AHOTs would take up so much "cognitive space" (i.e. neural mechanisms). No neurophysiological evidence is offered to show that our brains aren't "big enough" to handle the job. After all, it is not just the number of neurons in our brains, but also the numerous connections

between them. Second, it is unclear just how Carruthers' own DHOT theory would avoid this problem (assuming there is one), since even he acknowledges the presence of various *actual* brain structures (e.g. theory of mind mechanism) which are required to fill out the DHOT theory. For example, Carruthers is forced to acknowledge that his theory requires the presence of "something categorical [= actual] taking place in me whenever I have a conscious experience, on the [DHOT model] – the perceptual contents are actually there in the short-term memory store C, which is defined by its relation to HOT." (2000: 233) Rowlands (2001: 305–309) makes a similar point and then goes on to argue, convincingly in my view, that if these actual structures play such an important role in making mental states conscious, then the result is, in effect, to abandon the DHOT view. Dispositional states, on at least some interpretations, also require similar actual brain structure or "cognitive space." This is not of course to say, however, that I agree with Rowlands' unrelated criticisms of the AHOT theory.

Of course, part of the reason that Carruthers is so convinced by this argument against the AHOT theory has to do with his acceptance of what he calls the "richness of phenomenally conscious experience" (2000: 299ff.; cf. 224) which is often related to a belief in the nonconceptual content of experience mentioned earlier. This is a major topic that I cannot pursue here, except to note that I think a very strong case can be made for the conclusion that the phenomenally conscious aspects of our experiences are frequently much less "rich" than is commonly supposed (see e.g. Dennett 1991; Weisberg 1999). As Carruthers himself discusses, objects in the periphery of our visual field seem to lack the kind of rich determinacy that Carruthers seems to have in mind. But even within areas of one's visual focus (e.g. looking at a large painting), a case can be made that one's conscious attention is more fragmented than is often believed. Therefore, the HOTs in question again need not be as complex as Carruthers seems to think.

With regard to evolution, Carruthers tells us that he finds no evolutionary reason to suppose that actual HOTs be present in the case of conscious mental states: "What would have been the evolutionary pressure leading us to generate, routinely, a vast array of [actual] HOTs concerning the contents of our conscious experience?" (2000: 225). But, I suggest, that there are at least three good reasons overlooked by Carruthers: (1) On the AHOT theory, actual nonconscious HOTs, we may suppose, can more quickly become conscious HOTs resulting in *introspective* conscious mental states. On the AHOT theory, introspection occurs when a nonconscious HOT becomes conscious and is thus directed internally at another mental state. The ability for an organism to shift

quickly between outer-directed and inner-directed conscious states is, I believe, a crucial practical and adaptive factor in the evolution of species. For example, an animal that is able to shift back and forth between perceiving other animals (say, for potential food or danger) and introspecting its own mental states (say, a desire to eat or a fear of one's life) would be capable of a kind of practical intelligence that would be lacking otherwise. It seems reasonable to suppose that such quick transitions are more easily accomplished by having *actual* HOTs changing from being unconscious to conscious. (2) Even if we suppose that some lower animals are only capable of first-order conscious states (and thus only nonconscious HOTs), the evolutionary foundation has been laid for the yet more sophisticated introspective capacities enjoyed by those of us at the higher end of the evolutionary chain. Thus, the presence of actual (nonconscious) HOTs can be understood, from an evolutionary perspective, as a key stepping stone to the capacity for introspective consciousness. Such an evolutionary history is presumably mirrored in the layered development of the cortex. (3) Finally, as Rolls points out, having actual HOTs allows for the correction of plans that result from first-order processing. Rolls puts forth his own modified version of an AHOT theory, and suggests that "part of the evolutionary significance of this type of higher-order thought is that it enables correction of errors made in first-order linguistic or in non-linguistic processing." (Rolls 1999: 249)

1.4 The moral issue

Linking his discussion of animal consciousness back to moral issues, Carruthers explains that he had previously argued "that non-conscious pains – pains which would lack any subjective qualities, or *feel* – could not be appropriate objects of sympathy and moral concern." (2000: 203; cf. Carruthers 1989, 1992). But Carruthers has had a change of heart. He first imagines a conscious, language-using, agent called Phenumb "who is unusual only in that satisfactions and frustrations of his conscious desires take place without the normal sorts of distinctive phenomenology." (2000: 206; see also Carruthers 1999). Without becoming bogged down in the details of Carruthers' questionable thought experiment, he ultimately argues that Phenumb is an appropriate object of moral concern and that the example shows "that the psychological harmfulness of desire-frustration has nothing (or not much) to do with phenomenology, and everything (or almost everything) to do with thwarted agency." (2000: 207) In essence, Carruthers is attempting to separate desire frustration from consciousness in order to make room for the idea that ani-

mals can indeed be the objects of sympathy and moral concern, contrary to his previously held position. He explains that his "present view is that it is *first-order* (not necessarily-phenomenal) disappointments and frustrations of desire which are the most basic objects of sympathy and (possible) moral concern. And these can certainly be undergone by many species of non-human animal." (2000: 205)

I am frankly very puzzled by Carruthers' view here for several reasons. First, it seems to me that any *actual* organism (that we know of) capable of "desire frustrations" will also be capable of phenomenally conscious pains and would thereby also have the ability to *suffer*. Even the hypothetical Phenumb begins as a conscious agent. It seems to me that desire frustration is an even more sophisticated and intellectual psychological capacity than the mere ability to subjectively *feel* pains. Even if the two capacities are somehow *theoretically* distinct, I fail to see what positive reason we could ever have to attribute *only* the former to any known animal. Second, when Carruthers speaks of "desire frustrations," it is unclear to me how could they be *non*-conscious at all. I am not sure that I even understand the idea of a *non*-phenomenal "disappointment" or "desire frustration." As Cavalieri and Miller (1999: 3) put it, "[s]o nonhumans can be disappointed, can have their desires frustrated. . .but these disappointments and frustrations are non-phenomenal. This is just incoherent. . . Disappointments and frustrations are possible only for the sentient." Of course, there can be non-conscious desires (and even, pains), but it does not follow that there are non-conscious desire *frustrations*, especially in organisms who are supposed to be utterly (phenomenally) non-conscious.[1] Third, even if we can imagine the possibility of some Spock-like character only able to have such purely intellectual frustrations (as Carruthers suggests in his 1999: 478, Note 26), it does not follow that such frustrations would be entirely non-phenomenal. They may be devoid of the typical accompanying *emotions*, but there must, at minimum, be conscious *thoughts* and *beliefs* about the objects of those desires. Carruthers is now curiously and suddenly comparing (what he takes to be) non-conscious animals to a very sophisticated intellectual hypothetical character. Thus, in the end, I do not believe that Carruthers' current moral stance is any more tenable than his previously held view.[2]

1.5 Against the dispositional HOT theory

Although my main concern has been to reply to Carruthers on the topic of animal consciousness, I will conclude this section with a few critical observations

of his DHOT theory as such. (1) It seems to me that Carruthers is not always careful in the way that he phrases his theory. Perhaps best is when he says that "[i]n contrast with the actualist form of HOT theory, the HOTs which render M conscious are not necessarily actual, but potential. . .There need not *actually* be *any* HOT occurring. . ." (2000: 227) There is, of course, much more to Carruthers' theory and I cannot hope to do justice to all of the subtleties here (see especially his chapters eight and nine), but we should at least note the following elaboration of his view:

> We can propose that conscious experience occurs when perceptual contents are fed into a special short-term buffer memory store, whose function is (*inter alia*) to make those contents available to cause HOTs about themselves, where the causation in question will be direct, not mediated by any form of inference. (2000: 228)[3]

According to Carruthers, then, it would seem that the HOTs in question are not actual HOTs directed at a mental state M, but dispositional HOTs. However, there is significant ambiguity in the way that Carruthers phrases his view. Consider, for example, his slogan at the end of the book: "A disposition to get higher makes consciousness phenomenal." (2000: 329) Of course, like most slogans, it may necessarily be somewhat oversimplified and even misleading, but it seems to me that the problem runs deeper. The slogan might suggest that the disposition in question belongs to the *mental state M* instead of the HOT. What exactly has the "disposition to get higher"? If the DHOT is some form of HOT, then the HOT is, after all, already "higher" in some sense. This would seem to leave the *lower-order* state as having the disposition in question, which does not seem to be Carruthers' considered view nor does it seem to make very much sense. One might be inclined to dismiss this problem in the context of a mere slogan; however, in one of Carruthers' more formal statements of DHOT theory, he says the following:

> Any occurrent mental state M, of mine, is conscious = M is disposed to cause an activated belief (possibly a non-conscious one) that I have M, and to cause it non-inferentially." (2000: 227)

I find this definition unclear and, again, perhaps even at odds with what is supposed to be the DHOT theory. Carruthers is here clearly attributing the disposition in question *to M*, not to any higher-order state. M becomes conscious when *it* has a disposition to cause some kind of higher-order state directed at M. This not only contradicts other statements of the DHOT theory, but it is also

very difficult to understand on its own terms. At minimum, some clarification is needed.

(2) There is also a crucial distinction in the AHOT theory which seems lost, or at least unaccounted for, on Carruthers' DHOT theory. Indeed, in Gennaro 1993, I criticized Carruthers for conflating this distinction which, in turn, led him to some of his conclusions regarding animal consciousness. This well-known distinction, mentioned briefly earlier, is the difference between first-order (i.e. world-directed) conscious mental states and introspective (i.e. inner-directed) conscious states. On the AHOT theory, the former will be accompanied by nonconscious HOTs; in the latter case, there will be conscious HOTs accompanied by yet higher (third-order) nonconscious HOTs. Now this distinction is noticeably absent from Carruthers' alternative DHOT account, except perhaps for one very brief mention of third-order states (2000: 251–252). Very little is said about this critical difference in conscious mental states. It is therefore left unaccounted for and unclear on his DHOT theory. My sense is that Carruthers is once again conflating, or just ignoring, this important distinction. For example, Carruthers speaks of *focusing* "on my experience of a colour. . .and this is to focus on the subjectivity of my experiential state." (2000: 184) This suggests introspective consciousness, not just a first-order conscious color experience accompanied by an nonconscious HOT. One reason that this is so important is that this conflation also leads erroneously to the conclusion that animals cannot have conscious mental states. If, for example, one mistakenly supposes that the HOT theory requires *introspective* states to accompany first-order conscious states, then one may very well doubt the possibility of animal consciousness. In any case, Carruthers needs to answer the following questions: How does he explain the difference between first-order conscious and introspective conscious states on the DHOT model? Are the HOTs potential (or dispositional) HOTs only in the former case? If so, do they become actual conscious HOTs in the introspective case? If not, then how can the DHOT model account for the difference? Would there be an additional level of dispositional HOT in the introspective case? Whether Carruthers can answer these questions in a satisfying way without making major modifications to, or even abandoning, his DHOT theory is, in my opinion, very doubtful. Perhaps he can.[4]

(3) Part of the reason for Carruthers' attack on animal (and infant) consciousness has to do with his allegiance to the so-called "theory of mind" theory, whereby understanding mentalistic notions presupposes having a "folk-psychological" theory of mind (see Carruthers & Smith 1996). Once again, however, Carruthers builds a great deal into having such a theory and then

explicitly ties it to the capacity for having any HOTs at all. For example, he explains that "... the evidence is that children under, say, the age of three lack the concepts of *appearance* or *seeming* – or equivalently, they lack the idea of perception as involving *subjective* states of the perceiver – which are necessary for the child to entertain higher-order thoughts about its experiences." (2000: 202) But, once again, Carruthers seems to have in mind *conscious* HOTs, which are (once again!) *not* necessary for having first-order conscious states according to the AHOT theory. He also soon thereafter makes a similar point about autistic people who have been thought to be "mind-blind" in certain ways (Baron-Cohen 1995). But then Carruthers goes on to make the same error by telling us that "if autistic subjects are blind to their own mental states, then that will mean they are incapable of *self-directed* HORs; which in turn will mean that they lack phenomenally conscious mental states, if any form of HOR theory is correct." (2000: 202, emphasis added) If by "self-directed HORs" Carruthers means "introspective states," then he is plainly mistaken about what follows.

In line with many 'theory-theorists,' Carruthers also holds that animals with HOTs should be able to have HOTs about the mental states of *other creatures* as, for example, we might expect to find when animals engage in deceptive behavior. But even if some or most animals cannot engage in deceptive behavior and so do not have HOTs about the mental states of others, it still does not seem to follow that they cannot have HOTs about *their own* mental states (Ridge 2001). I therefore agree that Carruthers' view rests on the false assumption "that there could not be an agent capable of having HOTs about its own mental states but incapable of having HOTs about the mental states of others." (Ridge 2001: 333)[5] One might still believe, with Kant, that having HOTs presupposes some sort of implicit "I-thought" which distinguishes the thinker from outer objects (Gennaro 1996: 78–84, Ch. 9), but those outer objects need not always include the *mental states* of other conscious beings.

Overall, then, I believe that the AHOT theory remains the leading HOT theory, and it is consistent with the view that most animals have conscious mental states.

2. Misrepresentation and the division of phenomenal labor (Levine/Neander)

Joseph Levine (2001) raises an important objection to all higher-order theories of consciousness. He credits Karen Neander (1998) for an earlier version of this type of objection under the heading of the "division of phenomenal la-

bor." The main idea behind the objection is that when "we are dealing with a representational relation between two states, the possibility of misrepresentation looms." (Levine 2001: 108; cf. Neander 1998: 418ff.) Levine then argues that the HOT theory cannot explain what would occur when the higher-order state misrepresents the lower-order state. The main example used is based on color perception, though the objection could presumably be extended to other kinds of conscious states. Levine says:

> Suppose I am looking at my red diskette case, and therefore my visual system is in state R. According to HO, this is not sufficient for my having a conscious experience of red. It's also necessary that I occupy a higher-order state, say HR, which represents my being in state R, and thus constitutes my being aware of having the reddish visual experience. . .Suppose because of some neural misfiring (or whatever), I go into higher-order state HG, rather than HR. HG is the state whose representation content is that I'm having a greenish experience, what I normally have when in state G. The question is, what is the nature of my conscious experience in this case? My visual system is in state R, the normal response to red, but my higher-order state is HG, the normal response to being in state G, itself the normal response to green. Is my consciousness of the reddish or greenish variety? (Levine 2001: 108)

Levine (2001: 108) initially rightly points out that we should reject two possible answers, each of which is very problematic (cf. Neander 1998: 420ff):

Option one: The resulting conscious experience is of a *greenish* sort.
Option two: The resulting conscious experience is of a *reddish* sort.

Options one and two are both arbitrary and poorly motivated. Even worse, option one would make it seem as if "the first-order state plays no genuine role in determining the qualitative character of experience. . ." (Levine 2001: 108) The problem here is that one wonders what the point of having both a lower and higher-order state would be if only one of them determines the conscious experience. On the other hand, if we choose option two, then we have the same problem, except now it becomes unclear what role the higher-order state plays. It would then seem that higher-order states are generally not needed for conscious experience, which would also be disastrous for any HO theorist. Thus, Levine says later on: "When the higher-order state *misrepresents* the lower-order state, which content – higher-order or lower-order – determines the actual quality of experience? What this seems to show is that one can't divorce the quality from the awareness of the quality." (Levine 2001: 168, italics added)

However, both Levine (2001: 108–109) and Neander (1998: 429–430) recognize that other options are open to the HO theorist, but then dismiss

them as well. I will focus on Levine's treatment of these alternatives, and argue that these options are more viable than he thinks and that they are also closely related.

> **Option three**: "...when this sort of case occurs, there is no consciousness at all." (Levine 2001: 108)
> **Option four**: "A better option is to ensure correct representation by pinning the content of the higher-order state directly to the first-order state." (Levine 2001: 108)

First of all, it is a bit unclear what Levine means, in option three, by "no consciousness at all." Presumably, he does not mean that the hypothetical *person* in question would be completely unconscious. This would be a very unnecessary and puzzling consequence of any HO theory. It would also be to confuse "creature" consciousness with "state" consciousness, to use Rosenthal's terms. So it would seem that Levine's option three is really saying that, in such cases of misrepresentation, the person has *neither* the greenish *nor* the reddish conscious experience. But then it becomes unclear why Levine rejects option three as *ad hoc* (2001: 108). What exactly is so *ad hoc* about that reply? The HOT theory says that when one has a conscious mental state M, it is accompanied by a HOT that "I am in M." If there isn't a "match" between the contents of the lower-order and higher-order states, then it seems perfectly appropriate for the HOT theorist to hold that something like option three is a legitimate possibility. After all, this is an abnormal case where applying the HOT theory could not be expected to result in a normal conscious state. We are not told just how unlikely or abnormal such a scenario would be. There is an important lack of detail in Levine's thought experiment; recall that we are simply told to "suppose because of some neural misfiring (or whatever)...". Perhaps there would be no resulting conscious experience of the diskette case at all. Alternatively, if there are certain brain lesions involved, perhaps there would be some kind of loss of color vision (achromatopsia) with respect to the diskette case. The diskette case might then be experienced as neither green nor red.

This brings us to the "better option" in option four. In a sense, then, a defense of option three leads us naturally into option four. Indeed, they seem to be two sides of the same coin because defending option three is, in essence, arguing that a match between the higher-order and lower-order states must be "ensured" in order to result in a conscious experience with respect to the relevant concepts involved, e.g. "by endowing [the content of the higher-order state] with demonstrative content." (Levine 2001: 108) Levine does mention two problems with this fourth approach, but I am very puzzled by his re-

marks. He first asks "what if the higher-order state is triggered randomly, so that there's no first-order sensory state it's pointing at? Would that entail a sort of free-floating conscious state without a determinate character?" (Levine 2001: 109) The answer to the second question is clearly no because, in that case, you would have merely an unconscious HOT without a target state. An unconscious HOT, by itself, cannot result in a conscious state of any kind. Second, Levine simply expresses puzzlement about just how option four "overcomes the basic problem. . .[which] is that it just doesn't work to divide phenomenal labor." (Levine 2001: 109) This is not really *another* objection to option four or to the HOT theory; it merely repeats Levine's conclusion.

In addition, to revisit the initial thought experiment, when Levine says that my "visual system is in state R. . .but my higher-order state is HG," this is highly misleading and perhaps even begs the question against the HOT theory. What encompasses the "visual system"? Levine assumes that it is only the lower-order state R. However, if the HOT theory is true, it seems much more plausible to treat the *entire* system (including both the lower-order and higher-order state) as parts of the "visual system" in this case. Thus, the visual system (or at least the *conscious* visual system) would have to contain R *and HR*, so that there *would be* a conscious reddish experience (even if an idle HG state *also* exists). Perhaps option two is thus not so arbitrary after all. If so, then the hypothetical scenario would seem to be misdescribed or just assumes the falsity of the HOT theory. HOTs should be understood as part of the "visual system" when one is having a conscious perception of any kind. We should also say, then, in this case, that R and HR are each *necessary* for having a reddish experience, but neither one is *sufficient* by itself. R and HR are jointly sufficient. As I will urge later, empirical evidence also exists for such a characterization of the HOT theory.

It might be useful here to contrast Levine's example with two other abnormal cases. (1) In cases of *visual agnosia* subjects suffer from the inability to recognize perceived objects, but the disorder is neither due to any intellectual deficiency nor to any basic sensory dysfunction. The visual agnosic consciously perceives things, but cannot recognize what they are. They often mistake one object for another; for example, even mistaking one's wife for a hat (Sacks 1987). This differs from Levine's case in that the mere perception of objects remains intact, including the color perception. Visual agnosics are not blind and do not have damage to area 17 of the visual cortex. However, a HOT theorist might view this case as one where a HOT does not "match up" with the first-order visual input. Thus, it seems reasonable to view this as a case where the "normal" concept in the HOT does not accompany the input received through the visual modality (Gennaro 1996: 136–138). (2) On the other hand, there

is also the well-known phenomenon of *blindsight* (Weiskrantz 1986, 2000) whereby patients can sometimes accurately answer questions about objects without having any conscious perceptions of those objects, such as detecting the movements and shapes of objects in the blind visual field. Unlike Levine's case and the visual agnosic, the blindsight patient does not have a conscious perception *at all* of objects in her visual field (due to lesions of the visual cortex), though the common wisdom is that some visual information is processed in other "secondary" parts of the brain in ways that explain the behavioral data. In this case, there would be no HOT at all directed at the lower-order visual input because there is no conscious perception at all. At best, we have a case of a nonconscious mental state or an informational state influencing behavior (Gennaro 1996: 129–134).

Nonetheless, I do think that Levine and Neander have, in a somewhat indirect way, hit upon one very important and potentially troubling issue regarding the nature of HOT theory. There may indeed be an element of truth in Levine's argument; namely, that it is difficult to make sense of entirely splitting off the lower-order state from the HOT. Thus, I do believe that he is grappling with a deeper issue that must be addressed by any HOT theorist. It is perhaps best expressed by Levine when he says that the HOT theory has a difficulty with

> the paradoxical duality of qualitative experiences: there is an awareness relation, which ought to entail that there are two states serving as the relevant relata, yet experience doesn't seem to admit of this sort of bifurcation. Let's call this the problem of "duality." (Levine 2001: 168)

The problem of duality is an important problem, but I think it can be handled best by adopting a variation of Rosenthal's HOT theory. According to the variation that I have defended elsewhere (Gennaro 1996), first-order conscious mental states are complex (or global) states comprised of both the lower and higher-order states. Consequently, I think that consciousness is an intrinsic and essential feature of conscious mental states, unlike Rosenthal who holds that the HOT is an inessential and extrinsic property of conscious states. I have called this position the "wide intrinsicality view" (WIV) and have argued for its superiority over Rosenthal's model (Gennaro 1996: 24–30). In this context, though, I believe that the WIV can help to alleviate some of the puzzlement expressed by both Neander and Levine. For example, in the quote above, a proponent of the WIV can respond that conscious experience (from the first-person point of view) does not seem to allow for a split ("bifurcation") between the lower-order and higher-order states. However, the "awareness relation" still does not entail the existence of two entirely separate states. Instead, on the

WIV, we have two parts of a single conscious state with one part directed at ("aware of") the other. In short, there is a complex conscious mental state with an inner intrinsic relation between parts. There is, therefore, a kind of "self-referential" element in consciousness. This general idea is most closely linked to Brentano (1874/1973), but I have also argued at length that it is the best way to understand Sartre's theory of consciousness (Gennaro 2002).

Moreover, this variation allows us to avoid one controversial aspect of Rosenthal's theory which is also a target of Levine's critique: Rosenthal's theory "splits off subjectivity from qualitative character, and . . . it is precisely this feature that seems so implausible." (Levine 2001: 105). Unlike Rosenthal (1991), I do not hold that there can, for example, be unconscious *sensory* or *qualitative* states. Some of the disagreement here is purely terminological, e.g. how to use the terms 'sensory,' 'experience,' and the like. However, there is also substantial disagreement. Since Rosenthal believes that HOTs are extrinsic and inessential to their target states, he (unlike me) holds that the lower-order states can exist without HOTs *and continue to have their qualitative properties*. On my view, however, it is the HOTs which bring the intrinsic qualitative properties into the conscious state. So, for example, on my view there can be unconscious pains and perceptions, but they are not 'sensory' or 'qualitative' states while unconscious. When a pain or perception becomes conscious by virtue of becoming the target of an appropriate HOT, it then becomes a qualitative state.

In a striking passage, Neander credits Barry Loewer for an "ingenious suggestion that might be worth pursuing" (1998: 430):

> the suggestion is that the two levels of representation might be collapsed into one level that is *self-referential*. . ..This suggestions also rids us of the division of phenomenal labor, while still allowing us to maintain that the difference between conscious and unconscious sensory representations is that some of them are meta-represented and some are not. Since the first and second-order representings no longer involve two separate representations. . .the two cannot come apart, so mis-(meta-)representation is in principle impossible.
>
> (Neander 1998: 429–430)

This sounds very familiar, but Neander unfortunately also dismisses this option too quickly (1998: 430). I hope I have shown that this is a truly viable option which can help to counter what Levine calls "the problem of duality" along the lines of his option four.

Furthermore, it is also now possible to address some of Levine's other concerns about whether or not qualia can be explained as *either* intrinsic *or* relational features of conscious states (2001: 93–107; cf. Levine 1995). With the

WIV alternative, we can now see that this is a false dichotomy. Qualitative states can be complex states with both intrinsic and (inner) relational features. This solution is perhaps similar to what Levine calls "the complexity gambit," (2001: 95) but I have already argued that such a move can address his challenge about the nature of qualitative states while, at the same time, not admitting "that no progress can be made [in explaining consciousness] if we consider qualitative character to be an intrinsic property of experience. . ." (Levine 2001: 94) This is yet another advantage of the WIV over Rosenthal's HOT theory: Consciousness can be both an intrinsic and relational property of experience without giving up on any further explanation of its nature. A case could be made that Rosenthal mistakenly infers that treating consciousness as an intrinsic quality of experience forces one into an allegiance with the unsatisfying Cartesian position whereby consciousness is an unanalyzable property of conscious states (see Gennaro 1996: 21–24).[6]

Finally, I wish briefly to mention some very suggestive and relevant empirical evidence which I think supports the HOT theory in general and the WIV in particular. Gerald Edelman and others have argued that loops (or *re-entrant pathways* or *back projections*) in the neural circuitry of the brain are essential for conscious awareness (e.g. Edelman & Tononi 2000). As Churchland puts it, "the idea is that some neurons carry signals from more peripheral to more central regions, such as from V1 to V2, while others convey more highly processed signals in the reverse direction. . .it is a general rule of cortical organization that forward-projecting neurons are matched by an equal or greater number of back-projecting neurons." (Churchland 2002: 148–149) I cannot go into further neurological detail here, but it is worth making the following observations in this context: (1) The brain structures involved in loops seem to resemble the structure of at least some form of HOT theory; namely, that lower-order and higher-order states are combining to produce conscious states. (2) The importance of higher-order *concepts* in conscious experiences is also readily apparent. Part of the reason why Edelman and others believe that back projections play a prominent role in consciousness is that "perception *always* involves classification; conscious seeing is *seeing as*." (Churchland 2002: 149) This is a key aspect of any HOT theory. (3) More specifically, such evidence seems to support the WIV version of the HOT theory because of the intimate and essential relationship between the "higher" and "lower" areas of the brain involved. There is essential and mutual interaction between the relevant neuronal levels. Edelman and Tononi, for example, emphasize the global nature of conscious states and it is reasonable to interpret this as the view that conscious states are composed of both the higher and lower order states. They refer to what they call the "dy-

namic core"as generally "spatially distributed and thus cannot be localized to a single place in the brain." (Edelman & Tononi 2000: 146) It seems to me that their description of the neural correlates of consciousness fits more naturally with the WIV version of the HOT theory. (4) With respect to Levine's objection, then, it is at best misleading to treat the lower and higher level "parts" of a conscious state as potentially "bifurcated." While the standard AHOT theory *might* have a problem with respect to possible misrepresentation, it seems to me that the WIV is better able to handle this objection. This is because the "two" levels cannot really come apart in the way that Levine describes in his thought experiment. Levine's options three and four, therefore, seem open to the HOT theorist and particularly open to a defender of the WIV.

3. Conclusion

In closing, then, the AHOT theory of consciousness is consistent with animal consciousness. Those who think otherwise, such as Peter Carruthers, typically build too much into the HOTs which render mental states conscious. Moreover, the AHOT theory remains preferable to the DHOT alternative. Secondly, the charge that the HOT theory cannot handle cases of misrepresentation can be answered. Authors such as Levine and Neander do not properly explore the options and resources available to the HOT theorist, and they hastily conclude that the HOT theorist must choose between arbitrary or clearly untenable alternatives.

Notes

1. No doubt that part of the problem here is terminological, as is often the case in the literature on consciousness. One can, I suppose, *speak* of non-conscious 'sufferings,' 'feelings,' and 'desire frustrations,' and we are perhaps all entitled to use our own terminology to some extent. However, it seems to me that there is a point where using terms in this way becomes more of a provocative attempt to redefine them and can even simply add to the terminological confusion. It is most important, however, to keep our sights set on the substantial disagreement about whether or not animals have phenomenally conscious mental states.

2. For another more detailed criticism of Carruthers' recent moral argument (see McKinnon 2002).

3. Another important feature of Carruthers' account is that some perceptual contents acquire a dual analog content, one purely first-order (e.g. 'green') and the other higher-order (e.g. 'experience of green'). For Carruthers, it is the higher-order contents which confer sub-

jectivity on the perceptual state; conscious experiences (percepts) are made conscious by virtue of their being available to the subject's HOT forming module. See Carruthers (this volume) for more details.

4. Peter Carruthers has informed me (via email correspondence) that he is an "actualist" about introspection. Perhaps this is his best option for various reasons, though it may sound a bit surprising at first. However, I now wonder (a) why so little was said in his book about the structure of introspection, and, more importantly, (b) whether or not this detracts from his initial motivation for the dispositional account of first-order consciousness, especially when combined with the questions raised earlier about his "cognitive overload" argument against AHOT theory.

5. It actually seems to me that such tests for 'other-attributing' thoughts in the cognitive ethology and theory of mind literature are really aimed at determining whether or not animals or infants can have *conscious* HOTs directed at another's mental state. Children are often even asked to verbalize their attitudes toward another's beliefs or perceptions, and animals seem to be tested for behavioral signs that they are consciously thinking about another's beliefs, e.g. in cases of deception. In my view, even if the evidence suggests that these subjects fail such tests, it causes no problem for the HOT theory since the theory certainly allows for the presence of conscious states in the absence of (either self-attributing or other-attributing) conscious HOTs. This is also one very brief line of reply to Seager (this volume).

6. It is worth noting that Rosenthal defends Levine's "option 1" in several places. See, for example, Levine (2001:190, Note 24) for one place where Rosenthal says that if the HOT occurs without the target state the resulting conscious state might just be subjectively indistinguishable from one in which both occur. Once again, I find this option implausible partly because, as Levine says, "...doesn't this give the game away?...then conscious experience is not in the end a matter of a relation between two (non-conscious) states." (Levine 2001:190). Second, it seems to me that since the HOT is nonconscious, there would not be a conscious state anyway unless there is also the accompanying lower-order state. Thanks to Alex Byrne for calling my attention to this point.

References

Baron-Cohen, S. (1995). *Mindblindness: an essay on autism and theory of mind*. Cambridge, MA: MIT Press.

Brentano, F. (1874/1973). *Psychology from an empirical standpoint*. New York: Humanities

Carruthers, P. (1989). Brute experience. *Journal of Philosophy*, *86*, 258–269.

Carruthers, P. (1992). *The animals issue*. Cambridge: Cambridge University Press.

Carruthers, P. (1999). Sympathy and subjectivity. *Australasian Journal of Philosophy*, *77*, 465–482.

Carruthers, P. (2000). *Phenomenal consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. (this volume). HOP over FOR, HOT theory.

Carruthers, P. & P. Smith (Eds.). (1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Cavalieri, P. & H. Miller (1999). Automata, receptacles, and selves. *PSYCHE*, 5. <http://psyche.cs.monash.edu.au/v5>

Churchland, P. S. (2002). *Brain-wise: studies in neurophilosophy*. Cambridge, MA: MIT Press.

Dennett, D. (1991). *Consciousness explained*. Boston: Little Brown.

Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Edelman, G. & G. Tononi (2000). Reentry and the dynamic core: neural correlates of conscious experience. In T. Metzinger (Ed.), *Neural correlates of consciousness* (pp. 139–151). Cambridge: MIT Press.

Gennaro, R. (1993). Brute experience and the higher-order thought theory of consciousness. *Philosophical Papers*, *22*, 51–69.

Gennaro, R. (1996). *Consciousness and self-consciousness*. Amsterdam: John Benjamins.

Gennaro, R. (2002). Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy*, *32*, 293–330.

Gunther, Y. (Ed.). (2003). *Essays on nonconceptual content*. Cambridge, MA: MIT Press.

Güzeldere, G. (1995). Is consciousness the perception of what passes in one's own mind? In T. Metzinger (Ed.), *Conscious Experience* (pp. 335–357). Schöningh: Imprint Academic.

Kim, J. (1996). *Philosophy of mind*. Boulder, CO: Westview Press.

Levine, J. (1995). Qualia: intrinsic, relational or what? In T. Metzinger (Ed.), *Conscious Experience* (pp. 277–292). Schöningh: Imprint Academic.

Levine, J. (2001). *Purple Haze*. Cambridge, MA: MIT Press.

Lurz, R. (2002). Reducing consciousness by making it HOT: A review of Peter Carruthers' *Phenomenal Consciousness*. *PSYCHE*, 8. <http://psyche.cs.monash.edu.au/v8>

McKinnon, C. (2002). Desire-frustration and moral sympathy. *Australasian Journal of Philosophy*, *80*, 401–417.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–450.

Neander, K. (1998). The division of phenomenal labor: a problem for representational theories of consciousness. In James Tomberlin (Ed.), *Language, mind, and ontology* (pp. 411–434). Oxford: Blackwell.

Ridge, M. (2001). Taking solipsism seriously: nonhuman animals and meta-cognitive theories of consciousness. *Philosophical Studies*, *103*, 315–340.

Rolls, E. (1999). *The brain and emotion*. Oxford: Oxford University Press.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329–359.

Rosenthal, D. (1990). A theory of consciousness, Report No. 40 on MIND and BRAIN, Perspectives in Theoretical Psychology and the Philosophy of Mind (ZiF), University of Bielefeld. A version of this paper is reprinted in N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness* (pp. 729–753). Cambridge, MA: MIT Press.

Rosenthal, D. (1991). The independence of consciousness and sensory quality. In E. Villanueva (Ed.), *Consciousness* (pp. 15–36). Atascadero, CA: Ridgeview.

Rosenthal, D. (1993). Thinking that one thinks. In M. Davies & G. Humphreys (Eds.), *Consciousness* (pp. 197–223). Oxford: Blackwell.

Rosenthal, D. (2004). *Consciousness and Mind*. New York: Oxford.

Rosenthal, D. (this volume). Varieties of higher-order theory.

Rowlands, M. (2001). Consciousness and higher-order thoughts. *Mind and Language*, *16*, 290–310.

Sacks, O. (1987). *The man who mistook his wife for a hat and other clinical tales*. New York: Harper and Row.
Seager, W. (this volume). A cold look at HOT theory.
Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: MIT Press.
Van Gulick, R. (2000). Inward and upward: reflection, introspection, and self-awareness. *Philosophical Topics*, *28*, 275–305.
Weisberg, J. (1999). Active, Thin and Hot! An actualist response to Carruthers' Dispositionalist HOT view. *PSYCHE*, 5. <http://psyche.cs.monash.edu.au/v5>
Weiskrantz, L. (1986). *Blindsight*. Oxford: Clarendon.
Weiskrantz, L. (2000). *Consciousness lost and found*. Oxford: Oxford University Press.

Chapter 4

Higher-order global states (HOGS)

An alternative higher-order model of consciousness

Robert Van Gulick

1. Historical and contemporary background

Whatever may have been the case before the 17th century, in modern thought about the mind, consciousness and self-awareness have been typically aligned. Indeed by 1644 Descartes had defined thought itself in terms of self-awareness, as in *The Principles of Philosophy* IX where he wrote, "By the word 'thought' ('pense,') I understand all that of which we are conscious as operating in us." Half a century later John Locke took a more cautious line about equating the mind *per se* with conscious, but he still linked self-awareness with thought, which he clearly conceived of as conscious or experiential thought. In *An Essay Concerning Human Understanding* Locke put it thus,

> I do not say there is no soul in a man, because he is not sensible of it in his sleep, but I do say he cannot *think* at any time waking or sleeping, without being sensible of it. Our being sensible of it is not necessary to anything but our thoughts; and to them it is, and to them it always will be necessary.
>
> (Book II, 1:10)

The link between the two runs through much of modern thought down to the present. Few today would limit mind, or even thought, to that which is conscious; the existence of meaningful mental processes that occur beneath our awareness is taken as a given. But the notion of a conscious mental state remains tightly bound to that of self-awareness, a link that finds current expression in the so called "higher-order theory" of consciousness. Higher-order theories come in many versions, but all agree in treating the distinction between conscious and nonconscious mental states as a relational difference in-

volving the presence or absence of a further meta-intentional state. "Higher-order" here means meta-intentional, i.e. involving states that are about other states, such as thoughts about one's thoughts or about one's desires. A conscious mental state M is defined as one that is accompanied by a simultaneous meta-intentional state whose content is that one is in M. What makes a desire into a conscious desire is our awareness of it. Having a conscious desire for a cup of coffee requires being simultaneously in two mental states: a first-order desire for a cup of coffee, plus a second order thought or perception of oneself as having that desire. Thus on the higher-order view, the difference between conscious and nonconscious mental states turns not on any contrast in their intrinsic properties, but rather on the extrinsic relational fact of whether or not the relevant state is (or is not) accompanied by a simultaneous meta-state directed at it.

Two main variants of the higher-order view appear in the recent literature. Both agree that what makes a conscious state conscious is the simultaneous occurrence of meta-intentional state directed at it, but they disagree about the psychological modality of the relevant meta-states. Some take them to be thought-like, and others regard them as more perception-like. Thus the proponents of the two versions are said to advocate either a higher-order thought (HOT) model of consciousness or a higher-order perception (HOP) model. David Rosenthal (1986, 1992) has been the most prominent champion of the first, and William Lycan (1987, 1996) and David Armstrong (1980) have been major supporters of the latter. My current aim is not to assess the comparative merits of the two alternatives. I have done that at some length elsewhere (2001) and reached a mixed conclusion: the HOT and HOP views each have their relative strengths and difficulties, and neither is the clear overall winner (though see Lycan, this volume, for a comparative evaluation that not surprisingly favors the HOP approach.) My focus here is on problems or weaknesses that are shared by both versions, and how dealing with those difficulties may lead us to an alternative model that appeals to higher-order global states (HOGS).

I will begin by reviewing some of the most important reasons in favor of the higher-order theory in its standard HOT or HOP form, and then turn to several major objections that have been raised against it by its critics. In response to those problems, I will consider how we might weaken or discard some shared aspects of the standard model, which will lead us to the alternative HOGS model which reformulates the higher-order view in a way that appears better able to answer its critics.

2. Standard HOT and HOP models

The higher-order theory's has several basic strengths, including the following three.

1. It accords well with our common use of the adjective "conscious" as applied to mental states. Both in everyday use and in the generalized neo-Freudianism that pervades contemporary culture, the divide between conscious mental states and unconscious ones is a matter of epistemic access. Our unconscious wants, desires and memories are hidden from us; they lie "buried" beneath our awareness, and bringing them to consciousness is a matter of becoming aware of them, whether in a thought-like or perception-like way. In so far it is this distinction that it aims to capture, the higher-order view seems pretty much on target.
2. The higher-order theory also accords well with the empirical methods that are used to determine when subjects are having conscious mental states and when not. Researchers typically rely upon a first person report criterion. If the subject can report that she saw or perceived a stimulus or had a specific thought, then she is counted as having had a conscious mental state of the relevant sort. As proponents of the higher-order theory have noted (especially Rosenthal) such reports are the linguistic expressions of the sort of higher-order thought or beliefs that the theory posits. We generally assume that if a subject has the linguistic capacity to report on her mental states at all, then she can do so whenever she is in a conscious mental state. This is neatly explained as an immediate consequence of the higher-order view. For example, if I have a conscious desire then I must have a higher-order meta-state whose content is that I have that desire. My ability to make the relevant report is a simple matter of my being able to give linguistic expression to the higher-order state that constitutes my relevant self-knowledge.
3. The higher-order theory offers a potentially demystifying account of consciousness. If conscious states differ only in being accompanied by simultaneous meta-intentional states, then the problem of consciousness may reduce to the seemingly more tractable problem of intentionality. Many theorists of mind believe it possible to give a naturalistic account of intentionality and the aboutness relation, i.e. to explain the natural basis for a state's having content or being about a given state of affairs. If so, then it should be possible to apply that general account to the specific case of meta-states whose contents are about other mental states. In so far as we can expli-

cate that relation in naturalistic or physicalistically acceptable terms, then according to the higher-order theory we would also have succeeded in giving a naturalistic account of conscious mental states, a goal which many have argued is far more difficult if not impossible. By thus reducing the problem of conscious mentality to that of explicating intentional relations among nonconscious mental states, the higher-order theory promises to offer a solution to what many regard as the greatest obstacle to the project of naturalizing the mind.

Despite these substantial virtues the higher-order theory has met with many objections from its critics, of which four interrelated concerns are worth noting here.

1. Although the higher-order theory may do a good job of capturing one common use of the adjective "conscious", there seems to be another sense in which we speak of conscious states for which it seems less adequate. Moreover, according to some critics it is this other notion of a "conscious" mental state which is problematic for the naturalist's program. Being a conscious state in this latter phenomenal sense involves the intuitive idea that there is "something that it's like to be in that state" (Nagel 1974), some subjective experiential aspect present from the first person point of view. According to this line of objection, the crucial version of the conscious/unconscious distinction is not that between states we know we are in versus those we do not, but rather the distinction between those states which there is something that it's like to be in vs those for which there is not. Of course, higher-order proponents may try to show that the two distinctions collapse into one – indeed some have tried to do just that (Rosenthal 1991, 1992) – but the viability of the higher-order view as a general account of conscious states would then depend upon the plausibility of that alleged reduction.
2. Higher-order theories need to include certain further conditions to rule out obvious counter examples to the sufficiency of the higher-order analysis, but those conditions call into question the theory's basic idea of explaining consciousness in terms of meta-intentional content. For example, all the standard higher-order models require that the meta-state be roughly simultaneous with the state that it makes conscious and that the meta-state be arrived at noninferentially. The latter conditions is needed to rule out cases in which I come to know that I am in a given mental state by inferring it from observations of my behavior or other forms of third person evidence. If Tom infers that he wants to ingratiate himself with his supervisor

by observing the silly compliments that he pays her, that would not automatically make his desire into a conscious mental state in the sense that the higher-order theory aims to capture. The conscious-making higher-order states need to arise through some internal noninferential channel, one through which we just know directly what we want, remember or think. Including a noninferential condition in the analysis may help to exclude counter examples and perhaps preserve its extensional adequacy, but it does so at a significant price. The guiding idea of the higher-order theory is that what is crucial to consciousness is meta-intentional content; conscious states are states *about* which we have thoughts or awareness. But including the non-inferential condition calls that into doubt. States arrived at inferentially need be no different in content from those arrived at directly, yet according to the higher-order theory their path of origin deprives them of the power to make their lower-order mental objects conscious. Were it the meta-intentional content that really matters in making an unconscious state into a conscious one, then it would seem that the route by which the meta-state was produced would not make any difference. The fact that it does matter seems to imply that something other than the meta-intentional relation plays a major role in transforming nonconscious states into conscious ones. The real work may well be done not by the noninferential condition *per se* but by some other condition that correlates or covaries with it. Perhaps meta-states produced by the noninferential channel involve different vehicles of content; even if they can share their contents with inferentially produced meta-states they may use a different medium or system of representation, one that is crucial for making a state conscious.

To a lesser degree, similar problems affect the simultaneity condition. Both HOT and HOP models require that the meta-state be simultaneous with its lower order object in order to make it conscious. As with the noninferential requirement, the simultaneity condition may be needed for extensional adequacy, but it raises similar doubts about the degree to which the meta-intentional relation is doing the real work in the explaining the conscious/unconscious distinction. If what matters for making a state conscious is having a higher level meta-representation of it, then it seems that such states should work as well for nonsimultaneous mental objects. Since they clearly do not, it seems that something other than meta-intentionality also plays a major role.

3. A related objection which we may label the "Generality Problem," has been raised by Fred Dretske (1993) among others (Byrne 1997). In general, having a thought or perception of some object of type F, does not make it into a

conscious F. So why should having such a thought or perception of a mental object transform it into a conscious one. If my perceiving or thinking of a stone, a pencil or my nose does not turn any of them into conscious objects, then why should my thinking of one of my own mental states suffice to make it conscious? And even if it suffices as a matter of extensional adequacy, what is it about the relevant intentional relation or its correlates that explains why it does so in the meta-mental case even though perceiving or thinking of some object x does not in general make x conscious. It will not do to simply appeal to usage and say, "We apply 'conscious' to mental states we know of but not to others things we know." What is required is some explanation of why we do so, some account of the intuitive difference we feel between the two cases which grounds the difference in usage.

Another way to frame the worry is in terms of intrinsic vs merely relational properties and differences. According to the higher-order theory, the conscious/ unconscious distinction turns on a purely extrinsic relational difference: their being or not being a further state that is about the occurrence of the lower-order state. Thus the transformation from unconscious to conscious state involves no change in the state itself but only the addition of a purely external and independent state which has it as its intentional object. Yet in so far as one interprets the conscious/unconscious distinction as dividing states in terms of whether they are conscious in the sense of there being "something that it is like to be in them" it seems difficult to accept the idea that the division involves no differences in the intrinsic properties of the states themselves but only differences in purely relational facts about which other states if any are intentionally directed at them. Higher-order theorists have of course attempted to reply to this objection (Rosenthal 1991; Lycan 1996) but it not clear that their efforts have succeeded in solving the problem or dispelling the worry.

4. The question of how qualia fit into the higher-order theory presents another closely related way to raise the concerns voiced in the earlier objections. Those critics who claim that the higher-order theory fails to capture the distinctive first person nature of conscious states may also complain that it fails to account for the difference between those mental states that have phenomenal qualia and those that lack them. Qualia are often invoked as a way of unpacking the "something that it's like to be" notion of a conscious state. Qualia or phenomenal feels are taken to be features or aspects of the "what it's like"-ness of a conscious state, modes of its experiential character. My experience's being of phenomenal red is part of what it's like for me to be visually conscious of a ripe tomato. Thus the transition from

> states that are not conscious in the "what it's like" sense to states that are may seem to require the addition of qualia. But that in turn may seem to conflict with the basic higher-order claim that a state's becoming conscious does not involve changes in its intrinsic properties but only the addition of an extrinsic relational factor, namely a meta-state that is directed at it. How can the addition of such a merely relational element make a state into one such that there is something that it's like to be in it?

Some higher-order theorists have a surprising answer. They distinguish having qualia from having "what it's likeness". They restrict the latter to states that are conscious in the relevant higher-order sense, but they allow qualia to occur as properties of unconscious states that there is nothing that it's like to be in. On this view, one can have qualitative but unconscious mental states such as unconscious color perceptions or pains, but there is nothing that it is like to be in such a state. (Rosenthal 1991; Lycan 1996) Since one is not aware of being in a state with qualitative properties there is nothing that it is like be in it; it's not like anything to have an unconscious pain. Only when one becomes aware of the state by having the requisite HOT or HOP does one become aware of its qualitative properties. Thus according to the higher-order theorist the transition from states that are unconscious in the "what it's like" sense to those that are conscious, does not involve adding qualia to the state but of becoming aware via a meta-state of the qualia that the lower-order state already had.

The proposed solution is ingenious and has a certain appeal, but it also has its problems. It requires us to accept the idea of unconscious qualia, which many find incoherent or contradictory. Moreover, it seems to strand the notion of phenomenal or experiential feel in a "no man's land" unable to quite take hold at either the lower or the meta level. According to the proposal, qualia can be fully present in unconscious states which there is nothing that it is like to be in. Whatever one might say to explain the notion of having qualia in that sense, they seem unable to account for the central "what it's likeness" of experience, since the proposal acknowledges that there is nothing that's like to be in such a state. Conversely and almost paradoxically the meta states whose addition supposedly accounts for first person experiential feel are themselves devoid of any qualia. Thus the proposal requires a commitment both to qualia of a sort that do not by themselves produce any "what it's likeness" and to states that produce what it's likeness but themselves lack any qualia. It is in that sense that the qualia seem stranded.

The higher-order theorist denies that qualia are properties of the higher-order state, but accepting his claim that they are properties of the lower-order

state requires us to regard qualia as properties whose presence is compatible with the total absence of experiential feel. Thus qualia, commonly thought of as modes of experience, seem present at neither level in the proposed scheme. That in itself does not prove the proposal is mistaken, but it does show that its acceptance would require a major revision in our normal ways of thinking about qualia and experiential feel. Some may be willing to pay the price given the other benefits the higher-order theory promises, but it would be nice to find some way to modify the theory to avoid having to do so.

3. The HOGS model: Higher-order global states

The standard versions of both the HOT and HOP views agree in treating the conscious-making meta-states as distinct and separate from their lower-order objects. This is perhaps more explicit on the HOP model since its higher-order perceptions are realized by distinct token representations that occur only within one or another intra-mental monitoring system that is distinct from the states and processes that it monitors. However, the HOT model, at least in its typical form, is also implicitly committed to the HOTs being separate and distinct from their lower-order objects, if for no other reason than the difference in their intentional content. However, though distinctness is standardly assumed by higher-order theories it is not strictly entailed by it, and at least one higher-order theorist (and editor of this volume) has denied it (Gennaro 1996).

Rejecting or weakening the distinctness assumption creates the possibility for a different sort of the higher-order theory, which I have elsewhere (2001) called the HOGS model as an acronym for Higher-Order Global States. The first part of this section, which closely tracks that earlier presentation, explains the main features of the HOGS model as well as three sources of inspiration on which it draws (with apologies to those who already know that part of the story.) The latter part of this section then explains in greater detail how those three strands are woven into the HOGS model. The following two sections specifically address the respect in which the HOGS model implicitly incorporates higher-order reflexive mentality as a basic feature of the structure of phenomenal experience rather than locating it in a distinct explicit meta-state.

The basic idea of the HOGS model is that lower–order object states become conscious by being incorporated as components into the higher-order global states (HOGS) that are the neural and functional substrates of conscious self-awareness. The transformation from unconscious to conscious state is not a matter of merely directing a separate and distinct meta-state onto the lower-

order state but of "recruiting" it into the globally integrated state that is the momentary realization of the agent's shifting transient conscious awareness.

The model draws on several sources for its inspiration: Dan Dennett's notion of consciousness as cerebral celebrity and of the self as virtual focus of one's ongoing inner serial narrative, Chris Hill's view of introspection or conscious attention as a matter of volume control, and the empirical evidence for taking globally integrated neural states as the substrates of conscious experience. Let me say something brief about each.

1. Consciousness as cerebral celebrity. According to Dennett's multiple drafts theory (1991), the distinction between conscious and nonconscious mental states is blurry, admits of degrees and turns on two principal dimensions. The first concerns the degree to which a mental state (or content fixation) influences the subsequent development of the system's states and its outputs. This is what is meant by "cerebral celebrity"; to put it crudely, the more effect a given content fixation has on what other content fixations occur, the more "famous" it is. Conscious states take a more powerful and broader range of content-relative effects throughout the agent's mind; a conscious perception (thought or desire) and its content will be accessible to other processing areas, more able to affect other states (thoughts, desires, memories) and have more impact on those states driving the system's output, especially on the system's reports about its state of mind since conscious state are normally ones that we can report ourselves as being in. All these aspects of influence admit of degree, and in general the greater the impact of any given state the greater its level of cerebral celebrity. Thus in so far as being a conscious state is a matter of such "intra-mental fame", whether or not a state is conscious need not have a strict yes or no answer. The other dimension of consciousness on the multiple drafts model is the degree to which a given content gets integrated into what Dennett describes as the ongoing serial narrative the system constructs from the "stream of consciousness". This is not a separate meta-narrative that is produced independently or over and above the system's lower-order content fixations. Rather it is an assemblage of activated lower-order contentful states that cohere together in such a way that they form a more or less integrated set from the perspective of a unified self. Dennett denies that there is any separate self that constructs or views the sequence; rather it is the other way round. It is the coherent serial narrative that is fundamental and the self is merely a virtual entity that exists as the perspectival point which is implicit in the narrative and from which the narrative hangs together as unified.

Dennett's multiple drafts theory is thus a higher-order theory of a sort, though it differs greatly from more mainstream HOP and HOT models. A state with a high degree of cerebral celebrity will typically be one that the agent can report being in, and such a report would express the relevant higher-order thought. Indeed Dennett, like Rosenthal, relies heavily on a tight link between a state's being reportable and its being conscious. The second aspect of his theory also has a decidedly higher-order slant, since incorporation into the serial narrative carries with it the status of being represented as a state in the stream of the (virtual) self, which at least implicitly involves higher-order representation.

2. Introspection as volume control and activation. Chris Hill (1991) has faulted the "inner eye" model of introspection as overly passive. He has argued that introspection is active in the sense that it often alters its lower-order mental object. In a case of paradigmatic external perception, as when I see the lamp on my desk, my awareness of the object does not change it. The lamp is unaffected by being seen. However, inner awareness does often seem to alter its objects. When I turn my inner attention to the lingering taste of the olive that I ate a few minutes ago or to the ache of a running pull in my left Achilles tendon, the act of directing my awareness upon them can change many of their features. The sensation often gains in intensity and vividness; various sensory properties may become more specific, shift from one specific character to another or even emerge where no detailed character was previously present.

   The shift is importantly different from what happens in external perception. There too a redirection of attention typically leads to changes, but it is usually only the perceptual state that changes not its object. When I visually scrutinize my desk lamp, I become aware of many details that were previously unnoticed but the properties of the lamp itself remain unchanged. Admittedly in some external cases, the act of observation does change its object. That is apparently so at the quantum mechanical level, and obviously so in many social situations; indeed designing non-obtrusive measures is a perennial problem in the social sciences. However, in the interpersonal case, it is not the act of observation *per se* that produces the change but rather the subject's awareness at some of level of being observed that does so. By contrast, in the intra-mental case the mere act of observation does seem to alter its introspective objects.

   Hill thus contrasts the "inner eye" model of introspection with alternatives that he refers to as "volume control" and "activation" to emphasize the respects in which the intensity, character, or even the existence of a sensation

(or other lower-order state) can be affected by the occurrence of a higher-order awareness directed at it. He seems to regard this as a problem for the perceptual view of introspection and thus for the HOP model of consciousness. Lycan (1996), however, denies any such negative consequence follows for the HOP view. He accepts the active nature of inner awareness and the many ways in which it may alter its lower-order object, but denies that the HOP view is committed to a passive model of inner perception as the "inner eye" analogy might suggest. Thus Lycan accepts the data Hill presents but claims they are fully consistent with the HOP theory. For present purposes, we need not settle that latter dispute over consistency; it is the active nature of introspection that matters, and about that they agree.

3. Globally distributed neural correlate of consciousness. Current scientific evidence on the neural correlates of consciousness indicates that there is no special local brain area(s) that is (or are) the unique or special basis of conscious experience. Rather any given conscious state appears to be realized by a globally distributed pattern involving many different cortical and subcortical regions that are simultaneously active and bound together in some way, perhaps by regular oscillations that entrain neural firing patterns in disparate areas of the brain. An important consequence of this result is that the very same regions that are involved in the processing and realization of nonconscious mental states are also among the correlates or realization bases of conscious mental states. For example the areas of visual or auditory cortex that are active when one nonconsciously perceives a stimulus are also components of one's conscious perception of such a stimulus. The difference between the neural correlates of the conscious and nonconscious states is not that the information gets passed on and re-registered or re-presented elsewhere but rather that those same areas get integrated into a larger unified pattern of global brain activity in the conscious case.

Let us consider then how these three strands can be woven into the HOGS model. Its basic idea is that transforming a nonconscious state into a conscious one is a process of recruiting it into a globally integrated complex whose organization and intentional content embodies a heightened degree of reflexive self-awareness. The meta-intentional content is carried not by a distinct and separate vehicle but rather by a complex global state that includes the object state as a component.

Some of the connections to our three themes are obvious. The neurological evidence for global substrates coincides nicely with the model's view of consciousness as recruitment into a suitably integrated and broadly coherent

state. The evidence reveals no privileged location for conscious experience in the brain; it seems instead to be realized by interaction effects between the same cortical regions that are used in processing and encoding nonconscious information and content. The part of visual cortex that is active when I nonconsciously register a red apple to the left is also active when I consciously perceive it, and not merely as the causal precursor of the conscious experience but as an ongoing part of the experience's realization or substrate. This is just what one would expect on the HOGS model.

The implications for cerebral celebrity are also immediate. The recruited state is dynamically linked in content sensitive ways with a much larger field of active mental processes. Its sphere of influence is correspondingly increased. The dynamics of "fame" within the brain are much like those in the social worlds of politics and entertainment. The more active connections you have and the stronger the links, the more influence you have on the evolving state of the system. The connections are often also mutually reinforcing. The activity of individual states within the global ensemble tend to reinforce each other in a mode of resonant amplification. The activity of the recruited state typically increases in intensity. For example fMRI studies using a binocular rivalry paradigm show spikes in neural activation correlated with attentional shifts. The method is to present different images to the subject's two eyes respectively (e.g. a face to the right and a place or scene to the left). Alternating task demands prompt attentional shifts from one image to the other. Both images continue to be processed to a significant degree throughout since the stimuli remain invariant, but the attentional shifts alter the subject's experience and correlate with patterns of increased activity in the neural substrates for the consciously attended perception. Thus the global integration that produces the shift from nonconscious to conscious perception also tends to amplify the activity in the recruited state, further enhancing its capacity to influence other active regions. Recruited states are thus both better connected and more potent. As in the social world, becoming part of the "in crowd" enhances one's fame and influence.

Such amplification also brings the volume control strand into the HOGS model. As that metaphor reminds us, a shift in conscious attention typically alters its intra-mental object through amplification, heightened resolution and a diversity of possible changes in content. When I focus on my sensation or my desire, I likely affect it in some way so that its content, while related to that which it had before the attentional shift, is not the same in all respects as what it was. The ache in my tendon may become more intense, more locally placed and more articulated in its painfully stretched "soreness". Thus when the HOGS

model talks of "the object state" being preserved or recruited into the global higher-order state, it would be more accurate to talk of some "near successor of the original object state" being preserved as a component of the HOG. The integration proposed by the HOGS model is a dynamic reciprocal process in which the various components of a global state both amplify and modify each other in content sensitive ways. Thus as the idea of volume control suggests, recruitment will often to some degree alter the object state's specific first order content, as happens in the example above of directing one's attention at the pain in one's sore Achilles tendon.

However, important as these first order content shifts may be, even more important are those that are induced by the change in the state's meta-intentional context. The HOGS model appeals to "higher-order" global states, but the features we have been discussing so far focus mostly on the global or integrative aspect of the wider state into which the lower order state is recruited. Distributed neural correlates, cerebral celebrity and reciprocal amplification are important features, but all can be pretty much explained in terms of global integration without any mention of "higher-orderness". The HOGS model is intended as an alternative development of the higher-order theory, but so far global integration may seem to be doing all the work with little regard for meta-intentionality. If we do not want the HOGS model to collapse into a mere GS model we need to say something about how meta-intentional content plays an essential role; we need to show what role the HO plays in the HOGS model. I have sometimes been urged to forgo the higher-order model and to opt instead for either a straight forward global state model that appeals only to first order representation (FOR). However, I am not inclined to abandon the higher-order approach. I believe that despite the problems of its standard HOT and HOP versions, it has something important to teach us about consciousness. Thus I hope to show that the higher-order model and the global state model need not compete but can in fact be combined in a way that adds to each, which is just what the HOGS model aims to do.

4. The higher-order aspect of HOGS

The most difficult aspect of the HOGS model is explaining how its two main features fit together. In what respect do the global states it posits involve higher-order reflexive meta-intentionality? Or to put the question in acronymic terms: How does one get the HO in the HOGS? On both the HOT and HOP models the conscious-making meta-intentional component is carried by an added

higher-order state distinct and separate from its object. By contrast, on the HOGS model, the transition from nonconscious to conscious state is more a matter of transforming the content of the object state itself in a meta-intentional way by embedding it within a different systemic context.

To see how such a transition might occur, we first need to back up a bit and consider some more general issues about the nature of content and the factors that determine it. If one is a functionalist of some sort or another about mental states – as I am – then one regards a state's content as a function of its functional role within the system of which it is a part (Van Gulick 1980). What makes a state a belief that sugar is sweet or a desire for a cup of coffee is the role that it plays within one's internal mental economy, the network of states that guides our successful purposive interaction with our world. In a crude slogan, "content is a function of functional role." Though not everyone accepts a functionalist view of content, I will take it as a working hypothesis for the balance of this paper. Indeed, if I can use it successfully to help explain the meta-intentional aspect of consciousness that may provide an important further reason in support of the general view.

Given a functionalist view of content, we should not be surprised if the content of a state shifts when it is embedded within a significantly different context of interactions. A shift in systemic context does not always produce a shift in content, but it can and often does. If content is a function of functional role, then new contexts that induce states to play new roles may shift their content as well.

How does that general and abstract point apply to the specific case of recruiting lower-order states into a higher-order global ensemble? How does such recruitment alter their function, and might it change their content in ways that enable us to explicate the higher-orderness of the HOGS model?

The meta-intentional aspect that is added on the HOGS model is not a separate explicit meta-representation as on standard HOT and HOP models. Instead it is an implicit aspect of the structure of the globally integrated state into which the lower-order state is recruited. The general idea is that embedding the state in a larger more integrated context transforms its content in ways that involve an enhanced aspect of reflexive meta-intentionality. In what specific ways might the structure and organization of that global context embody meta-intentional aspects, and how in turn might they transform the contents of states embedded in it? At the neural level, those global ensembles are the substrates of the subject's transient flow of conscious experience, and thus it is to phenomenal experience (and especially to its coherence) that we should look for the nature of the relevant implicit meta-intentionality.

The global states underlying conscious experience supposedly display a high degree of integration and coherence, but what sorts of coherence are involved and how might they be relevant to the structure of experience and implicit meta-intentionality? Such global states may well cohere in a variety of straightforwardly physical ways, as does the coherent light in a laser. The global state may entrain the firing patterns of groups of neurons in diverse regions of the brain, perhaps in tandem with 40Hz oscillations as has been proposed on some models of intra-cortical binding. They may fire synchronously or exercise reciprocal causal influences on each other's activities in ways that bring them into some shared or harmonic pattern of physical action. But even if this is so, it would not by itself provide us with any sort of coherence that helps us with the meta-intentional aspects that we are trying to understand. At most, it may serve as the realization base for some more high-level features that provide the real explanation. Thus we should consider what types of coherence exist at the phenomenal level and whether they might better meet our explanatory needs.

Although this is not the place to do in-depth phenomenological analysis, we can perhaps get what we need by considering some of the most basic forms of phenomenological coherence. Two interdependent unities pervade the realm of phenomenal experience: the *unity of the experienced world* and the *unity of the experiencing self*. We do not normally experience isolated patches of color, sensations floating in a void or phantom shapes. Our phenomenal experience is of a world of independently existing and relatively stable objects located in a unified arena of space and time within which we ourselves are located as conscious perceivers. The objects we experience and the world within which they exist have a phenomenal and intentional "thickness", not merely in the literal three dimensional sense but in the richness of their properties and relations and in the diversity and density of the connections that hold among them in constituting "a world" of objects. On the correlative side of the experiential relation, is the more or less unified self that coheres as a single subject of experience both at a time and diachronically, bound by links of content, inference, access and memory.

As has been evident since the time of Kant, self and world are two inextricably bound aspects of phenomenal reality. The phenomenal or empirical world is one of objects perceived and experienced from a perspective or point of view, that of the self within that world. Correspondingly the self is always a self set over against a world of objects with which it is engaged. Each side of the phenomenal structure is incoherent without the other. Experience is always the experience *of a self* and *of a world of objects.* It is both the experience *of a self located within a world of objects*, and also *of a world of objects as they appear and*

present themselves from the point of view of the self. The phenomenal world and phenomenal objects are always apprehended from some self-like point of view, and conversely the phenomenal reality of such points of view consists largely in the access they afford for objects to appear or to present themselves within the experienced world.

The relevant notion of "perspective" here is broader and rich than that of the physical perspective from which the subject's sense organs monitor the environment, though that is of course one important aspect of the phenomenal perspective. The experiential viewpoint is more than the spatial locus from which we look and see. It includes as well those many aspects of the self through which it actively engages the objects of its experienced world. It that nexus of engagement and the density and diversity of modes of reciprocal access it provides that gives objective thickness to experienced reality. The thick phenomenology of experience as being *of a world of enduring independent objects* depends upon the richness of the viewpoint of the self to which they are present. And in turn the clarity and richness of that viewpoint depends on the complex structure of the experienced world that defines the self's location within it. Self and world are thus interdependent structures within experience, at least within anything that we can make empathetically coherent to ourselves as experience.

For illustration, consider some mundane stretch of ordinary experience, an everyday time slice of phenomenal reality. I sit at my desk, tapping at the keys and drinking strong hot coffee, but with my focus mostly on the words with which I make this sentence. A bright pink highlighter, though in my visual field for minutes, has only just now attracted my attention, which soon returns to the screen and then wanders off again to see the bubbled bud suspended almost frozen within a blue glass paperweight of supercooled transparent glass. The small flat bottomed globe has a phenomenal heft and thickness to match its physical weight and density. My experience of it is both "my experience" and "of it". It is appears to me from my point of view, from my location in physical, conceptual, and qualitative space. But it appears as well as an object within a world of objects: It is to the right of the keyboard beneath the lamp, darker – both more cobalt and more blackened – than the blue of the summer sky, made in China and purchased a while back on a clearance sale in a Pier One store, composed I know of silicates that flow too slowly for my eyes to see. And though I have not touched it on this day, I know by memory the heavy solid way that it would feel and the satisfying thunk that it would sound if dropped from just one half an inch above the desk. All that and more is part of the sense and content with which I meet it in my experience. Its weight is present

phenomenally as an implicit component of my experiential state, even though there is no explicit tactile aspect to my merely looking at it. The phenomenal content of experience extends far beyond what is explicitly present in sensation.

The example is banal as autobiography and uninspired perhaps as introspection, but nonetheless a clear display of the way in which experience is both of self and object, of objects that precipitate out of a world of dense connections apprehended from a single point of view. The example could be multiplied by millions with little effort, though with no great benefit. The one quick snapshot hopefully conveys the picture and the nature of the structure that repeats over and again in every phenomenal frame: self and world, self within world, world from point of view of self.

We can now see, at least in part, how and why the shift from the mere perception to conscious experience involves an increase in the quality and quantity of implicit meta-intentionality. To transform the processing and registration of visual information that had been unconscious (though perhaps available in various ways to guide action) into conscious visual experience, we must embed the information carried by the recruited state into the integrated structure of the phenomenal self-world. The global states that constitute the transient realizations of the flow of consciousness hang together and cohere at the level of phenomenal intentionality as experiential states jointly of world and self: of self located in a world of objects present to it, and of a world of objects present from that self's perspective.

One might object that this duality trades on a punning use of "of" that hides a privileged status for the self. Experience is *of a self* in the *constitutional* sense that each experiential state is ontologically dependent on the self of which is a state or mode, just as the states of the economy ontologically depend upon the economic systems of which they states. By contrast, experience may seem to be *of a world* in only an *intentional* sense that concerns its content and what if anything it aims at or represents.

But I would counsel against succumbing to the lure of this objection, attractive and plausible though it may seem. We must not fall into subjective idealism nor fail to remember what Kant taught us long ago about the roles of world and self in the phenomenal scheme. The *phenomenal self* is no less dependent on the intricately constructed organization of experience than are the *phenomenal objects* it confronts within it. The self as point of view has no reality apart from all that can appear from that perspective. The order of phenomenal reality is just as much a structure of the viewer as the viewed.

It is apt here to bring in the other strand of Dennett discussed above, that of the self as center of narrative gravity, the self as seeming point within the

story from which things cohere. Though stories are often told from a first person stance within the tale, one makes a bad mistake if one confuses the narrator with the author who constructs both narrative and narrator in one coherent interlinking structure. So too within the conscious realm and phenomenal reality; the two sides of the structure – world and self – cohere and intertwine in ways by which each resolves itself exhibiting an ordered organization that emerges only in their mutually inter-constituting linkage. The intentionality found at the phenomenal level – the intentionality of experience, appearance and presentness – exists within that overarching structure of world and self. And it is that structure, that context of meaning, within which individual experiences get embedded that is the main source of the reflexive meta-intentionality associated with conscious states. It derives not from the addition of distinct explicit self-ascriptive meta-states, but from the implicit self-perspectuality that is built into the intentional structure of conscious experience itself.

An important qualification is needed to avoid confusion. Although the notion of self that I am proposing is in part inspired by Dennett's idea of the self as the center of narrative gravity, it differs from his in significant ways. Most importantly, it is not solely, nor even primarily, connected with the serial personal level narrative that gets constructed on his theory. On Dennett's view, the self is the focus of this personal autobiographical tale that is spun out of the many strands of the multiple drafts model. That sort of self may indeed be a crucial aspect of human conscious experience, and may share key features with other influential ideas such as Michael Gazzaniga's notion of the interpretative self (Gazzaniga 1988). But the idea of self, I am proposing is intended to be more general, one that would apply not only to adult human beings, but to many non-human animals and to young pre-linguistic children.

I suppose it might be taken as a generalization of Dennett's idea. On his account, the self is the narrative center from which the serial narrative is told, whereas on the view I am offering the self is the perspective from which all of phenomenal experience is apprehended. It is not something outside experience, some homunculus or Cartesian ego that observes, but rather the perspectival point that is built into the structure of experience itself. The proposed model thus agrees with Dennett's view of self as perspective of the story rather than author of the tale, but the relevant "tale" is not just the serial narrative but all of phenomenal experience.

It is in that sense that the self, and even more importantly an implicit understanding of the self, is embedded in the very structure of phenomenal experience. The very act of having a conscious visual experience of the dark blue

paperweight requires at least an implicit understanding of one's status as the subject to whom it appears or to whom it is present. The intentional content of any phenomenal experience always implies the existence of the subject – not merely [blue globe] or even [blue globe here now], but [blue globe seen here/now by me], [blue globe appearing or being present now to me as part of my experienced world].

This is not to say that one is in a state like that one would be in if one explicitly thought or said those words to oneself. The self is not the explicit object of experience in the ordinary case – the dark blue paperweight is that – but both the self and its relation to the object are implicit in the structure of the state's phenomenal content. The meta-intentional aspect is built right into the first order content of the experiential state: a dark blue paperweight is present to me as part of my world, i.e. as part of the world that is present from my point of view, which is in turn as self defined by its location in that world of objects and appearance. That sort of implicit reference to self is an essential component of *phenomenal content*, if not of intentional content in general. It is part of what distinguishes my *experiencing* the paperweight from merely *representing* it.

One way to think of this is in terms of satisfaction conditions for phenomenal experience. That is, consider some extended stretch of a conscious creature's experiential life, and ask what conditions would have to hold in the world to satisfy the model of reality inherent in that sequence of experience. It will not be enough that the outer world contains any given set of objects and properties, for part of what that stretch of experience presents as real is also the existence of a conscious self to which those objects are present or to which they appear.

Thus we come full circle. The content of the perceptual experience is directed outward at objects in the world, but the experienced reality of that world, its objectivity, its thing-liness, its *res*-ality and *de re*-ness are phenomenally represented in the presentness of those objects as they appear as here and now from the perspective of the self. This experienced presentness phenomenologically captures or embodies their reality, and since such presentness is always a matter of presence to self (and from the self perspective), it follows that understanding the reality of experienced objects requires at least implicitly understanding the nature of the self to which they are present. It is in this respect that conscious experience embodies a significant degree of reflexive meta-intentionality in its very organization and structure.

5. Implicit meta-intentionality and phenomenal structure – a second time

The HOGS way of thinking about the meta-intentional aspect of consciousness is admittedly non-standard. Despite my efforts in the last section to motivate and explain it, I suspect it may still strike some readers as strange or obscure. So at the risk of repetition, it may be worthwhile to restate the line of argument again in slightly different terms. (Those who already feel convinced of the implicit higher-order aspect of experience, or who at least feel they understand clearly how such a result is supposed to follow from the HOGS model, may wish to jump ahead to the next section.)

The key idea is that the meta-intentionality of conscious states is not primarily a matter of their being accompanied by distinct explicit meta-states, but rather of their being embedded within a larger integrative context whose organization and structure implicitly embodies a significant degree of meta-intentional content. A central element of that structure is the interdependence within phenomenal reality of self and world. Thus I will reprise the general argument twice: once with the focus on the experienced world of objects and once with it on the experiencing self. Doing so will hopefully clarify some of the specific respects in which conscious experience embodies meta-intentionality in its structure and organization.

Let us look first from the world of objects angle. As a matter of phenomenology, we experience the reality of perceived objects as their being present to us here and now. We do not merely represent them as such we experience them as such; they appear to us transparently and directly. The dark blue paperweight is *experienced as present* to me at this very moment here and now. Experiential presence in that sense is an essential feature of the phenomenal content and of the intentional structure of the experience.

Thus to be capable of having (or undergoing) such experiences I must have at least an implicit understanding of that relation of presence. Experience is not a process of blind representation. It requires some understanding of what is being experienced; one cannot have (or undergo) an experience without a grasp of its content. Thus in so far as *presence* is essential to the content of perceptual experience, I could not have (or undergo) any such experiences without a sufficient grasp of that relation. But understanding the relation requires understanding that *to which* it is a relation, i.e. understanding it as a relation of *presence to self*. Thus we see that a key element of reflexive meta-intentionality is built into the very structure of conscious experience itself, not as an additional explicit meta-state but as implicit feature of its intentional and phenomenal

structure. Perhaps this is part of what Kant was referring to when he wrote that the "I think" must at least potentially accompany all our experiences (cf. Gennaro 1992).

Starting from the other pole of phenomenal reality, that of the experiencing self, one reaches a similar conclusion about the implicitly meta-intentional nature of experience. As a matter of phenomenology, experiences are always the experiences of some conscious subject who has or undergoes them. There could not be a visual experience of a blue circle nor an experience of pain without some self or conscious subject who has or undergoes them. The idea of an experience of pain without a self or subject to experience it seems incoherent. However, that dependence need not commit one to the existence of a substantial self or to a conscious subject that exists distinct from experience and prior to it. The relevant link may be internal to the phenomenal realm and a matter of intentional interdependence rather than one way ontological dependence.

As noted above, the phenomenal self is perhaps best regarded as an internal feature of intentional structure of experience, as the coherent unified viewpoint from which the world is experienced. The phenomenal self is thus a crucial element of the internal structure of experience rather than an entity that stands outside it and brings it into existence. For example, the fact of a given visual experience or a specific experience of pain as being *my experience* is built right into the phenomenal structure of the experience itself. Indeed it is difficult and perhaps even impossible to understand how something could count as an experience unless such a reference to an experiencing self were built into its intentional structure and content. As a matter of phenomenology, it does not seem something could count as an experience at all if its nature as something undergone by an experiencing self were not built right into its intentional structure and content. Part of what I understand when I undergo an experience is that I am undergoing it.

Thus even if one regards the self as a virtual entity internal to the structure of experience rather than as an external substantial entity, the essential link between experience and self remains: the existence of an experience requires the existence of an experiencing self that undergoes it. However, the dependence turns out not to be an ontological fact about a relation between experience and its external basis, but rather a phenomenological fact about the intentional structure of experience itself. Thus it provides another key respect in which reflexive meta-intentionality is built into the very structure of phenomenal experience. Nothing could count as a form of experience unless its intentional structure embodied at least implicitly the idea of those experiences as the experiences of an experiencing self that has or undergoes them.

Thus our two lines of phenomenological inquiry – one starting from the objective pole of phenomenal reality and the relation of presence, and the other beginning from the subjective pole and the dependence of experience on the experiencing self – both reveal important respects in which reflexive meta-intentionality is built into the very structure of phenomenal experience. Thus at least in these implicit ways, experiential states embody a heightened form of meta-intentionality. In that respect the move from nonconscious forms of mind to conscious experiential states depends in part upon an increase in degree and quality of meta-intentional content. When a nonconscious mental state is recruited and integrated into a global state of the sort posited by the HOGS model, its content is transformed by its embedding within that larger context. And since such global states are the neural realizations of phenomenal experiences, the content of the recruited state is specifically transformed in ways that reflect or embody the meta-intentionality implicit in the structure of the phenomenal experience into which it gets incorporated.

To avoid confusion, let me add that I am not claiming that being conscious in the experiential sense requires having explicit conceptualized self-understanding of the sort that could be expressed by linguistic self-reports such as "I am having an experience of a dark blue paperweight" or "A dark blue paperweight is now phenomenally present to me." The required self-understanding may be entirely implicit and not conceptualized in ways that make explicit inferential thought or linguistic report possible. Non-human animals and even young pre-linguistic children likely lack any such explicit conceptualized understanding, but they nonetheless clearly seem to undergo phenomenal experiences and have rich conscious inner lives. The sort of self-understanding that I claim they must have is that which is built into the very structure of experience and phenomenal intentionality.

Consider a simple example. When a dog that saw a rabbit go behind a bush sees it come out the other side, it does not need to think any explicit thought of the form, "The rabbit now again appears to me." But the dog experiences the rabbit as present here and now, and in its phenomenal world I believe it also makes the link between the rabbit present a moment ago and that which is now present. Part of the content of its phenomenal state (and thus part of the meaning of that state that it grasps in undergoing it) is that the object that has appeared twice to the same self is one and the same rabbit. Making the link between the remembered object and the currently perceived object is a matter as well of making at least an implicit link between the remembered experience and present experience, and thus as linking the two experiences as from a single "point of view" that of the same self. Thus even the dog, despite its

limited capacity for explicit conceptualized meta-thoughts, embodies implicit meta-understanding in the dynamic organization and intentional structure of its experience.

Though a lot more phenomenological inquiry would be needed to spell out all the important details, I hope that what I have offered here has sufficed to indicate the sorts of ways in which the GS global states of the HOGS model involve implicit Higher-Order content and thus to have shown how to put the HO into the HOGS.

6. HOGS and general objections to the HO view

Before closing let me turn briefly to consider how the HOGS model might handle the four interrelated objections raised above against both HOT and HOP versions of the theory. The four were as follows:

1. Though the higher order theory may suffice to explicate one sense of "conscious state" as "state we are aware of", it seems to many not to capture the crucial sense that corresponds to "state it's something like to be in."
2. The HOT and HOP models must rely on seemingly *ad hoc* restrictions, such as the noninferential and simultaneity requirements, to rule out obvious counter examples.
3. The Generality objection: x's being the object of an intentional state does not in general make x conscious. Why then should mental states as intentional objects be any different?
4. The problem of qualia which seem to get stranded on the standard versions. Although HOT and HOP theorists claim qualia are properties of the lower-order object state, they could be so only in an oddly strained sense of "qualia", since such lower-order states can have all the properties they do without there being anything it's like to be in them. Nor according to the HOT or HOP models do qualia occur as properties of higher-order states. Thus they seem to find no satisfactory home on either model.

The HOGS model offers a unified response to all four objections. On the HOGS model, a nonconscious perception gets transformed into a conscious visual experience by being recruited into an integrated global state that realizes a momentary episode of experience. On the HOGS model the transition is not a matter of merely adding a separate meta-state directed at the lower-order state, so cases that involve meta-states based on third person evidence simply

do not count as even *prima facie* counter examples. Thus there is no need to adopt seemingly *ad hoc* restrictions, such as the noninferential requirement.

Answers to the qualia and generality objections similarly turn on the role played by recruitment. A nonconscious visual perception of a blue globe becomes a visual experience of it when the active region of visual cortex is recruited into the global substrate of a conscious episode. The recruitment preserves much of the prior content but in an altered context that transforms it into an aspect of phenomenal reality, e.g. into a visually experienced property of an object present to me in my world. The color of the globe as it appears to me thus become a quale, a property of phenomenal objects, i.e. of objects of my experienced world that are present to me from my point of view within it. No entailments follow about qualia being merely subjective features, nor purely mental as sense data have been supposed to be. One might, as some have recently proposed (Alston 1999), explain appearing as a form of direct awareness and perceptual experience as a broad relation that includes the outer object as well as what goes on within the subject's head. I offer no opinion of such issues here, but only wish to note that taking qualia to be properties of objects that appear to us in experience by itself carries no commitment to they're being purely private or only "in the head."

The reply to the Generality Problem falls out immediately. What makes my perception into a conscious one is not my merely having a state about it, but its being recruited into my integrated experiential state. Thus one need not worry about "conscious pencils" unless pencils too can be recruited into global phenomenal states, a possibility that does not even seem to make sense as far as I can see. The fact that my having intentional states about pencils does not make them conscious poses no difficulty for the HOGS model since the model does not claim that one transforms a state into a conscious one by having a separate meta-state about it.

Given what it has to say about qualia and in reply to the Generality Problem, the HOGS model seems more promising as way of explaining the the notion of a conscious state in the "what it's like" sense. I certainly do not claim that it does the job by itself or that it thereby solves the so called "hard problem". But I think it moves us significantly in the right direction. It focuses us on the thick and densely integrated structure of phenomenal intentionality and the fundamentally dual aspect of experience as of both self and world. It thus gives us opportunities to try and understand how such a structure of phenomenal intentional content might be realized by the globally integrated states that seem to provide its physical substrate.

7. Conclusion

The alternative HOGS model I have offered here is not as yet a fully worked out theory. There remains a lot of work to do, and perhaps in the end it may not shed the light I hope it will. Most paths to understanding consciousness come to dead ends or go only a little way before disappearing into the trackless grass. Yet the HOGS model seems more promising than many recent contenders, and I believe it merits serious attention and development which I hope others will join in pursuing. The HOGS theory is not likely to give us the full story of consciousness, but it has the potential to make an important contribution. Understanding consciousness requires understanding so many different things in so many diverse ways, that no single model could meet all our explanatory needs. Nonetheless, the HOGS model offers an insightful perspective that enables us to see important links that might otherwise be missed. Though success in understanding consciousness is not likely in the end to be a wholly HOGS affair, it seems likely to be partly so.

References

Alston, W. (1999). Back to the theory of appearing. In J. Tomberlin (Ed.), *Philosophical Perspectives 13 Epistemology.* Oxford: Blackwell Publishing.

Armstrong, D. (1980). What is consciousness? In D. Armstrong (Ed.), *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Byrne, A. (1997). Some like it HOT. *Philosophical Studies, 86*, 103–129.

Dennett, D. (1991). *Consciousness Explained.* New York: Little Brown.

Descartes, R. (1644/1972). *The Principles of Philosophy.* In E. Haldane & G. Ross (Eds.), *The Philosophical Works of Descartes.* London: Cambridge University Press.

Dretske, F. (1993). Conscious experience. *Mind, 102*, 263–283.

Gazzaniga, M. (1988). *How the Mind and the Brain Interact to Create our Conscious Lives.* New York: Houghton Mifflin.

Gennaro, R. (1992) Consciousness, self-consciousness, and episodic memory. *Philosophical Psychology, 5*, 333–347.

Gennaro, R. (1996) *Consciousness and Self-Consciousness.* Amsterdam and Philadelphia: John Benjamins Publishing Co.

Hill, C. (1991). *Sensations.* New York: Oxford University Press.

Locke, J. (1688/1959). *An Essay Concerning Human Understanding*, annotated by A. C. Fraser. New York: Dover.

Lycan, W. (1990). *Mind and Cognition*. Oxford: Basil Blackwell.

Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.

Lycan, W. (1996). *Conscious Experience.* Cambridge, MA: MIT Press.

Rosenthal, D. (1991). The independence of consciousness and sensory quality. In E. Villanueva (Ed.), *Consciousness: Philosophical Issues 1* (pp. 15–36). Atascadero, CA: Ridgeview Publishing.

Rosenthal, D. (1992). Thinking that one thinks. In M. Davies & G. Humphreys (Ed.), *Consciousness*. Oxford: Basil Blackwell.

Van Gulick, R. (1980). Functionalism, information and content. *Nature and System, 2*, 139–162. Reprinted in Lycan (1990).

Van Gulick, R. (2001). Inward and upward: Reflection, introspection and self-awareness. *Philosophical Topics, 28*, 275–305.

Chapter 5

The superiority of HOP to HOT

William G. Lycan

> [T]he perceptual model does not withstand scrutiny.
>
> David Rosenthal (1997: 740)

What is consciousness? – to coin a question. According to "higher-order representation" (HOR) theories of consciousness, a mental state or event is a conscious state or event just in case it (itself) is the intentional object of one of the subject's mental representations.

That may sound odd, perhaps crazy. In fact, because of the richly diverse uses of the word "conscious" in contemporary philosophy of mind, it is bound to sound odd to many people. So I must begin by specifying what I here mean by "conscious state or event" (hereafter just "state," for convenience).

1. The explanandum

A state is a conscious state iff it is *a mental state whose subject is (directly or at least nonevidentially) aware of being in it*. For the duration of this paper, the latter biconditional is to be taken as a stipulative definition of the term "conscious state." I think that definition is not *brutely* stipulative or merely technical, because it records one perfectly good thing that is sometimes meant by the phrase, as in "a conscious memory," "a conscious decision." But it is necessary to be careful and painfully explicit about usage, because the phrase "conscious state" has also been used in at least two entirely different senses, as respectively by Dretske (1993) and Block (1995). Failure to keep these and other senses straight not only can lead but has led to severe confusion.)

To come at my target subject-matter from a slightly different angle, mental or psychological states fall roughly into three categories: (a) States whose subjects are aware of being in them, such as felt sharp pains, or reciting a poem

carefully to oneself in an effort to recall its exact words. (b) States whose subjects are not aware of being in them, but could have been had they paid attention; you realize that you have been angry for some time but were unaware of it, or you do not notice hearing a tiny whine from your computer monitor until you concentrate on the quality of the ambient noise. (c) States that are entirely subterranean and inaccessible to introspection, such as language- or visual-processing states and just possibly Freudian unconscious ones. A theory is needed to explain the differences between states of these three kinds. Such a theory is what *I here* shall call a theory "of consciousness" – though there is no known limit to the variety of intellectual structures that have somewhere or another been called "theories of consciousness" (see Lycan 2002).

2. HOR theories

A "higher-order representation" theory of consciousness in the foregoing sense is one according to which: what makes a conscious state conscious is that the state is the intentional object of, or is represented by, another of the subject's mental states, suitably placed or employed.[1] Given our definition of "conscious state," that idea is not so strange after all. On the contrary; it may now seem too obvious: If I am aware of being in mental state M, then tautologically, my awareness itself is an awareness *of*, and that "of" is the "of" of intentionality; my state of awareness has M as its intentional object or representatum. Therefore, for M to be a conscious state is (at least) for M to be represented by another of my mental states, QED.

It is easy to see how HOR theories explain the differences between the foregoing three types of mental state. States of type (a) are states which are the objects of actual higher-order representations. States of type (b) are ones which are not the objects of actual higher-order representations, but which could have been and perhaps come to be so. States of type (c) are those which, for reasons of our psychological architecture, cannot be internally targeted by person-level metarepresentations ("internally" because, of course, cognitive psychologists come to represent the processing states of subjects in the course of third-person theorizing).

In addition to the argument given two paragraphs back, and the fact that HOR theories do explain the differences between our three types of mental state, there are other features that recommend HOR theories. I have expounded them in previous works (chiefly Lycan 1996: 15ff.), so I shall only mention a couple of them here. For example: Such theories explain a nagging ambiguity

and/or dialect difference regarding sensations and feeling; and they afford what I believe is the best going account of "knowing what it's like" to experience a particular sort of sensation (Lycan 1996: Ch. 3).

But of course further issues arise, chiefly about the nature of the putative higher-order representations. Notoriously, HOR theories subdivide according to whether the representations are taken to be perception-like or merely thought-like. According to "inner sense" or "higher-order perception" (HOP) versions, a mental state is conscious just in case it is the object of a kind of internal scanning or monitoring by a quasi-perceptual faculty (Armstrong 1968, 1981; Lycan 1987, 1996). But "higher-order thought" (HOT) theorists contend that the state need be represented only by an "assertoric" thought to the effect that one is in that sort of state (Rosenthal 1986, 1990, 1991b, 1993, 1997; Gennaro 1996; Carruthers 1996, 2000).

As is hinted by my title, my main purpose in this paper is to argue that HOP versions of HOR are better motivated and more promising than HOT versions.[2] Which I believe to be true; but I have written the paper reluctantly. For at least ten years I have meaning to write a piece exhibiting the superiority of HOP to HOT, but until now I have always put it off, because at bottom I regard the two views as more allies than competitors, both of them sensibly resistant to some of the craziness that is out there about consciousness. I doubt that their differences matter very much. But the quote from Rosenthal that heads this paper is inspiring.

Before proceeding to the task, though, we should note the main objections to HOR theories tout court.

3. Objections to HOR theories

In particular, it is worth heading off two bad and ignorant criticisms. These two have been scouted more than once before, but I have found that they still come up every time a HOR theory is propounded. First is the *Qualia-Creation* objection: "You say that what makes a state conscious is that the state is the intentional object of a higher-order perception or thought. But surely the mere addition of a higher-order representation cannot bring a sensory or phenomenal quality into being. If the original state had no qualitative or phenomenal character, the HOR theories do not in any way explain the qualitative or phenomenal character of experience; a mere higher-order representation could hardly bring a phenomenal property into being. So how can they claim to be theories of *con*sciousness?"

But it is no claim of either HOP or HOT per se to have explained anything about qualitative character; they are theories only of the distinction between mental states one is aware of being in and mental states one is not aware of being in. Some other theory must be given of the original qualitative character of a sensory state, such as the yellowy-orangeness of an after-image, the pitch of a heard tone, or the smell of an odor. (Neither Armstrong, Rosenthal nor I have ever suggested that a mere higher-order thought or even a higher-order quasi-perception could explain the sensory core of experience. Each of us has given an account of sensory qualities, but elsewhere.[3])

Second, the *Regress* objection: "If the second-order representation is to confer consciousness on the first-order state, it must itself be a conscious state; so there must be a third-order representation of it, and so on forever."

But HOP and HOT theorists reject the opening conditional premise. The second-order representation need not itself be a conscious state. (Of course, it may be a conscious state, if there does happen to be a higher-order representation of it in turn.)

Of course, HOR theories face other objections that are more serious, four in particular; I have tried to answer them in previous works.

(1) Some critics, such as Sydney Shoemaker (1994), have complained that HOR awards introspection too great a degree of fallibility. Also, HOR theories predict the theoretical possibility of false positives: Could I not deploy a *nonveridical* second-order representation of a nonexistent sensation? – and so, I might seem to myself to be in intense pain when in fact there was no pain at all, which seems absurd. Karen Neander (1998) has prosecuted an especially incisive version of this objection. I have rebutted Shoemaker's argument in Lycan (1996: 17–22), and I have replied to Neander, not altogether satisfactorily, in Lycan (1998). I argued against Shoemaker that introspection is certainly fallible to some degree; having been elaborately primed, the fraternity pledge mistakes the cold sensation produce on his bare skin by the ice cube for burning heat. And there are reasons why the egregious sorts of false positive cited by Neander would not happen.

(2) Georges Rey (1983) argued that higher-order representation comes too cheap, and that laptop computers can deploy higher-order representations of their first-order states; yet we would hardly award consciousness in any sense to a laptop. In response, Lycan (1996: 36–43) argued that consciousness in the present sense comes in degrees, and that *if* we suppose (as Rey's argument must) that laptop computers have mental states in the first place,

there is no reason to deny that some of those states are conscious to a very low degree.

(3) Carruthers (2000) has complained that HOR theories require a grade of computational complexity that lower animals and probably even human subjects do not attain. His argument assumes that at any given time, many of our mental states are conscious states; Lycan (1999) rejoins by rejecting that assumption.

(4) HOR theorists' standard reply to the Qualia-Creation objection separates matters of consciousness in the present sense (mental states one is aware of being in) from matters of qualia and qualitatative character, and accordingly HOR theorists hold that a qualitative perceptual or sensory state may go entirely unnoticed; thus there are, or could be, unfelt pains, and many other sensations that are entirely nonconscious. But to some people the idea of a sensation whose subject is unaware of having it is problematic to say the least.

Unfelt pains and nonconscious sensations generally have been defended at length by Armstrong (1968), Palmer (1975), Nelkin (1989), Rosenthal (1991a), Lycan (1996) and others.[4] (The matter is complicated by a dialect difference, and the dispute over nonconscious sensations is at least partly verbal; see Lycan (1996: 16–21).) But Joseph Levine (2001) has pushed a version of the present objection that I have not seen addressed in print, so I shall confront it here.

Levine considers three mental states involving visual redness:

> my state as I deliberately stare at the [red] diskette case, my perception of a red light while driving on 'automatic pilot,' and the clearly unconscious state of my early visual processing system that detects a light intensity gradient. According to HO[R], the first two states both instantiate the same qualitative character, and for the second and third states there is nothing it is like to occupy them; they are both unconscious, non-experiences. One oddity, then, is the fact that, on HO[R], the very same qualitative feature possessed by my perception of the diskette case can be possessed by a state that is as unconscious as the intensity gradient detector.
>
> Furthermore, does the intensity gradient detector itself possess qualitative character? HO[R] faces a dilemma. To say such states have qualitative character seems to rob the notion of any significance. To deny them qualitative character requires justification. What one would normally say is that to have qualitative character there must be something it is like to occupy a state, and the qualitative character is what it's like. But the advocate of HO[R] can't say this, since states there is nothing it is like to occupy have qualitative character on this view. (106)

I am not sure why Levine finds it odd that the same qualitative feature possessed by his perception of the diskette case can be possessed by a state that is "as unconscious as" the intensity gradient detector. Indeed, I am not sure how he intends the detector example in the first place, as an example of a nonconscious mental state that has a red qualitative character in virtue of its intensity-gradient detection, or rather as an example of a nonconscious psychological state that has no qualitative character. In light of the alleged dilemma, Levine does not seem to decide between these two interpretations, though I believe he intends the second, because he is trying to bring out the absurdity he sees in the idea of a qualitative state of which one is entirely unaware.

Moving on to the dilemma: What are its horns? That (a) one either says that the gradient detector state has qualitative character, and thereby "seems to rob the notion of any significance," or (b) one denies that the gradient detector state has qualitative character, which "requires justification." The first of those conditionals suggests that the detector is interpreted in the second way, as not being in any qualitative state, the idea being that if we say that the gradient detector state has qualitative character, we might as well say that a pancreatic state or a state of a brick has qualitative character. I am not inclined to choose horn (a), so I turn to the second conditional.

Since Levine himself seems to want to deny that the gradient detector state has qualitative character, I suspect what he really thinks requires justification is that the automatic-pilot visual state does have qualitative character when the detector state admittedly does not: "What one would normally say is that to have qualitative character there must be something it is like to occupy a state, and the qualitative character is what it's like." The argument is then, "but the advocate of HO[R] can't say this, since states there is nothing it is like to occupy have qualitative character on this view."

That argument is guilty of equivocation. For both expressions, "qualitative character" and "what it's like" (as well as "phenomenal character/property"), are now ambiguous and in just the same way. Each has been used to mean the sort of qualitative property that characteristically figures in a sensory experience, such as yellowy-orangeness, pitch, or smell as mentioned above – call these "Q-properties." Each has also been used to mean the higher-order property of "what it's like" for the subject to experience the relevant Q-property.

To see that these are quite distinct, notice: (i) The higher-order "what it's like" property is higher-order; it is a property *of* the relevant Q-property. (ii) A Q-property is normally described in one's public natural language, while what it is like to experience that Q-property seems to be ineffable. Suppose you are having a yellowy-orange after-image and Jack asks you, "How, exactly, does the

after-image look to you as regards color?" You reply, "It looks yellowy-orange." "But," persists Jack, "can you tell me, descriptively rather than comparatively or demonstratively, what it's like to experience that 'yellowy-orange' look?" At that point, if you are like me, words fail you; all you can say is, "It's like. . .*this*; I can't put it into words." Thus, the Q-property, the subjective color itself, can be specified in ordinary English, but what it is like to experience that Q-property cannot be.

(The distinction, which is also nicely elaborated by Carruthers (2000), is obscured by some writers' too quickly defining "sensory quality" in terms of "what it's like," and by others' using the phrase "what it's like" to mean merely a Q-property; Dretske (1995) and Tye (1995) do the latter.)

I hold that Levine's own argument further illustrates the distinction. We champions of nonconscious sensory qualities have argued that a Q-property can occur without its subject being aware of it. In such a case, as Levine says, there is a good sense in which it would not be like anything for the subject to experience that Q-property. (Of course, in the Dretske-Tye sense there would be something it was like, since the Q-property itself is that. For what it is worth, I think Levine's usage is better.) So even in the case in which one is aware of one's Q-property, the higher-order type of "what it's like" requires awareness and so is something distinct from the Q-property itself.

Thus, Levine is right that the advocate of HOR cannot say that "to have qualitative character there must be something it is like to occupy a state, and the qualitative character is what it's like." And no one else should say that either, if by "qualitative character" they mean a Q-property. For we HORists hold that one can be in a state featuring a Q-property without being aware of it, hence without there being anything it is like to be in that state. And unless I have missed it, Levine has provided no good argument against that claim.

4. My HOP theory

Armstrong states the Inner Sense doctrine as follows.

> Introspective consciousness...is a perception-like awareness of current states and activities in our own mind. The current activities will include sense-perception: which latter is the awareness of current states and activities of our environment and our body. (1981:61)

As I would put it, consciousness is the functioning of internal *attention mechanisms* directed upon lower-order psychological states and events. I would

also add an explicit element of teleology: Attention mechanisms are devices which have the *job* of relaying and/or coördinating information about ongoing psychological events and processes.

But that is not all. To count in the analysis of my consciousness in particular, the monitor must do its monitoring *for me*. A monitor might have been implanted in me somewhere that sends its outputs straight to to the news media, so that the whole world may learn of my first-order mental states. Such a device would be functioning as a monitor, but as the media's monitor rather than mine. More importantly, a monitor functioning within one of my subordinate homunculi might be doing its distinctive job for that homunculus rather than for me; e.g., it might be serving the homunculus' proprietary event memory rather than my own event memory. This distinction blocks what would otherwise be obvious counterexamples to HOP as stated so far.[5]

5. Rosenthal's objection to HOP

To justify his curt dismissal, Rosenthal (1997: 740) makes a direct argument against HOP, roughly: While perceiving always involves some sensory quality, an inner sensing would itself involve some sensory quality. But attention to one's own sensory state does not involve any such quality over and above that of the sensory state itself.

I reply (as in Lycan 1996: 28–29) by granting the disanalogy. No HOP theorist has contended that inner sense is like external-world perception in every single respect. Nor, in particular, should we expect inner sense to involve some distinctive sensory quality at its own level of operation. The reason is that "outer" sense organs have the function of feature detection. The sensory properties involved in first-order sensory states are, according to me (Lycan 1987: Ch. 8; 1996: Ch. 4), the represented features of physical objects; e.g., the color featured in a visual perception is the represented color of a (real or unreal) physical object. Being internal to the mind, first-order states themselves do not have ecologically significant features of that sort, and so we would not expect internal representations of first-order states to have sensory qualities representing or otherwise corresponding to such features. As Rosenthal himself puts it, "Whereas a range of stimuli is characteristic of each sensory modality, [first-order] mental states do not exemplify a single range of properties" (p. 740).

Rosenthal will reply that if inner sense is disanalogous in that way from external perception, what is the positive analogy? ("[O]therwise the analogy with perception will be idle" (p. 740).) And that brings me, finally, to my announced

agenda: I mean to show that the higher-order representation of mental states is more like perception than it is like mere thought.[6]

6. HOP vs. HOT

6.0 Intuitive priority

(I number this point "0" because it is not really an argument, but only a preludial sniff.) One would suppose that awareness psychologically preceded thinking, in that if X is real and "aware" is taken to be factive, S can think about X only if S is independently (if not previously) aware of X. (Of course I here mean "aware" in its ordinary, fairly general sense, not anything like direct or perceptual awareness.) It is hard to imagine the reverse, S thinking about X and *thereby* becoming aware of X. By contrast, to perceive X is precisely a way of becoming aware of X.

No doubt Rosenthal would reply that I am inappropriately importing a model based on thinking about external objects; awareness of those does require epistemic contact prior to thought. But it does not follow that the same holds for our awareness of our own mental states. Perhaps when S is in an attentive frame of mind, S's mental state M itself simply causes in S the thought that S is in M; no intermediating sort of awareness is required. Fair enough. (And yet, what is the role of the "attentive" frame of mind? "Attentive" means, disposed to attend – which again suggests that S first attends, and then has thoughts about what attending has turned up.)

6.1 Phenomenology

Wittgenstein and Ryle pooh-poohed the etymology of "introspection," contending that the metaphor of "looking inward" to "see" one's own mental states is a poor metaphor and an even worse philosophical picture of consciousness. But I say it is at least a fine metaphor. When we attend to our own mental states, it feels like that is just what we are doing: focusing our internal attention on something that is there for us to discern. Now, from this fact about introspect*ing*, it does not follow that the phenomenology of normal state consciousness is also one of quasi-perceptual discernment, because in normal state consciousness we are not doing any active introspecting, but are only passively and usually nonconsciously aware of our first-order states. But even in the passive case, as Van Gulick (2001: 288–289) points out, our first-order states are

phenomenologically *present to* our minds and not (or not just) represented by them, much as in external vision, physical objects are present to us without seeming to be represented by us.[7] (It takes philosophical sophistication to see that vision really is representation; indeed, some philosophers still dispute the latter thesis. So too, it takes philosophical sophistication to reject Descartes' conviction that our own conscious states are simply and directly present to our minds, rather than represented by them.)

Moreover, consider what happens when you have only been passively aware of being in a mental state M but are then moved, by curiosity or by someone else's suggestion, to introspect. You both attend more closely to M and become aware of your previously passive awareness. But the sense of presence rather than representation remains. As Byrne (1997: 117) observes, in such a case you do not suddenly notice one or more *thoughts* about M, but only the discernment of M through M's presence to you. (I emphasize that all this is intended as phenomenology, not as insisting that how things seem is how they are. The phenomenology of presence may be illusory. My claim is only that it *is* the phenomenology of awareness of one's own mental state.)

6.2 Voluntary control

Consciousness of our mental states is like perception, and unlike ordinary thought, in that we have a considerable degree of control over what areas of our phenomenal fields to *make* conscious. This is particularly true of active introspecting, as I have emphasized elsewhere (Lycan 1996, 1999). If I ask you to attend first to the feeling in your tongue, then to the sounds you can hear, then to the right center of your visual field, and then to what you can smell right now, you can do those things at will, and in each case you will probably make a sensory state conscious that may not have been conscious before.

Here again, that introspect*ing* is a highly voluntary activity does not entail that ordinary passive representation of first-order states is as well. But it seems to show that the introspectors, the monitors or scanners, are there to be mobilized and that they may function in a less deliberate way under more general circumstances. I contend, then, that the higher-order representations that make a first-order state conscious are (etiologically) more like perceptions than they are like thoughts. They are characteristically the outputs of an attention mechanism that is under voluntary control; thoughts are not that.

Carruthers (2000: 212–213) has trenchantly objected to one of my previous appeals to the voluntary control of introspecting. (That appeal (Lycan 1996: 16–17) was made in defense of a different thesis, that our brains actually

do contain mechanisms of internal attention.) He suggests that "when we shift attention across our visual field, this is mediated by shifting the first-order attentional processes which are at work in normal perception, and which deliver for us richer contents in the attended-to region of visual space. . ." Carruthers proposes that what one does in response to the command, "Shift your attention to the right center of your visual field" is exactly what one would do in response to "Without changing the direction in which you are looking, shift your visual attention to the state of the world just right of center." "The process can be a purely first-order one."

As a staunch representationalist about first-order sensory states (Lycan 1987, 1996), I am in no position to insist on a contrast between attending to one's Q-properties and attending to the real or putative perceptible features of external objects; I believe Q-properties *just are* the real or putative perceptible features of external objects. However, notice that it is not just the intentional qualitative contents of my first-order states that I can voluntarily attend to. I can also focus my attention, at will, on further properties of those contents, and properties of the containing state as well. (Thus, I deny the full-blown "transparency" thesis defended by Tye (2002).) I introspect a green patch in my visual field. Let us grant that that is in part to detect external, physical greenness; but that property itself determines nothing about any sensory modality, much less any finer-grained mode of presentation. I also introspect that the greenness is visual rather than something I learned of in some other way. I can also tell (fallibly) by introspection how compelling a visual experience it is, how strongly it convinces me that there is a real green object in front of me.

Or consider pains. I am firmly convinced by Armstrong (1968) and Pitcher (1970) that pains are representational and have intentional objects, real or inexistent, which objects are unsalutary conditions of body parts. But those objects are not all I can introspect about a pain. I can also introspect its awfulness and the urgency of my desire that it cease. (I distinguish between the Q-property of a pain, the pain's specifically sensory core – say the throbbing character of a headache – and the pain's affective aspect that constitutes its awfulness/hurtfulness (Lycan 1998). Those are not normally felt as distinct, but two different neurological subsystems are responsible for the overall experience, and they can come apart.[8] The Q-property is what remains under morphine; what morphine blocks is the affective aspect – the desire that the pain stop, the distraction, the urge to pain-behave. I contend that the affective components are functional rather than representational.)

Finally, for any Q-property, I can introspect the higher-order property of what it is like to experience that Q-property.

The moral is that I can focus my inner attention more finely (even) than on particular Q-properties, and in particular on purely mental properties that, unlike the Q-properties themselves, are not just features of the external world.

In any case, notice that although Rosenthal makes free (and perfectly reasonable) use of the notions of introspection and introspective awareness, he has no obvious account to give of introspect*ing*, the voluntary and substantive sort of activity I have described. The having of thoughts on a given subject-matter is only partly, and not directly, under voluntary control.

Rocco Gennaro has offered a HOTist reply to the current argument. It takes the form of a dilemma: Either (a) the relevant higher-order representation is itself conscious or (b) it is not. Suppose (a). HOP has no obvious advantage over HOT for this case, Gennaro says, because the HOT theorist can equally talk of a subject S's "actively searching for" an object of the higher-order thought, or "deliberately thinking about" such an object. (Indeed, we often do such things in what HOT theorists might call "reflection.") As in the anticipated reply to point 0 above, perhaps when S is in an attentive frame of mind, S's first-order mental state itself simply causes the thought that S is in that state, with no intermediating sort of awareness required.

Now suppose (b), that the relevant higher-order representation is not a conscious one. Then, when S controls where to focus S's attention, that does not seem to be the result of S's controlling S's unconscious higher-order representations. Even though S does have significant voluntary control over S's first-order perceptual states, the higher-order representations produced by the attention mechanism are in no way contributing to the voluntariness in question. Thus, even if we agree that the relevant higher-order representations are *produced by* a mechanism that is perception-like, there is no reason to think that *what is* produced by such a mechanism is more perception-like than thought-like. And so HOP has no advantage over HOT for case (b) either. Net result: No relevant advantage for HOP over HOT.

I am not persuaded on either count. For case (a), I do not concede that the HOT theorist can *equally* talk of "active searching" etc. Compare the voluntary control of first-order attending to external objects: At will, we can selectively attend to environmental region R and see whatever there is in R. We do not *in the same facile way* control what things in the environment we have thoughts about; thought is more spontaneous and random. The only obvious way in which we control what to have thoughts about is first to attend (perceptually) to a region R and *thereby* cause ourselves to have thoughts about whatever there is in R.

The same goes for the voluntary control of attending to first-order mental states. At will, we can selectively attend to phenomenal region R and detect whatever sensory qualia there are in R. We do not in the same facile way control what regions of or things in the phenomenal field we have thoughts about; the only obvious way in which we control what to have thoughts about is first to attend to a region R and thereby cause ourselves to have thoughts about whatever sensory qualia there are in R. I can *try* to have thoughts about contents of R only by attending to R and detecting qualia there.

Anent case (b): Agreed, S's control of where to focus S's attention is not the result of S's controlling S's unconscious higher-order representations, and my argument does not show that the representation produced is *in its own nature and structure* more perception-like. (So far as has been shown, representations spit out by the attention mechanisms and representations that just well up thought-like from the marshmallow may be otherwise just alike – say, neural states having the structures of predicate-calculus formulas.) But I want to say that even if the putative higher-order perceptions and higher-order thoughts are thus "intrinsically" alike, they still differ importantly and distinctively in their etiological properties: The relevant higher-order representations are characteristically produced by the exercise of attention. That makes them more like perceptions than like thoughts, since it is not characteristic of thoughts to be directly produced by the exercise of attention (though of course thoughts can happen to be produced by the exercise of attention, normally by way of a mediating perception).

6.3 Nonvoluntary results

There is a sort of obverse point to be made in terms of voluntariness. It is characteristic of external-world perception that, once the subject has exerted her/his voluntary control in directing sensory attention, e.g., chosen to look in a particular direction or to sniff the air inside a cupboard, the result is up to the world. Perhaps what one then sees or smells is conditioned in some ways by expectation or by interest, but for the most part we must see what is there in that direction to be seen and smell whatever miasma the world has furnished. The same is true of awareness of our own mental states. Though there too, awareness is to some degree conditioned by expectation (recall the frat-boy type of example), the first-order states that present themselves to our attention are primarily as they are independently of our will. If you concentrate on your right lower canine, you may find a slight ache there, or you may feel nothing

there, but in either case (normally) that is up to your sensory system and what it has or has not delivered.

Van Gulick (2001: 286–287) argues that the case of higher-order thought is not actually so different. He reminds us that the thoughts appealed to by HOT theorists are *assertoric* thoughts. "Once the assertoric requirement comes to the fore, our degree of voluntary control seems to shrink if not altogether disappear," because it is controversial at best to suppose that we can control our beliefs. We should agree that if I choose to direct my attention to certain nonperceptual but factual questions, I will nonvoluntarily be caused to make assertoric judgments one way or the other. (What is my daughter's name? – Jane. Is foreign-language study required for a Ph.D. in philosophy at my university? – No. How fast does light travel in a vacuum? – Around 186,000 mps.)

But in those cases, I *already* know, or am confident of, the answers. On questions I have not previously investigated, if I raise them and do not thereupon investigate (perceptually or otherwise), I will normally not be confronted by a fait accompli. How many people are sitting right now in the Carolina Coffee Shop? What piece of music is being played on our local classical radio station? Who was Vice-President of the United States in 1880? Occasionally, I may consider a novel question and the answer may force itself upon me, in the manner of a thought experiment. (If we lower admissions requirements and increase the student body by 4,000, will our average student be better or worse?) But that sort of case is not the norm, and this still distinguishes thought from perception, though perhaps the distinction is now less in nonvoluntariness per se than in the comparative *range of* nonvoluntariness: For perceptual investigation, the answer comes right up nonvoluntarily almost every time, but for thought-eliciting questions the percentage is a lot smaller.

6.4 Degrees of awareness

As Lycan (1996: 39–43) emphasized, awareness of our own mental states comes in degrees. Given a mild pain that I have, I may be only very dimly and peripherally aware of it (assuming I am aware of it at all); or I may be more than dimly aware of it though still only mutedly; or I may be well aware of it; or I may be quite pointedly aware of it, thank you; or in a fit of hypochondria I may be agonizingly aware of it and aware of little else. This range of possibilities is of course characteristic of external-world perception as well. It is not characteristic of mere thought. That is not to deny that the HOT theorist might construct a notion of degree-of-awareness; for example, the *number* of distinct higher-

order thoughts I have about my pain might be invoked. The point is only that the HOP theorist has such a notion already and does not need to construct one.

6.5 Epistemology

Our awareness of our own mental states *justifies* our beliefs about them. Indeed, my only justification for believing that I now have a slight ache in my tailbone and that I am hearing sounds as of a car pulling into the driveway is that I am aware of both states. Active introspection justifies our beliefs about our mental states even more strongly, though by no means infallibly. This is just what we should expect if awareness is a perception-like affair. By contrast, merely having an assertoric thought to the effect that one is in state M by itself does nothing to justify the belief that one is in M, and having a metathought about that thought adds no justification either.

Van Gulick (2001: 280–281) puts a reliabilist spin on this contrast. We think of our perceptual capacities as reliable channels of information, and, barring evil-demon and other skeptical scenarios, for the most part they are. We do not think of the having of thoughts, per se, as reliable information channels, but count a thought as justified or justifying only when it is itself the output of some other reliable channel – paradigmatically a perceptual one. Introspective awareness is like perception in that respect: We think of our internal awareness as a reliable source of information, and (again barring skeptical scenarios) for the most part it is.

6.6 Grain

Thoughts, as ordinarily conceived, differ from perceptual states in being more thoroughly and more discretely conceptual. Their contents are reported in indirect discourse using complement clauses made of natural-language words, and it is fair to say that we usually think in concepts that correspond to, and are naturally expressed by, words. But words are fairly coarse-grained representations. Now, consider a phenomenal region that may be an object of consciousness for you at a time, say, a subarea of your visual field. Even if it is not a particularly large region, it is rich, in irregular outlines, textures, gradations of shading and the like. The phenomenal contents of this region would be either impossible to describe in words at all, or possible to describe only in that one could devise a description if one had gabillions of words and days or weeks of time in which to deploy them all. (Byrne (1997: 117) refers to roughly these disjuncts, respectively, as "the inexpressibility problem" and "the problem of the

unthinkable thought.") Unless thoughts are significantly less "conceptual" and subtler than sentences of natural languages, your consciousness of the contents of the phenomenal region cannot be constituted by a higher-order thought or set of same. Perhaps thoughts are more subtle than sentences in some helpful respect, but I think it is up to the HOT theorist to make that case.

Rosenthal emphasizes (1997: 742, 745) that in areas of subtle sensory discrimination, such as wine-tasting and music appreciation, an increase in conceptual sophistication often makes for new experiences, or at least for finer-grained experiences than were previously possible. He contends, and I agree, that this is a mark in favor of HOT theories. But it does not help him against the present argument in particular. The nuances of wine tastes and of musical experience still outrun the verbalizable.

One might suggest on HOT's behalf that a conscious state need not be *fully* verbalizable. Indeed Rosenthal fleetingly acknowledges the problem (1993: 210): "No higher-order thoughts could capture all the subtle variations of sensory quality we consciously experience." He responds, "So higher-order thoughts must refer to sensory states demonstratively, perhaps as occupying this or that position in the relevant sensory field." But, taken literally, that will not do; as Byrne points out (117–118), one can *demonstrate* only what one is already aware of, or some thing determinately related to another relevant thing one is aware of. (Byrne adds that relaxing the requirement still further by allowing the higher-order thought to designate the first-order state by any directly referring term will not help either, because to be aware of being in a state is to be aware of it in some characterizing way from one's first-person point of view, not just to token a directly referring name, such as "Alice," that in fact designates the state.)

6.7 The ineffability of "what it's like"

What is it like to experience the phenomenal yellowy-orangeness of a yellowy-orange after-image? As I said in Section III above, we cannot generally express such things in words. HOP explains that ineffability. When you introspect your after-image and its phenomenal color, your quasi-perceptual faculty mobilizes an introspective concept, one that is proprietary to that introspector and (for a reason I have given elsewhere (Lycan 1996: 60–61))) does not translate into English or any other natural language. That is why you cannot say what it is like for you to experience the visual yellowy-orangeness, though you still, and rightly, feel that there *is* something it is like and that its being like that is a fact if only you could express it verbally.

HOT affords no comparable explanation. The HOT theorist would agree that the way in which you are aware of what it is like to experience the greenness is by deploying a higher-order representation of the after-image and its color, but s/he has no systematic explanation of the ineffability. (Though of course it is entirely *compatible* with HOT that some or all of the relevant higher-order thoughts might be inexpressible in natural language.) Rosenthal's passing contention that higher-order thoughts must refer to sensory states demonstratively would help if it were tenable, but Byrne's point remains, that one can demonstrate only what one is already aware of or what is determinately related to something else one is aware of.

6.8 Purely recognitional concepts

Brian Loar (1990) and others have argued for the existence of "purely recognitional" concepts, possessed by subjects and applying to those subjects' experiences. Such concepts classify sensations without bearing any analytic or otherwise a priori connections to other concepts easily expressible in natural language, and without dependence on the subject's existing beliefs about the sensation in question. (It is in part because we have concepts of this sort that we are able to conceive of having a sensation of *this* kind without the sensation's playing its usual causal role, without its representing what it usually represents, without its being physical at all, etc., and we are able to conceive of there being a body just like ours in exactly similar circumstances but is not giving rise to a sensation of *this* kind – all facts fallaciously made much of by opponents of materialism.)

Carruthers (2001: Section 3) notes that HOP can explain how it is possible for us to acquire such phenomenal concepts. "For if we possess higher-order perceptual contents, then it should be possible for us to learn to recognize the occurrence of our own perceptual states immediately – or 'straight off' – grounded in those higher-order analog contents. And this should be possible without those recognitional concepts thereby having any conceptual connections with our beliefs about the nature or content of the states recognized, nor with any of our surrounding mental concepts." HOT can make no parallel claim about the acquisition of recognitional concepts, at least not straight away. For one thing, how could we have a higher-order thought about a sensory state without already having the pertinent concept? But in any case, merely having a thought about something is not generally recognized as a means of *acquiring* the concept of that thing.

6.9 HOT's problem of sufficiency

Rosenthal has always acknowledged a prima facie problem about higher-order thoughts that are *mediated* in the wrong ways. For example, I might learn of a nonconscious mental state that I am in by noting my own behavior, or through testimony from someone else who has been observing my behavior, or through Freudian analysis; and so I could then have thoughts about that state even though it is not a conscious one. To preëmpt that objection, Rosenthal requires of a conscious state that its subject's higher-order thought not be the result of "ordinary inference."

For some reason, several commentators have reported that Rosenthal has imposed a stronger, causal requirement as well, e.g., that the higher-order thought be directly caused by the first-order state. To my knowledge Rosenthal has never done that. However, on the basis of a series of hypothetical cases, Francescotti (1995) contends that he should have. Further, Francescotti argues that this gets Rosenthal in trouble, in that the causal requirement must be either too weak or too strong.

No doubt chisholming will ensue, and perhaps Rosenthal can solve the problem. But HOP has no such problem in the first place.

I do not regard any or all of the foregoing as a proof that HOP is superior to HOT. I do think I have shown that HOP can withstand considerably more scrutiny than at least the leading proponent of HOT has been prepared to allow.[9]

Notes

1. As in Lycan (2001), I ignore the dubious possibility that the state is its own intentional object, represented by itself alone. I know of no HOR theory that does not understand a higher-order representation as being numerically distinct from its representatum. (Though on Gennaro's (1996) "Wide Intrinsicality View" and on Van Gulick's (2001) "Higher-Order Global State" picture, the relevant higher-order states are not *entirely* distinct from the first-order states they target.)

2. My opponent here will be what Carruthers (2000) calls "actualist" HOT, principally Rosenthal's own version. I shall not address Carruthers own "dispositionalist" HOT, because I believe it is quite a different sort of theory and incurs different sorts of objections. And see his paper in this volume.

3. E.g., Armstrong and Malcolm (1984); Rosenthal (1991a); Lycan (1996: Ch. 4). It is true that Rosenthal and I have pressed our respective HOR views into service in helping to explain other things about consciousness more generally, including the phenomenon of its

being "like" something to experience a sensory quality, but we did that only by conjoining the views in an ancillary way with other, independent theoretical apparatus.

4. Armstrong's famous example is that of the long-distance truck driver who is daydreaming and driving on autopilot; the driver must have perceived stop lights as red, for example, or he would not have stopped at them. This example has become a poster child for HOP theories, but I now believe incorrectly so; I do not think it is an example of a subject who *characteristically* lacks higher-order perceptions. See Lycan and Ryder (2003).

5. One such is offered by Levine (2001:106): States of the early visual system that detect intensity gradients are (obviously) not conscious states. "[B]ut it's not obvious that there aren't higher-order states of the requisite sort within the computational system that carries out visual processing."

6. Güzeldere (1997) offers another argument specifically against my version of HOP (the view he calls "Option 2" on its first interpretation (p. 794)). The argument is hard for me to parse, because – understandably – it is stated in terms of Armstrong's truck-driver example, which I now reject (note 4 above). He points out, in my view correctly, that higher-order representation of perceptual states would not help explain how those brain states constitute the representings of external objects that they do. More generally, he warns against the fallacy of confusing features of the representer with the represented. But these points are immanent to his diagnosis of what he and I both consider an error of Armstrong's about the truck driver. They do not extrapolate to HOP per se. If Güzeldere has given any further argument against HOP itself, I have missed it.

7. In his article (279), Van Gulick does us the excellent service of listing some paradigm features of perceiving, and asking the question of which of those features are shared by the meta-mental states that HOP and HOT theorists agree make for consciousness.

8. Besides a sensory, nociceptive system, it seems there is also an inhibitory system that for occasional reasons damps the nociceptive signals. See Hardcastle (1999) and the references given therein.

9. Thanks to Zena Ryder for very detailed comments on a previous draft.

References

Armstrong, D. M. (1968). *A materialist theory of the mind*. London: Routledge and Kegan Paul.

Armstrong, D. M. (1981). What is consciousness? In *The Nature of Mind and Other Essays*. Ithaca, NY: Cornell University Press.

Block, N. J. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18,* 227–247.

Block, N. J., O. Flanagan, & G. Güzeldere (Eds.). (1997). *The nature of consciousness*. Cambridge, MA: Bradford Books/MIT Press.

Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies, 86,* 103–129.

Carruthers, P. (1996). *Language, thought and consciousness*. Cambridge: Cambridge University Press.
Carruthers, P. (2000). *Phenomenal consciousness*. Cambridge: Cambridge University Press.
Carruthers, P. (2001). Higher-order theories of consciousness. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/archives/fall1999/entries/consciousness-higher/>.
Dretske, F. (1993). Conscious experience. *Mind, 102,* 263–283; reprinted in Block, Flanagan and Güzeldere (1997).
Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: Bradford Books/MIT Press.
Francescotti, R. M. (1995). Higher-order thoughts and conscious experience. *Philosophical Psychology, 8,* 239–254.
Gennaro, R. (1996). *Consciousness and self-consciousness*. Amsterdam & Philadelphia: John Benjamins.
Güzeldere, G. (1997). Is consciousness the perception of what passes in one's own mind? In Block, Flanagan and Güzeldere (1997).
Hardcastle, V. G. (1999). *The myth of pain*. Cambridge, MA: MIT Press.
Levine, J. (2001). *Purple haze*. Oxford: Oxford University Press.
Loar, B. (1990). Phenomenal properties. In J. Tomberlin (Ed.), *Philosophical Perspectives: Action theory and philosophy of mind*. Atascadero, CA: Ridgeview Publishing.
Lycan, W. G. (1986). *Consciousness*. Cambridge, MA: Bradford Books/MIT Press.
Lycan, W. G. (1995). *Consciousness and experience*. Cambridge, MA: Bradford Books/MIT Press.
Lycan, W. G. (1998). In defense of the representational theory of qualia (replies to Neander, Rey and Tye). In Tomberlin (1998).
Lycan, W. G. (1999). A response to Carruthers' 'Natural theories of consciousness'. *Psyche, 5,* <*http://psyche.cs.monash.edu.au/v5*/psyche-5-11-lycan.html>.
Lycan, W. G. (2001). A simple argument for a higher-order representation theory of consciousness. *Analysis, 61,* 3–4.
Lycan, W. G. (2002). The plurality of consciousness. *Philosophic Exchange, 32,* 33–49. A briefer version will appear in J. M. Larrazabal & L.A. Perez Miranda (Eds.), *Language, knowledge, and representation* (Dordrecht: Kluwer Academic Publishing).
Lycan, W. G. & Z. Ryder (2003). The loneliness of the long-distance truck driver. *Analysis, 63,* 132–136.
Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. In Tomberlin (1998).
Nelkin, N. (1989). Unconscious sensations. *Philosophical Psychology, 2,* 129–141.
Palmer, D. (1975). Unfelt pains. *American Philosophical Quarterly, 12,* 289–298.
Pitcher, G. (1970). Pain perception. *Philosopical Review, 79,* 368–393.
Rey, G. (1983). A reason for doubting the existence of consciousness. In Davidson, Schwartz, & Shapiro (Eds.), *Consciousness and self-regulation, Vol. 3* (pp. 1–39). New York: Plenum Press.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49,* 329–359.
Rosenthal, D. (1990). A theory of consciousness. Report No. 40, Research Group on Mind and Brain, Zentrum für Interdisziplinäre Forschung (Bielefeld, Germany).

Rosenthal, D. (1991a). The independence of consciousness and sensory quality. In Villanueva (1991).
Rosenthal, D. (1991b). Explaining consciousness. Unpublished MS, presented at the Washington University conference in Philosophy of Mind (December, 1991).
Rosenthal, D. (1993). Thinking that one thinks. In M. Davies & G. Humphreys (Eds.), *Consciousness*. Oxford: Basil Blackwell.
Rosenthal, D. (1997). A theory of consciousness. In Block, Flanagan and Güzeldere (1997). (An expanded and updated version of Rosenthal (1990).)
Rowlands, M. (2001). Consciousness and higher-order thoughts. *Mind and Language, 16,* 290–310.
Shoemaker, S. (1994). Self-knowledge and 'inner sense', Lecture II: The broad perceptual model. *Philosophy and Phenomenological Research, 54,* 271–290.
Tomberlin, J. E. (Ed.). (1998). *Language, mind, and ontology (Philosophical Perspectives, Vol. 12)*. Atascadero, CA: Ridgeview Publishing.
Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: Bradford Books/MIT Press.
Tye, M. (2002). Representationalism and the transparency of experience. *Noûs, 36,* 137–151.
Van Gulick, R. (2001). Inward and upward: Reflection, introspection, and self-awareness. *Philosophical Topics, 28,* 275–305.
Villanueva, E. (Ed.). (1991). *Philosophical issues, I: Consciousness*. Atascadero, CA: Ridgeview Publishing.

Chapter 6

HOP over FOR, HOT theory

Peter Carruthers

Following a short introduction, this chapter begins by contrasting two different forms of higher-order perception (HOP) theory of phenomenal consciousness – inner sense theory versus a dispositionalist kind of higher-order thought (HOT) theory – and by giving a brief statement of the superiority of the latter. Thereafter the chapter considers arguments in support of HOP theories in general. It develops two parallel objections against both first-order representationalist (FOR) theories and actualist forms of HOT theory. First, neither can give an adequate account of the distinctive features of our recognitional concepts of experience. And second, neither can explain why there are some states of the relevant kinds that are phenomenal and some that aren't. The chapter shows briefly how HOP theories succeed with the former task. And it then responds (successfully) to the challenge that HOP theories face the latter charge too. In the end, then, the dispositionalist HOT version of HOP theory emerges as the overall winner: only it can provide us with a reductive explanation of phenomenal consciousness which is both successful in itself and plausible on other grounds.

1. Introduction

I should begin by explaining the bad joke that forms my title. (It is bad because it does need some explanation, unfortunately.) On the one hand, I shall be arguing in this chapter for the superiority of higher-order perception (HOP) theories over both first-order representationalist (FOR) and actualist higher-order thought (HOT) theories. (That is, I shall be arguing that *HOP* theories win out *over* both *FOR* theory and actualist *HOT theory*.) But on the other hand, I shall be arguing that the theory on which we should all converge (the theory that we should all *hop over for*) is actually a dispositionalist form of *HOT*

theory (a form of HOT theory that, when combined with consumer semantics, can also count as a kind of HOP theory, as we shall see).

The topic of this chapter is phenomenal consciousness: the sort of conscious state that it is *like something* to have, or that has *feel*, or *subjective phenomenology*. More specifically, this chapter is about whether (and how) phenomenal consciousness can be reductively explained, hence integrating it with our understanding of the rest of the natural world. I shan't here pause to distinguish phenomenal consciousness from other forms of state-consciousness (specifically, from various forms of *access*-consciousness), nor from a variety of kinds of *creature*-consciousness; for these distinctions have been adequately drawn elsewhere, and should by now be familiar (Rosenthal 1986; Block 1995; Lycan 1996; Carruthers 2000: Ch. 1). Nor shall I pause to consider 'mysterian' arguments that phenomenal consciousness lies beyond the scope of reductive explanation (McGinn 1991; Chalmers 1996; Levine 2000). And accounts that attempt to explain phenomenal consciousness directly in terms of neurology or brain-function (e.g. Crick & Koch 1990) are similarly excluded from discussion. (For direct critiques of both mysterian and neurological approaches to consciousness, see Carruthers 2000: Chs. 2–4.) Somewhat more narrowly, then, this chapter is concerned with attempts to provide a reductive explanation of phenomenal consciousness in terms of some combination of *intentional* (or representational) *content* and *causal* (or functional) *role*.

Representationalist theories of phenomenal consciousness can be divided into two broad categories, each of which then admits of several further subdivisions. On the one hand there are first-order theories of the sort defended, in different ways, by Kirk (1994), Dretske (1995) and Tye (1995, 2000). (For discussion of a number of other variants on this first-order theme, see Carruthers 2000: Ch. 5.) Such theories reduce phenomenal consciousness to a certain sort of intentional content (*analog* or *fine-grained*, perhaps; or maybe *nonconceptual* – these differences won't concern us here) figuring in a distinctive place in the causal architecture of cognition (perhaps as the output of our perceptual faculties, *poised* to have an impact on conceptual thought and behavior control). And then on the other hand there are a variety of higher-order theories that reduce phenomenal consciousness to some sort of higher-order awareness of such first-order analog / non-conceptual intentional states.

Here is one way of carving up the different forms of higher-order representational accounts of phenomenal consciousness. The basic contrast is between theories that claim that the higher-order states in question are themselves perceptual or quasi-perceptual, on the one hand, and those that claim that they are conceptualized thoughts, on the other. Higher-order perception

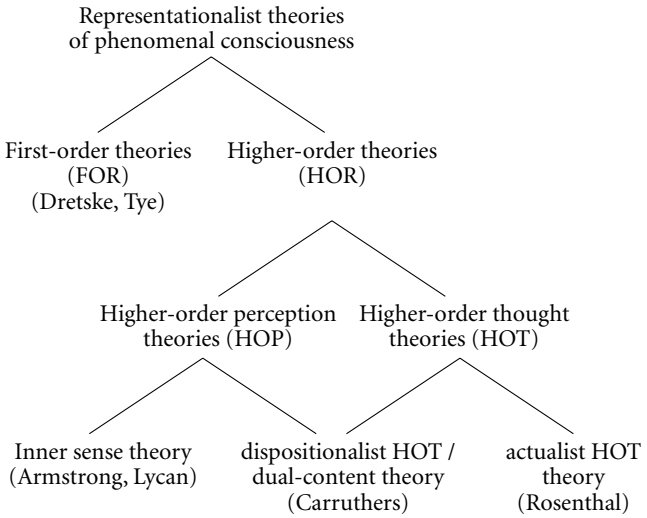

Figure 1. Representationalist theories of phenomenal consciousness

(HOP) theories propose a reduction of phenomenal consciousness to analog / non-conceptual intentional content which is itself the target of (higher-order) analog / non-conceptual intentional contents (Armstrong 1968; Lycan 1996; Carruthers 2000). Actualist higher-order thought (HOT) theory, on the other hand, reduces phenomenal consciousness to analog / non-conceptual contents which are the actual target, at the time, of a higher-order belief or thought. Or otherwise put, actualist HOT theory reduces phenomenal consciousness to analog / non-conceptual contents of which the subject is conceptually aware (Rosenthal 1986, 1993, 1997).

One somewhat surprising thesis to be advanced in the present chapter is that both FOR theories and actualist HOT theories (which superficially look very different from one another) turn out to be subject to quite similar kinds of difficulty. In point of fact, essentially the same arguments that can be used to defeat the one can also be used to defeat the other. Which then leaves HOP theory as the only representationalist account left standing. But HOP theory, too, admits of a pair of sub-varieties, one of which turns out to be, at the same time, a (dispositionalist) form of HOT theory. This is where we begin, in Section 2. But the range of different representationalist alternatives can be seen laid out in Figure 1.

2. Two kinds of HOP theory

The form of higher-order perception (HOP) theory that will be familiar to most people is so-called 'inner sense' theory, generally credited to John Locke (1690). It was reintroduced in our era by Armstrong (1968), and has been defended more recently by Lycan (1987, 1996). On this account, we not only have a set of first-order senses charged with generating analog / non-conceptual representations of our environments and the states of our own bodies, but we also have a faculty of *inner* sense, which scans the outputs of those first-order senses and generates higher-order analog / non-conceptual representations of (some of) them in turn. And while terminology differs, it would seem that it is these higher-order representations that are responsible for the *feel* of our phenomenally conscious states.[1] That is to say, our first-order perceptual states get to be phenomenally conscious by virtue of being targeted by higher-order perceptions, produced by the operations of our faculty of inner sense.

The contrasting, less familiar, form of HOP theory is a dispositionalist version of HOT theory (Carruthers 2000), although it might equally be called a 'dual-content theory'. On this account, some of our first-order perceptual states acquire, at the same time, a higher-order analog / non-conceptual content by virtue of their availability to a faculty of higher-order thought (HOT), combined with the truth of some or other version of consumer semantics – either teleosemantics, or functional / conceptual role semantics.[2] (It is because it proposes a set of higher-order analog – or 'experiential' – states, which represent the existence and content of our first-order perceptual states, that the theory deserves the title of 'higher-order *perception*' theory, despite the absence of any postulated *organs* of higher-order perception.)

There is no faculty of 'inner sense' on this account; and it is one and the same set of states that have *both* first-order *and* higher-order analog / non-conceptual contents. Rather, a set of first-order perceptual states is made available to a variety of down-stream 'consumer systems' (Millikan 1984), some concerned with first-order conceptualization and planning in relation to the perceived environment, but another of which is concerned to generate higher-order thoughts, including thoughts about those first-order perceptual states themselves. And it is by virtue of their availability to the latter consumer system that the perceptual states in question acquire a dual content. Besides being first-order analog / non-conceptual representations of redness, smoothness, and so on, they are now also second-order analog / non-conceptual representations of seeming-redness, experienced-smoothness, and so forth; hence acquiring a

dimension of *subjectivity*. And it is this dimension that constitutes those states as phenomenally conscious, on this account.

How can we adjudicate between these two very different versions of HOP theory? There are a pair of significant problems with inner sense theory. One is that it is very hard to see any evolutionary reason for the development of an organ of inner sense. Yet such a faculty would be by no means computationally trivial. Since it would be costly to build and maintain, we need a good story about the adaptive benefits that it would confer on us in return. But in fact there are no such stories on offer. All of the various proposed functions of inner sense turn out, either not to require inner sense at all, or to presuppose a faculty for higher-order thought (HOT), or both (Carruthers 2000). In contrast, it isn't difficult for dispositionalist HOT theory to explain why a HOT faculty should have evolved, nor why it should have access to perceptual contents. Here the standard stories about the adaptive benefits of sophisticated social – and perhaps 'Machiavellian' – thinking will surely suffice (Byrne & Whiten 1988, 1998; Carruthers 2000).

The other main problem with inner sense theory is that it ought to be possible for such a sense-organ to malfunction, just as our other senses sometimes do (Sturgeon 2000). I can be confronted with a surface that is actually red, but which (due to unusual lighting conditions, or whatever) I perceive as orange. So, too, then, it ought to be possible for me to be undergoing an experience with the first-order analog / non-conceptual content *red* while my inner-sense faculty is producing the higher-order analog / non-conceptual content *seems orange* or *experienced orange*. In such circumstances I would be disposed to make the first-order recognitional judgment, 'It is red' (spontaneously, without inferring that the surface is red from background knowledge or beliefs about my circumstances), while at the same time being inclined to say that my experience of the object *seems orange to me*. Yet nothing like this ever seems to occur.[3]

In contrast once again, no parallel difficulty arises for dispositionalist HOT theory. For it is one and the same state that has both first-order and higher-order analog / non-conceptual content, on this account. (So there can be no question of the higher-order content existing in the absence of the first-order one.) And the higher-order content is entirely parasitic upon the first-order one, being produced from it by virtue of the latter's availability to a faculty of higher-order thought. There therefore seems to be no possibility that these contents could ever 'get out of line' with one another. On the contrary, the higher-order analog / non-conceptual state will always be a *seeming* of whatever first-order analog / non-conceptual content is in question.

There are difficulties for inner sense theory that don't arise for dispositionalist HOT theory, then. Are there any comparable costs that attend the dispositionalist HOT version of HOP theory? Two are sometimes alleged; but neither seems to me very real or significant. It is sometimes said in support of inner sense theory that this approach makes it more likely that phenomenal consciousness will be widespread in the animal kingdom (Lycan 1996). Whereas it is rightly said that dispositionalist HOT theory will restrict such consciousness to creatures capable of higher-order thought (humans, and perhaps also the other great apes). But this alleged advantage is spurious in the absence of some account of the evolutionary function of inner sense, which might then warrant its widespread distribution. And our temptation to ascribe phenomenal consciousness quite widely amongst non-human animals is easily explained as a mere by-product of our imaginative abilities (Carruthers 1999, 2000), and/or by our failure to be sufficiently clear about what really carries the explanatory burden when we explain other people's behavior by attributing phenomenally conscious states to them (Carruthers forthcoming).

The other 'cost' of preferring dispositionalist HOT theory to inner sense theory is that we are then required to embrace some form of consumer semantics, and must give up on any pure causal-covariance, or informational, mere input-side semantics. But this strikes me as no cost at all, since I maintain that all right-thinking persons should embrace consumer-semantics as at least one determinant of intentional content, quite apart from any considerations to do with phenomenal consciousness (Botterill & Carruthers 1999).

I conclude, then, that once the contrast is clearly seen between inner sense theory and dispositionalist HOT / dual-content versions of higher-order perception (HOP) accounts of phenomenal consciousness, then the latter should emerge as the winner overall. For there are powerful arguments against inner sense theory, while there exist no significant arguments against dispositionalist HOT theory (which aren't just arguments against the higher-order character of the account, which both approaches share, of course).

This result is important, since many people are inclined to reject HOP accounts of phenomenal consciousness too easily. In fact, they see the weaknesses in inner sense theory without realizing that there is an alternative form of HOP theory (dispositionalist HOT theory plus consumer semantics) which isn't really subject to those problems. The remainder of this chapter will now argue in support of HOP approaches in general, as against both first-order representationalist (FOR) and actualist HOT accounts. Combining those arguments with the points made briefly in the present section will then amount to an overall argument in support of a dispositionalist HOT form of HOP theory.

3. Explaining higher-order recognitional judgments

There is something of a consensus building amongst philosophers opposed to 'mysterian' approaches to phenomenal consciousness. It is that the right way to undermine the various thought experiments (zombies, inverted experiences, and such-like) that are supposed to show that phenomenal properties don't supervene logically on physical, functional, or intentional facts, is to appeal to our possession of a set of *purely recognitional concepts* of experience (Loar 1990, 1997; Papineau 1993, 2002; Sturgeon 1994, 2000; Tye 1995, 2000; Carruthers 2000).

The idea is that we either have, or can form, recognitional concepts for our phenomenally conscious experiences that lack any conceptual connections with other concepts of ours, whether physical, functional, or intentional. I can, as it were, just recognize a given type of experience as *this* each time it occurs, where my concept *this* lacks any conceptual connections with any other concepts of mine – even the concept *experience*. My possession of the concept *this* can consist in nothing more nor less than a capacity to recognize a given type of phenomenal state as and when it occurs.[4]

Given that I possess such purely recognitional concepts of experience, then it is easy to explain how the philosophical thought experiments become possible. I can think, without conceptual incoherence or contradiction, '*This* type of state [an experience *as of* red] might have occurred in me, or might normally occur in others, in the absence of any of its actual causes and effects; so on any view of intentional content that sees content as tied to normal causes (i.e. to information carried) and/or to normal effects (i.e. to teleological or inferential role), *this* type of state might occur without representing redness.' Equally, I can think, '*This* type of state [an experience] might not have been, or might not be in others, an *experience* at all. Rather it might have been / might be in others a state of some quite different sort, occupying a different position within the causal architecture of cognition.' Even more radically, I can think, 'There might have been a being (a zombie) who had all of my physical, functional, and intentional properties, but who lacked *this* and *this* and *that* – indeed, who lacked any of *these* states.'

Now, from the fact that we have *concepts* of phenomenally conscious states that lack any conceptual connections with physical, functional, or intentional concepts, it of course doesn't follow that the *properties* that our purely recognitional concepts pick out aren't physical, functional, or intentional ones. So we can explain the philosophical thought experiments while claiming that phenomenal consciousness is reductively explicable in physical, functional, or in-

tentional terms. Indeed, it increasingly looks to me, and to others, that any would-be naturalizer of phenomenal consciousness needs to buy into the existence of purely recognitional concepts of experience.

Higher-order perception (HOP) theorists of phenomenal consciousness are well placed to explain the existence of purely recognitional concepts of experience. We can say the following. Just as our first-order analog perceptual contents can ground purely recognitional concepts for secondary qualities in our environments (and bodies), so our higher-order analog perceptual contents can ground purely recognitional concepts for our first-order experiences themselves. The first-order perceptual contents *analog-green*, *analog-smooth* and so on can serve to ground the recognitional concepts, *green*, *smooth* and so forth. Similarly, then, the higher-order perceptual contents *analog-experienced-green* and *analog-experienced-smooth* can serve to ground the purely recognitional concepts of experience, *this state* and *that state*. And such concepts are grounded in (higher-order) awareness of their objects, just as our recognitional concepts *green* and *smooth* are grounded in (first-order) awareness of the relevant secondary properties.

Neither first-order (FOR) theories of the sort defended by Dretske (1995) and Tye (1995, 2000), nor actualist higher-order thought (HOT) theories of the kind proposed by Rosenthal (1993, 1997) can give an adequate account of our possession of purely recognitional concepts of experience, however. Or so I shall briefly argue. (For further elaboration of some of these arguments, especially in relation to FOR theories, see Carruthers 2004a.)

According to FOR theories, phenomenal consciousness consists in a distinctive kind of content (analog or non-conceptual) figuring in a distinctive position in cognition (poised to have an impact upon thought and decision making, say). Such contents are appropriate to ground first-order recognitional applications of concepts of secondary qualities, such as *green*, *smooth*, and so on. But what basis can they provide for higher-order recognition of those first-order experiences themselves? The perceptual content *analog-green* can ground a recognitional application of the concept *green*. But how could such a content ground a recognitional application of the concept *this* [experience of green]? It isn't the right *kind* of content to ground an application of a higher-order recognitional concept. For if such concepts are to be applied recognitionally, then that means that they must be associated with some analog or non-conceptual presentation of the properties to which they apply. And that means, surely, a higher-order analog content or HOP.

One option for a FOR theorist here would be to say, as does Dretske (1995), that the higher-order concept applies to experience indirectly, via recognition

of the property that the experience is an experience *of*. On such an account the putative recognitional concept *this* [experience *as of* green] is really a concept of the form, *my experience of this* [green]. But this then means that the concept is not, after all, a purely recognitional one. On the contrary, it is definitionally tied to the concept *experience*, and also to the presence of greenness. And then we can no longer explain the seeming coherence of the thoughts, '*This* [experience *as of* green] might not have been an experience, and might not have been *of this [green].*'

Another option for a FOR theorist would be to defend a form of *brute-causal* account, as Loar (1990) seems tempted to do. On this view the higher-order recognitional concept *this* [experience *as of* green] wouldn't have the quasi-descriptive content assumed by Dretske. Rather, applications of it would be caused by the presence of the appropriate kind of experience [*as of* green] without the mediation of any mental state, and more specifically, without the mediation of any higher-order perceptual state. But this view gets the phenomenology of higher-order recognitional judgment quite wrong. When I judge recognitionally, 'Here is *this* type of experience again', I do so on the basis of *awareness of* that which my judgment concerns – a given type of experience. I do not, as it were, judge *blindly*, as the brute-causal account would have it.

Finally, a FOR theorist might allow that we do have higher-order perceptions (HOPs) to ground our recognitional concepts of experience, while denying that this is what constitutes those experiences as phenomenally conscious ones. (Tye 1995, sometimes seems tempted to adopt a position of this sort.) On the contrary, it might be said, all first-order analog perceptual contents are phenomenally conscious, but only some of these are targeted by higher-order perceptual contents in such a way as to ground purely-recognitional concepts of experience.

One problem with this proposal, however, is that it requires us to accept the existence of phenomenally conscious states that are inaccessible to their subjects. That is, it requires us to believe that there can be phenomenally conscious states of which subjects cannot be aware. For as I shall note briefly in Section 4 below, there is now robust evidence for the existence of perceptual systems whose outputs are inaccessible to consciousness, in the sense of being unavailable to higher-order awareness or verbal report. And it is one thing to claim that there can be phenomenally conscious states that we happen not to be aware of through other demands on our attention (some have wanted to describe the 'absent minded driver' type of example in these terms), but it is quite another thing to claim that there are phenomenally conscious states in

us that we *cannot* be aware of, or that we are *blind to*. This would be very hard to accept.

Another difficulty with the proposal is that it appears to confuse together two distinct forms of subjectivity. Any first-order perceptual state will be, in a sense, subjective. That is, it will present a subjective take on the organism's environment, presenting that environment in one way rather than another, depending on the organism's perspective and its discriminatory abilities. Thus any form of perception will involve a kind of subjectivity in the way that the world is presented to the organism. But phenomenal consciousness surely involves a much richer form of subjectivity than this. It involves, not just a distinctive way in which the world is presented to us in perception, but also a distinctive way that our perceptual states themselves are presented to us. It isn't just the world that seems a certain way to us, but our experiences of the world, too, appear to us in a certain way, and have a distinctive *feel* or phenomenology. And this requires the presence of some form of higher-order awareness, which would be lacking in the first-order representational (FOR) proposal made above.

FOR theories face severe difficulties in accounting for our possession of purely-recognitional concepts of experience, then. Similar difficulties arise for actualist higher-order thought (HOT) theory, of the sort defended by Rosenthal (1993, 1997). On this account, an experience gets to be phenomenally conscious by virtue of the subject's conceptual awareness of the occurrence of that experience – that is to say, provided that the experience causes (and causes immediately, or non-inferentially) a higher-order thought to the effect that such an experience is taking place. But in the absence of any higher-order perceptual contents to ground such higher-order thoughts, this approach provides just another version of the 'brute-causal' account discussed above, and suffers from the same difficulties.

Specifically, actualist HOT theory cannot account for the way in which our higher-order thoughts about our experiences appear to be grounded in some sort on non-conceptual awareness of those experiences. Nor, in consequence, can it explain how purely recognitional (higher-order) concepts of experience are possible which preserves their similarity to (first-order) recognitional concepts of color. Just as my judgments of 'green' are grounded in perceptual awareness of greenness (guided in their application by the content *analog-green*), so too my judgments of 'this state' are grounded in awareness of the state in question, which requires that they should be guided by a higher-order perceptual content such as *analog-experienced-green*.

According to actualist HOT theory, there is a sense in which my recognitional judgments of experience are made *blindly*.[5] I find myself making higher-order judgments about the occurrence of experience, but without those judgments being grounded in any other awareness of those experiences themselves. It is rather as if I found myself making judgments of color (e.g. 'Red here again') in the absence of any perceptual awareness of color. But higher-order judgment doesn't appear to be like that at all. When I think, 'Here is *that* experience again', I think as I do because I am aware of the experience in question. I can reflect on the appropriateness of my judgment, given the properties of the experience, for example. This requires the presence of higher-order perceptions of that experience – namely, HOPs.

4. Why some states are phenomenal and some aren't

One difficulty for both first-order (FOR) theories and actualist higher-order thought (HOT) theories of phenomenal consciousness, then, is that neither can account adequately for the existence of purely recognitional judgments of experience. Another, equally powerful, objection is that neither can explain why some perceptual states are phenomenal and some aren't. That is to say, neither can give an adequate account of that in virtue of which some analog / non-conceptual states have the properties distinctive of phenomenal consciousness and some don't. But in order to make this point, I first need to say just a little about the conscious / non-conscious distinction as it applies to perceptual states.

The evidence for non-conscious perceptual states in all sensory modalities is now quite robust (Baars 1997; Weiskrantz 1997). Here let me concentrate on the case of vision. Armstrong (1968) uses the example of absent-minded driving to make the point. Most of us at some time have had the rather unnerving experience of 'coming to' after having been driving on 'automatic pilot' while our attention was directed elsewhere – perhaps day-dreaming or engaged in intense conversation with a passenger. We were apparently not consciously aware of any of the route we have recently taken, nor of any of the obstacles we avoided on the way. Yet we must surely have been *seeing*, or we would have crashed the car. Others have used the example of blindsight (Weiskrantz 1986; Carruthers 1996). This is a condition in which subjects have had a portion of their primary visual cortex destroyed, and apparently become blind in a region of their visual field as a result. But it has now been known for some time that if subjects are asked to *guess* at the properties of their 'blind' field (e.g. at whether

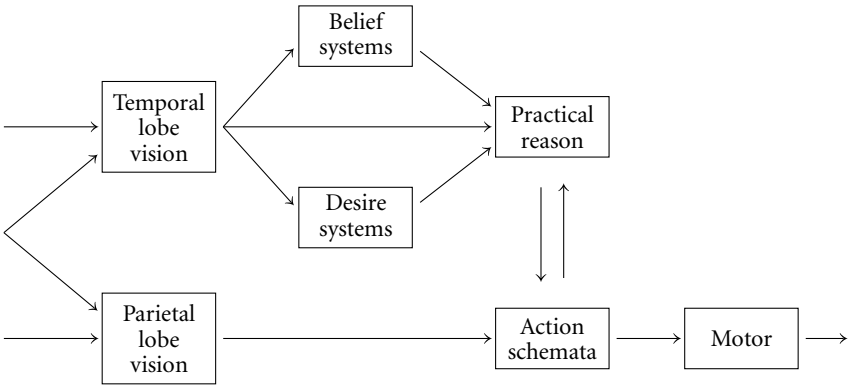

Figure 2. The dual visual systems hypothesis

it contains a horizontal or vertical grating, or whether it contains an 'X' or an 'O'), they prove remarkably accurate. Subjects can also reach out and grasp objects in their 'blind' field with something like 80% or more of normal accuracy, and can catch a ball thrown from their 'blind' side, all without conscious awareness. (See Weiskrantz 1997, for details and discussion.)

More recently, a even more powerful case for the existence of non-conscious visual experience has been generated by the *two visual systems theory* proposed and defended by Milner and Goodale (1995). (See Figure 2.) They review a wide variety of kinds of neurological and neuro-psychological evidence for the substantial independence of two distinct visual systems, instantiated in the temporal and parietal lobes respectively. They conclude that the parietal lobes provide a set of specialized semi-independent modules for the on-line visual control of action; whereas the temporal lobes are primarily concerned with more off-line functions such as visual learning, object recognition, and action-planning in relation to the perceived environment. And only the experiences generated by the temporal-lobe system are phenomenally conscious, on their account.[6]

To get the flavor of Milner and Goodale's hypothesis, consider just one strand from the wealth of evidence they provide. (For more extensive philosophical discussion, see Carruthers 2000; Clark 2002.) This is a neurological syndrome called *visual form agnosia*, which results from damage localized to both temporal lobes, leaving primary visual cortex and the parietal lobes intact. (Visual form agnosia is normally caused by carbon monoxide poisoning, for reasons that are little understood.) Such patients cannot recognize objects

or shapes, and may be capable of little conscious visual experience; but their sensorimotor abilities remain largely intact.

One particular patient (D.F.) has now been examined in considerable detail. While D.F. is severely agnosic, she is not completely lacking in conscious visual experience. Her capacities to perceive colors and textures are almost completely preserved. (Why just these sub-modules in her temporal cortex should have been spared isn't known.) As a result, she can sometimes guess the identity of a presented object – recognizing a banana, say, from its yellow color and the distinctive texture of its surface. But she is unable to perceive the shape of the banana (whether straight or curved); nor its orientation (upright or horizontal; pointing towards her or across). Yet many of her sensorimotor abilities are close to normal – she would be able to reach out and grasp the banana, orienting her hand and wrist appropriately for its position and orientation, and using a normal and appropriate finger grip.

Under experimental conditions it turns out that although D.F. is at chance in identifying the orientation of a broad line or letter-box, she is almost normal when posting a letter through a similarly-shaped slot oriented at random angles. In the same way, although she is at chance when trying to discriminate between rectangular blocks of very different sizes, her reaching and grasping behaviors when asked to pick up such a block are virtually indistinguishable from those of normal controls. It is very hard to make sense of this data without supposing that the sensorimotor perceptual system is functionally and anatomically distinct from the object-recognition / conscious system.

There is a powerful case, then, for thinking that there are non-conscious as well as conscious visual percepts. While the perceptions that ground your thoughts when you plan in relation to the perceived environment ('I'll pick up *that* one') may be conscious, and while you will continue to enjoy conscious perceptions of what you are doing while you act, the perceptual states that actually guide the details of your movements when you reach out and grab the object will *not* be conscious ones, if Milner and Goodale are correct.

But what implications does this have for *phenomenal* consciousness, as opposed to *access* consciousness (Block 1995)? Must these non-conscious percepts also be lacking in *phenomenal* properties? Most people think so. While it may be possible to get oneself to believe that the perceptions of the absent-minded car driver can remain phenomenally conscious (perhaps lying outside of the focus of attention, or being instantly forgotten), it is very hard to believe that either blindsight percepts or D.F.'s sensorimotor perceptual states might be phenomenally conscious ones. For these perceptions are ones to which the subjects of those states are *blind*, and of which they *cannot* be aware. And the ques-

tion, then, is: what makes the relevant difference? What is it about a conscious perception that renders it *phenomenal*, that a blindsight perceptual state would correspondingly lack? Higher-order perception (HOP) theorists are united in thinking that the relevant difference consists in the presence of a higher-order perceptual content in the first case that is absent in the second, in virtue of the presence of which a phenomenally conscious state is a state *of which the subject is perceptually aware*.

First-order (FOR) theories, by contrast, face considerable difficulties on this point. Unless the proponents of such theories choose to respond by denying the data, or by insisting that even blindsight and sensorimotor percepts are actually phenomenally conscious ones, then there is really only one viable option remaining. This is to appeal to the functional differences between percepts of the different kinds in explaining why one set is phenomenally conscious while the others aren't. For notice that the percepts constructed by the temporal-lobe system are available to conceptual thought and planning, but not to guide detailed movement on-line; whereas the reverse is true of the percepts produced by the parietal system. It is therefore open to a FOR theorist to say that it is *availability to conceptual thought* that constitutes an otherwise non-conscious perceptual state as phenomenally conscious (Kirk 1994; Tye 1995).

If what were being proposed were a brute identity claim, then such a position might be acceptable (or as acceptable as such claims ever are, if what we really seek is an *explanation*).[7] But proponents of FOR theories are supposed to be in the business of reductively *explaining* phenomenal consciousness. And it is left entirely obscure why the presence of conceptual thought and/or planning should make such a difference. Why should a perceptual state with the content *analog-green*, for example, remain unconscious if it is available just to guide movement, but become phenomenally conscious if used to inform conceptualized thoughts (such as, 'That one is green' or 'I will pick up the green one')? Granted, there is a big difference between thinking and acting. But what reason is there for believing that this difference can explain the difference between phenomenality and its lack?

Actualist higher-order thought (HOT) theory faces essentially the same difficulty. In explaining why sensorimotor percepts aren't phenomenally conscious, a HOT theorist can point out that while such percepts guide movement, they aren't available to higher-order thought and judgment. In contrast, the percepts produced by the temporal-lobe visual system are available to conceptual thought and reasoning in general, and to higher-order thought in particular. And the claim can then be made that those perceptual states produced

by the temporal-lobe system are phenomenally conscious that are actually the target of a higher-order thought about themselves.

But why should the presence of a higher-order belief about the existence of a first-order perceptual state render that state phenomenally conscious? Why should higher-order *access* consciousness generate *phenomenal* consciousness? The first-order state remains the same, just as it was in the absence of the higher-order thought. (Or if it changes, this will be merely via a shifting of perceptual similarity-spaces, of the sort that is often caused by concept-application – as when I can make the aspect of a duck–rabbit figure alter by applying different concepts to it – not a change from the absence of subjective *feel* to its presence.) And the higher-order thought in question will characteristically not be a conscious one. It seems like actualist HOT theorists have no option but to advance a brute identity claim, saying that to be a phenomenally conscious state just *is* to be a perceptual state targeted by a HOT. But this is to give up on attempting a reductive *explanation* of the phenomena.[8]

5. Does HOP theory face the same objection?

I have argued then (in Section 3) that both first-order (FOR) theories and actualist HOT theories face essentially the same problem in explaining how we can have purely recognitional concepts of experience; whereas higher-order perception (HOP) theories are well placed to provide such an explanation. And I have now argued (in Section 4) that neither FOR theories nor actualist HOT theory can give an adequate account of the conscious / non-conscious distinction, explaining why some perceptual states are phenomenally conscious while some aren't. But how well do HOP theories perform in this latter respect? For ease of presentation, I shall now switch to framing the discussion in terms of the form of HOP theory that I actually endorse, namely dispositionalist HOT theory. On this account, phenomenal consciousness consists in the dual perceptual content (both first-order and higher-order) possessed by those perceptual states that are made available to HOT (given the truth of some or other form of consumer semantics).

Initially, the explanation of the difference between the states produced by the parietal and temporal-lobe visual systems is straightforward. The outputs of the sensorimotor system are first-order analog contents that are used merely to guide movement; and as such they aren't phenomenally conscious. The outputs of the temporal-lobe system, in contrast, are available to a variety of downstream consumer systems (such as action-planning), included in which is a

faculty for higher-order thought (HOT). And it is by virtue of their availability to this HOT faculty that the first-order analog states that are produced by the temporal-lobe system come to acquire, at the same time, higher-order analog contents (given the truth of consumer semantics). And it is by virtue of having such dual content that the perceptual states in question are phenomenally conscious.[9]

Here is how an objection might go, however (Byrne 2001). Even if we don't have any real examples, it surely *could* happen that dual-content perceptual states might occur without being accessible to their subjects (e.g. without being available to conscious thought and/or without being reportable in speech). For example, perhaps there could be a separate HOT faculty that monitors the outputs of the sensorimotor visual system for some reason, rendering those states, too, as dual-content ones. Then if I want to say that such states wouldn't really be phenomenally conscious ones, don't I have to appeal to functional-role considerations, just as FOR theory did above? Don't I have to say that phenomenally conscious states are dual-content perceptual states *that are reportable in speech*, or something of the sort? And then can't I, too, be charged with postulating a brute identity here, and giving up on reductive explanation?

An initial reply is that it is extremely unlikely that there should *actually* be such dual contents that aren't available to us. This is because higher-order thought doesn't come cheap. So far as we know, a capacity for it has evolved just once in the history of life on earth, somewhere in the great ape / hominid lineage (perhaps only with the appearance of *Homo* – see Povinelli 2000, for a skeptical look at the mentalizing abilities of chimps). The idea that there might be a capacity for HOT attached to the outputs of the sensorimotor system, or embedded someplace within our color-discrimination module or whatever, is unlikely in the extreme. So I don't think that there are any *real* examples of non-access-conscious dual-content analog states (in the way that there *are* lots of real examples of non-conscious first-order perceptual states).

I concede, however, that it is logically possible that there could be dual-content events that aren't (in a sense) conscious. But here it is important to emphasize the distinction between *phenomenal* consciousness and various forms of *access* consciousness. I bite the bullet, and commit myself to the view that it is logically possible for there to be phenomenally conscious events (analog perceptual states with dual content, hence perceptual states with a subjective dimension) that aren't access-conscious (that aren't available for reporting in speech or to figure in decision-making). And I think that intuitions to the contrary are easily explained away. I certainly don't see why one should *define* phenomenal consciousness in such a way as to entail access consciousness.

This is where the fact that there are no real examples of dual-content perceptual states that aren't also access-conscious becomes important. For within our experience and to the best of our belief these two properties are always co-instantiated. It might be natural for us, then, to assume that the two are somehow essentially connected with one another – especially since imagination, when conscious and reflectively guided, always deploys states that are access-conscious. It is hard for us to imagine a phenomenally conscious state that isn't access-conscious. But that may just be because any image that we reflectively form is *de facto* access-conscious, given the way in which our cognitive system is actually structured.

What matters is not what we can or can't imagine, but what we can or can't explain. And my contention is that dispositionalist HOT theory can reductively explain the distinctive features of phenomenal consciousness. In particular, by virtue of their dual analog content, perceptual states that are available to HOT will take on a subjective dimension. They will be both world-representing (or body-representing, in the case of pain and touch) and experience-representing at the same time. In such cases it isn't just the world that is presented in a certain way to us, but our own experience of that world will also be presented in a certain way to us. And by virtue of such higher-order presentings, we can form purely recognitional concepts targeted on those very experiential states.

This isn't the place to set out and explain in detail the way in which dispositionalist HOT theory / dual-content theory can provide a successful reductive explanation of the various distinctive features of phenomenal consciousness. (See Carruthers 2000.) But in addition to explaining how phenomenally conscious states possess a subjective dimension, this approach can also of course explain how such states possess properties that are available to introspective recognition, and how they can ground purely recognitional concepts, as we saw in Section 3 above. We can therefore explain why, to anyone employing such concepts, the so-called 'explanatory gap' will seem to be unbridgeable. For such a person will always be able to combine, without incoherence, any proposed theory (including dual-content theory itself) with the thought, 'But someone might satisfy the conditions of the theory without possessing *this* kind of state' (thereby deploying their purely-recognitional concept *this*). Our account can also explain, too (and in common with other representationalist approaches, it should be said), how phenomenally conscious properties have a 'fineness of grain' that gives them a richness well beyond our powers of description and categorization. And it can be shown how people will then be strongly inclined to think of phenomenally conscious states as possessing intrinsic – that is, non-relational and non-intentional – properties; that people will be inclined

to think of these properties as ineffable and private; and that we will be inclined to think that we have incorrigible, or at least privileged, knowledge of them.

Although there hasn't here been the space to develop these points in any detail, it should nevertheless be plain that it is the dual-content aspect of the theory, rather than wider features of functional role (availability to planning and to speech, for example), that does the work in these explanations. This seems to me adequate motivation for the claim that phenomenal consciousness is constituted by dual-content perceptual states, wherever they might occur. To the best of our knowledge such states are also always actually accessible to the reasoning processes and reporting systems of their subjects. But there is nothing in my account of phenomenal consciousness as such that logically requires it.

6. Conclusion

I have argued that amongst higher-order perception (HOP) theories of phenomenal consciousness, dispositionalist HOT theory / dual-content theory is preferable to inner sense theory. I have also argued that HOP theories are preferable to both first-order (FOR) theories and to actualist HOT theory. Neither of the latter can give an adequate account of purely recognitional concepts of experience, nor of the distinction between conscious and non-conscious perceptual states; whereas HOP theories are well placed in both these respects. In the end, then, a dual-content theorist is what everyone ought to be.[10]

Notes

1. Lycan (1996) describes first-order perceptual states as possessing *qualia*, irrespective of their targeting by higher-order perception; and the terminology of 'qualia' is normally reserved for states that are phenomenally conscious. But I think that what he has in mind is just that first-order perceptual states represent fine-grained colors, textures and so forth; and that those states only acquire a dimension of subjective *feel* (hence becoming phenomenally conscious) when they are higher-order perceived. At any rate this is what I shall assume in what follows. (Inner sense theory seems to me devoid of interest otherwise.)

2. For teleosemantics (see Millikan 1984, 1989; Papineau 1987, 1993). For functional or inferential role semantics (see Loar 1981; Block 1986; McGinn 1989; Peacocke 1992).

3. Another variant on this theme, is that according to inner sense theory it ought to be possible for me to undergo a higher-order perception with the analog / non-conceptual content *seems orange* while I am undergoing no relevant first-order perceptual state at all. (Just as,

in the case of hallucination, my first-order senses can sometimes produce a state with the analog / non-conceptual content *red*, while there is nothing colored in my environment at all.) In such circumstances I would be inclined to make the first-order spontaneous judgment that I see *nothing* colored, while at the same time saying that I have an experience that seems orange to me. This combination of judgments seems barely coherent. Note, too, that similar problems can arise for actualist HOT theory; see Levine 2000.

4. Proponents of the existence of such concepts are then committed, of course, to rejecting the (quite different) arguments put forward by Wittgenstein (1953) and Fodor (1998) against the very possibility of purely recognitional concepts. Fortunately, neither set of arguments is at all compelling, though I shan't attempt to demonstrate this here.

5. By this I *don't* mean that my higher-order judgments are *non-conscious*. For this isn't problematic. It is granted on all hands that the higher-order representations that render our first-order percepts conscious aren't themselves conscious ones, in general. Rather, I mean that for actualist HOT theory, higher-order judgments of experience aren't grounded in awareness of their objects; which debars them from counting as genuinely *recognitional*.

6. Note that this isn't the old and familiar distinction between *what* and *where* visual systems, but is rather a successor to it. For the temporal-lobe system is supposed to have access both to property information and to spatial information. Instead, it is a distinction between a combined *what-where* system located in the temporal lobes and a *how-to* or action-guiding system located in the parietal lobes. And note, too, that the two-visual-systems hypothesis has the resources to explain the blindsight data.

7. Block and Stalnaker (1999) argue that identities aren't the kinds of facts that *admit* of further explanation. Consider the identity of water and H_2O, for example. If someone asks, '*Why* is water H_2O?', it looks like we can only reply (vacuously), 'Because it is'. You can't *explain* the identity of water and H_2O. Rather, identity facts are *brute* ones (not further explicable). Now, it is true that identities can't be explained as such. But it is also true that, if the identity is to count as a successful reduction of the higher-level property involved, then it must be possible to deploy features of the property that figures on the reducing-side of the identity-claim in such a way as to explain the features distinctive of the property on the other side (the reduced property). Consider the identity of water and H_2O again. Don't we think that it must be possible to deploy facts about H_2O, as such, in order to explain the distinctive properties of water – why it is colorless and odorless; why it is liquid at room temperatures; why it boils at 100° Centigrade; and so forth? Likewise, then, with phenomenal consciousness. A postulated identity, here, can only be acceptable if we can deploy the properties involved in such as way as to explain some of the distinctive features of phenomenality.

8. For discussion of the demands placed on successful reductive explanation in general, and as applied to reductive explanations of phenomenal consciousness in particular, see Carruthers (2004b).

9. Am I here holding dispositionalist HOT theory to a standard less demanding than that just imposed upon actualist HOT and FOR theories? No, because the dual-content idea *can* reductively explain the various features of phenomenal consciousness, particularly the latter's *subjective aspect* and the way in which it can ground purely recognitional concepts of experience.

10. Thanks to Zoltan Dienes, Rocco Gennaro and Bill Lycan for comments on an earlier draft of this chapter.

References

Armstrong, D. (1968). *A Materialist Theory of the Mind*. London: Routledge.

Baars, B. (1997). *In the Theatre of Consciousness*. Oxford: Oxford University Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, *10*, 615–678.

Block, N. (1995). A confusion about the function of consciousness. *Behavioral and Brain Sciences*, *18*, 227–247.

Block, N. & Stalnaker, R. (1999). Conceptual analysis, dualism and the explanatory gap. *The Philosophical Review*, *108*, 1–46.

Botterill, G. & Carruthers, P. (1999). *The Philosophy of Psychology*. Cambridge: Cambridge University Press.

Byrne, A. (2001). Review of *Phenomenal Consciousness* by Peter Carruthers. *Mind*, *110*, 440–442.

Byrne, R. & Whiten, A. (Eds.). (1988). *Machiavellian Intelligence*. Oxford: Oxford University Press.

Byrne, R. & Whiten, A. (Eds.). (1998). *Machiavellian Intelligence II*. Cambridge: Cambridge University Press.

Carruthers, P. (1996). *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. (1999). Sympathy and subjectivity. *Australasian Journal of Philosophy*, *77*, 465–482.

Carruthers, P. (2000). *Phenomenal Consciousness: A naturalistic theory*. Cambridge: Cambridge University Press.

Carruthers, P. (2004a). Phenomenal concepts and higher-order experiences. *Philosophy and Phenomenological Research*, *68*.

Carruthers, P. (forthcoming). Why the question of animal consciousness might not matter very much.

Carruthers, P. (2004b). Reductive explanation and the 'explanatory gap'. *Canadian Journal of Philosophy, 34*.

Chalmers, D. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Clark, A. (2002). Visual experience and motor action: Are the bonds too tight? *Philosophical Review*, *110*, 495–520.

Crick, F. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2*, 263–275.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Fodor, J. (1998). There are no recognitional concepts, not even RED. In his *In Critical Condition*. Cambridge, MA: MIT Press.

Kirk, R. (1994). *Raw Feels*. Oxford: Oxford University Press.

Levine, J. (2000). *Purple Haze*. Cambridge, MA: MIT Press.

Loar, B. (1981). *Mind and Meaning*. Cambridge: Cambridge University Press.
Loar, B. (1990). Phenomenal states. In J. Tomberlin (Ed.), *Philosophical Perspectives*, 4. Northridge, CA: Ridgeview.
Loar, B. (1997). Phenomenal states. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press.
Locke, J. (1690). *An Essay Concerning Human Understanding*. Many editions now available.
Lycan, W. (1987). *Consciousness*. Cambridge, MA: MIT Press.
Lycan, W. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.
McGinn, C. (1989). *Mental Content.* Oxford: Blackwell.
McGinn, C. (1991). *The Problem of Consciousness.* Oxford: Blackwell.
Millikan, R. (1984). *Language, Thought, and Other Biological Categories.* Cambridge, MA: MIT Press.
Millikan, R. (1989). Biosemantics. *Journal of Philosophy*, *86*, 281–297.
Milner, D. & Goodale, M. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.
Papineau, D. (1987). *Reality and Representation.* Oxford: Blackwell.
Papineau, D. (1993). *Philosophical Naturalism.* Oxford: Blackwell.
Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.
Peacocke, C. (1992). *A Study of Concepts.* Cambridge, MA: MIT Press.
Povinelli, D. (2000). *Folk Physics for Apes*. Oxford: Oxford University Press.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329–359.
Rosenthal, D. (1993). Thinking that one thinks. In M. Davies & G. Humphreys (Eds.), *Consciousness*. Oxford: Blackwell.
Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press.
Sturgeon, S. (1994). The epistemic view of subjectivity. *Journal of Philosophy*, *91*, 221–235.
Sturgeon, S. (2000). *Matters of Mind*. London: Routledge.
Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
Tye, M. (2000). *Consciousness, Color and Content*. Cambridge, MA: MIT Press.
Weiskrantz, L. (1986). *Blindsight.* Oxford: Oxford University Press.
Weiskrantz, L. (2000). *Consciousness Lost and Found*. Oxford: Oxford University Press.
Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Chapter 7

A higher order syntactic thought (HOST) theory of consciousness

Edmund T. Rolls

1. Background

The background to the HOST theory of consciousness described here is a theory of emotion based on the neuroscience of emotion (Rolls 1999a, 2002e, 1990), as described in this section.

1.1 A theory of emotion

Rolls' theory of emotion holds that emotions can usefully be defined as states elicited by rewards and punishers which have particular functions (Rolls 1999a). A reward is anything for which an animal (which includes humans) will work. A punisher is anything that an animal will escape from or avoid. An example of an emotion might thus be happiness produced by being given a reward, such as a pleasant touch, praise, or winning a large sum of money. Another example of an emotion might be fear produced by the sound of a rapidly approaching bus, or the sight of an angry expression on someone's face. We will work to avoid such stimuli, which are punishing. Another example would be frustration, anger, or sadness produced by the omission of an expected reward such as a prize, or the termination of a reward such as the death of a loved one. Another example would be relief, produced by the omission or termination of a punishing stimulus such as the removal of a painful stimulus, or sailing out of danger. These examples indicate how emotions can be produced by the delivery, omission, or termination of rewarding or punishing stimuli, and go some way to indicate how different emotions could be produced and classified in terms of the rewards and punishments received, omitted, or terminated.

This approach has been greatly extended to account for many types of emotion (Rolls 1999a, 2000e, 1990).

It is worth raising the issue that philosophers usually categorize fear in the example as an emotion, but not pain. The distinction they make may be that primary (unlearned) reinforcers do not produce emotions, whereas secondary reinforcers (stimuli associated by stimulus-reinforcement learning with primary reinforcers) do. They describe the pain as a sensation. But neutral stimuli (such as a table) can produce sensations when touched. It accordingly seems to be much more useful to categorise stimuli according to whether they are reinforcing (in which case they produce emotions), or are not reinforcing (in which case they do not produce emotions). Clearly there is a difference between primary reinforcers and learned reinforcers; but this is most precisely caught by noting that this is the difference, and that it is whether a stimulus is reinforcing that determines whether it is related to emotion.

1.2 The functions of emotion

The functions of emotion also provide insight into the nature of emotion. These functions, described more fully elsewhere (Rolls 1999a, 2000e, 1990), can be summarized as follows:

1. The *elicitation of autonomic responses* (e.g., a change in heart rate) *and endocrine responses* (e.g., the release of adrenaline). These prepare the body for action.
2. *Flexibility of behavioral responses to reinforcing stimuli*. Emotional (and motivational) states allow a simple interface between sensory inputs and action systems. The essence of this idea is that goals for behavior are specified by reward and punishment evaluation. When an environmental stimulus has been decoded as a primary reward or punishment, or (after previous stimulus-reinforcer association learning) a secondary rewarding or punishing stimulus, then it becomes a goal for action. The animal can then perform any action (instrumental response) to obtain the reward, or to avoid the punisher. Thus there is flexibility of action, and this is in contrast with stimulus-response, or habit, learning in which a particular response to a particular stimulus is learned. The emotional route to action is flexible not only because any action can be performed to obtain the reward or avoid the punishment, but also because the animal can learn in as little as one trial that a reward or punishment is associated with a particular stimulus, in what is termed "stimulus-reinforcer association learning".

To summarize and formalize, two processes are involved in the actions being described. The first is stimulus-reinforcer association learning, and the second is instrumental learning of an operant response made to approach and obtain the reward or to avoid or escape from the punisher. Emotion is an integral part of this, for it is the state elicited in the first stage, by stimuli which are decoded as rewards or punishers, and this state has the property that it is motivating. The motivation is to obtain the reward or avoid the punisher, and animals must be built to obtain certain rewards and avoid certain punishers. Indeed, primary or unlearned rewards and punishers are specified by genes which effectively specify the goals for action. This, Rolls proposes (Rolls 1999a) is the solution which natural selection has found for how genes can influence behavior to promote their fitness (as measured by reproductive success), and for how the brain could interface sensory systems to action systems.

Selecting between available rewards with their associated costs, and avoiding punishers with their associated costs, is a process which can take place both implicitly (unconsciously), and explicitly using a language system to enable long-term plans to be made (Rolls 1999a). These many different brain systems, some involving implicit evaluation of rewards, and others explicit, verbal, conscious, evaluation of rewards and planned long-term goals, must all enter into the selector of behavior (see Figure 1). This selector is poorly understood, but it might include a process of competition between all the competing calls on output, and might involve the basal ganglia in the brain (see Figure 1 and Rolls 1999a).

3. Emotion is *motivating*, as just described. For example, fear learned by stimulus-reinforcement association provides the motivation for actions performed to avoid noxious stimuli.
4. *Communication*. Monkeys for example may communicate their emotional state to others, by making an open-mouth threat to indicate the extent to which they are willing to compete for resources, and this may influence the behavior of other animals. This aspect of emotion was emphasized by Darwin (Darwin 1872) and has been studied more recently by Ekman (1982, 1993). As shown elsewhere (Rolls 2000a, 2000c), there are neural systems in the amygdala and overlying temporal cortical visual areas which are specialized for the face-related aspects of this processing.
5. *Social bonding*. Examples of this are the emotions associated with the attachment of the parents to their young, and the attachment of the young to their parents.

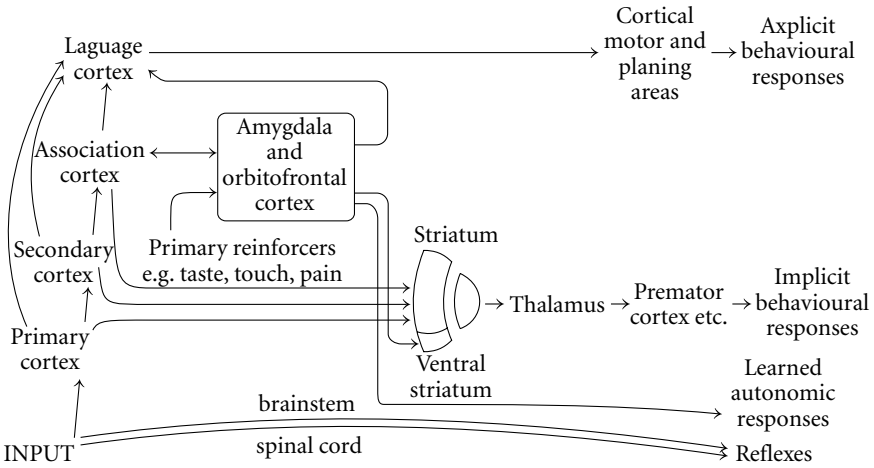

Figure 1. Dual routes to the initiation of action in response to rewarding and punishing stimuli. The inputs from different sensory systems to brain structures such as the orbitofrontal cortex and amygdala allow these brain structures to evaluate the reward- or punishment-related value of incoming stimuli, or of remembered stimuli. The different sensory inputs enable evaluations within the orbitofrontal cortex and amygdala based mainly on the primary (unlearned) reinforcement value for taste, touch and olfactory stimuli, and on the secondary (learned) reinforcement value for visual and auditory stimuli. In the case of vision, the 'association cortex' which outputs representations of objects to the amygdala and orbitofrontal cortex is the inferior temporal visual cortex. One route for the outputs from these evaluative brain structures is via projections directly to structures such as the basal ganglia (including the striatum and ventral striatum) to enable implicit, direct behavioural responses based on the reward or punishment-related evaluation of the stimuli to be made. The second route is via the language systems of the brain, which allow explicit (verbalizable) decisions involving multi-step syntactic planning to be implemented. (After Rolls 1999a: Figure 9.4).

6. The current mood state can affect the *cognitive evaluation of events or memories* (see Oatley & Jenkins 1996). This may facilitate continuity in the interpretation of the reinforcing value of events in the environment. A hypothesis that backprojections from parts of the brain involved in emotion such as the orbitofrontal cortex and amygdala implement this is described in *The Brain and Emotion* (Rolls 1999a).
7. Emotion may facilitate the *storage of memories*. One way this occurs is that episodic memory (i.e., one's memory of particular episodes) is facilitated by emotional states. This may be advantageous in that storing many details of the prevailing situation when a strong reinforcer is delivered may be use-

ful in generating appropriate behavior in situations with some similarities in the future. This function may be implemented by the relatively nonspecific projecting systems to the cerebral cortex and hippocampus, including the cholinergic pathways in the basal forebrain and medial septum, and the ascending noradrenergic pathways (Rolls 1999a; Rolls & Treves 1998). A second way in which emotion may affect the storage of memories is that the current emotional state may be stored with episodic memories, providing a mechanism for the current emotional state to affect which memories are recalled. A third way that emotion may affect the storage of memories is by guiding the cerebral cortex in the representations of the world which are set up. For example, in the visual system it may be useful for perceptual representations or analyzers to be built which are different from each other if they are associated with different reinforcers, and for these to be less likely to be built if they have no association with reinforcement. Ways in which backprojections from parts of the brain important in emotion (such as the amygdala) to parts of the cerebral cortex could perform this function are discussed by Rolls and Treves (1998).

8. Another function of emotion is that by enduring for minutes or longer after a reinforcing stimulus has occurred, it may help to produce *persistent and continuing motivation and direction of behavior*, to help achieve a goal or goals.
9. Emotion may trigger the *recall of memories* stored in neocortical representations. Amygdala backprojections to the cortex could perform this for emotion in a way analogous to that in which the hippocampus could implement the retrieval in the neocortex of recent (episodic) memories (Rolls & Stinger 2001; Rolls & Treves 1998).

1.3 To what extent is consciousness involved in the different types of processing initiated by emotional states?

It might be possible to build a computer which would perform the functions of emotions described above and in more detail by Rolls (Rolls 1999a, 2000e), and yet we might not want to ascribe emotional *feelings* to the computer. We might even build the computer with some of the main processing stages present in the brain, and implemented using neural networks which simulate the operation of the real neural networks in the brain (Rolls & Deco 2002; Rolls & Treves 1998), yet we might not still wish to ascribe emotional feelings to this computer. In a sense, the functions of reward and punishment in emotional behaviour are described by the above types of process and their underlying

brain mechanisms in structures such as the amygdala and orbitofrontal cortex as described by Rolls (Rolls 1999a, 2000c, 2000d), but what about the subjective aspects of emotion, what about the pleasure? A similar point arises when we consider the parts of the taste, olfactory, and visual systems in which the reward value of the taste, smell and sight of food are represented. One such brain region is the orbitofrontal cortex (Rolls 1999a, 2002, 2000d, 1997b). Although the neuronal representation in the orbitofrontal cortex is clearly related to the reward value of food, is this where the pleasantness (the subjective hedonic aspect) of the taste, smell and sight of food is represented? Again, we could (in principle at least) build a computer with neural networks to simulate each of the processing stages for the taste, smell and sight of food which are described by Rolls (Rolls 1999a; Rolls & Deco 2002; and more formally in terms of neural networks Rolls & Treves 1998), and yet would probably not wish to ascribe feelings of pleasantness to the system we have simulated on the computer. The point I am making here is that much of the processing related to the control of emotional and motivational behavior could take place without any need to invoke the possibility to be doing the type of processing that leads to reward decoding, and the initiation of behavioral responses. Part of the evidence for this hypothesis is that patients with orbitofrontal cortex damage may not be able to implement reward-related associative learning and reversal, yet may be able to comment verbally and consciously on the behavioral choices that they should be making (Hornak 2003a, 2003b; Rolls et al. 1994). This dissociation between different systems involved in the implementation of behavior is described further in Sections 1.4 and 3 on dual routes to action. (I do believe that provided that the right type of processing is implemented in a computer, it would be conscious. In this sense I am a functionalist. In this Chapter I develop a hypothesis on what would have to be implemented in a computer for it to be conscious.)

What is it about neural processing that makes it feel like something when some types of information processing are taking place. It is clearly not a general property of processing in neural networks, for there is much processing, for example that concerned with the control of our blood pressure and heart rate, of which we are not aware. Is it then that awareness arises when a certain type of information processing is being performed? If so, what type of information processing? And how do emotional feelings, and sensory events, come to feel like anything? These feels are called qualia. These are great mysteries that have puzzled philosophers for centuries. They are at the heart of the problem of consciousness, for why it should feel like something at all is the great mystery. Other aspects of consciousness, such as the fact that often when we "pay atten-

tion" to events in the world, we can process those events in some better way, that is process or access as opposed to phenomenal aspects of consciousness, may be easier to analyse (Allport 1988; Block 1995; Chalmers 1996). The puzzle of qualia, that is of the phenomenal aspect of consciousness, seems to be rather different from normal investigations in science, in that there is no agreement on criteria by which to assess whether we have made progress. So, although the aim of this Chapter is to address the issue of consciousness, especially of qualia, in relation to emotional feelings and actions, what is written cannot be regarded as being establishable by the normal methods of scientific enquiry. Accordingly, I emphasize that the view on consciousness that I describe is only preliminary, and theories of consciousness are likely to develop considerably. Partly for these reasons, this theory of consciousness, at least, should not be taken to have practical implications.

1.4 Dual routes to action, and consciousness

According to the present formulation, there are two types of route to action performed in relation to reward or punishment in humans (Rolls 1999a, 2003). Examples of such actions include emotional and motivational behaviour.

The *first route* is via the brain systems that have been present in non-human primates such as monkeys, and to some extent in other mammals, for millions of years. These systems include the amygdala and, particularly well-developed in primates, the orbitofrontal cortex. These systems control behaviour in relation to previous associations of stimuli with reinforcement. The computation which controls the action thus involves assessment of the reinforcement-related value of a stimulus. This assessment may be based on a number of different factors. One is the previous reinforcement history, which involves stimulus-reinforcement association learning using the amygdala, and its rapid updating especially in primates using the orbitofrontal cortex. This stimulus-reinforcement association learning may involve quite specific information about a stimulus, for example of the energy associated with each type of food, by the process of conditioned appetite and satiety (Booth 1985). A second is the current motivational state, for example whether hunger is present, whether other needs are satisfied, etc. A third factor which affects the computed reward value of the stimulus is whether that reward has been received recently, by the processes of incentive motivation and sensory-specific satiety (Rolls 1999a). A fourth factor is the computed absolute value of the reward or punishment expected or being obtained from a stimulus, e.g., the sweetness of the stimulus (set by evolution so that sweet stimuli will tend to be rewarding,

because they are generally associated with energy sources), or the pleasantness of touch (set by evolution to be pleasant according to the extent to which it brings animals of the opposite sex together, and depending on the investment in time that the partner is willing to put into making the touch pleasurable, a sign which indicates the commitment and value for the partner of the relationship). After the reward value of the stimulus has been assessed in these ways, behaviour is then initiated based on approach towards or withdrawal from the stimulus. A critical aspect of the behaviour produced by this type of system is that it is aimed directly towards obtaining a sensed or expected reward, by virtue of connections to brain systems such as the basal ganglia which are concerned with the initiation of actions (see Figure 1). The expectation may of course involve behaviour to obtain stimuli associated with reward, which might even be present in a chain.

Now part of the way in which the behaviour is controlled with this first route is according to the reward value of the outcome. At the same time, the animal may only work for the reward if the cost is not too high. Indeed, in the field of behavioural ecology, animals are often thought of as performing optimally on some cost-benefit curve (see, e.g., Krebs & Kacelnik 1991). This does not at all mean that the animal thinks about the rewards, and performs a cost-benefit analysis using a lot of thoughts about the costs, other rewards available and their costs, etc. Instead, it should be taken to mean that in evolution, the system has evolved in such a way that the way in which the reward varies with the different energy densities or amounts of food and the delay before it is received, can be used as part of the input to a mechanism which has also been built to track the costs of obtaining the food (e.g., energy loss in obtaining it, risk of predation, etc.), and to then select given many such types of reward and the associated cost, the current behaviour that provides the most "net reward". Part of the value of having the computation expressed in this reward-minus-cost form is that there is then a suitable "currency", or net reward value, to enable the animal to select the behaviour with currently the most net reward gain (or minimal aversive outcome).

The *second route* in humans involves a computation with many "if. . .then" conditional statements, to implement a plan to obtain a reward. In this case, the reward may actually be *deferred* as part of the plan, which might involve working first to obtain one reward, and only then to work for a second more highly valued reward, if this was thought to be overall an optimal strategy in terms of resource usage (e.g., time). In this case, syntax is required, because the many symbols (e.g., names of people) that are part of the plan must be correctly linked or bound in order to implement for example the conditional

statements involved in each step of the plan. Such linking might be of the form: "if A does this, then B is likely to do this, and this will cause C to do this . . .". The requirement of syntax for this type of planning implies that an output to language systems in the brain is required for this type of planning (see Figure 1). Thus the explicit language system in humans may allow working for deferred rewards by enabling use of a one-off, individual, plan appropriate for each situation. I emphasise that this type of processing would involve syntax, in that each step of the plan might have its own symbols that would need to be kept apart from those utilized in other steps of the plan, and further that each step of the plan might itself require syntax, of the type required to implement for example "if . . . then" conditionals. The need for syntax can be illustrated also by considering a neural network in which different populations of neurons are active, each representing different symbols. If each symbol is represented by a set of neuronal firings, how does the system implement the relations between the symbols, e.g. that symbol A acts on symbol B and not vice versa. This need for syntax in neural networks has been recognised, and solutions to this binding problem such as stimulus-dependent synchronisation of the firing of different neuronal populations that need to be related to each other have been proposed (Singer 1999), but do not seem to be adequate for the job required (Rolls & Deco 2002: Section 13.7). Another building block for such planning operations in the brain may be the type of short term memory in which the prefrontal cortex is involved. This short term memory may be for example in non-human primates of where in space a response has just been made. A development of this type of short term response memory system in humans to enable multiple short term memories to be held in place correctly, preferably with the temporal order of the different items in the short term memory coded correctly, may be another building block for the multiple step "if . . . then" type of computation in order to form a multiple step plan. Such short term memories are implemented in the (dorsolateral and inferior convexity) prefrontal cortex of non-human primates and humans (see Goldman-Rakic 1996; Petrides 1996), and may be part of the reason why prefrontal cortex damage impairs planning (Rolls & Deco 2002; Shallice & Burgess 1996).

Of these two routes (see Figure 1), it is the second which I suggest is related to consciousness (see Rolls 1999a). The hypothesis is that consciousness is the state which arises by virtue of having the ability to think about one's own thoughts, which has the adaptive value I argue of enabling one to correct long multi-step syntactic plans and thus solving a credit assignment problem, as described below. This second system is thus the one in which explicit, declarative, processing occurs. Processing in this system is frequently associated with rea-

son and rationality, in that many of the consequences of possible actions can be taken into account. The actual computation of how rewarding a particular stimulus or situation is or will be probably still depends on activity in the orbitofrontal cortex and amygdala, as the reward value of stimuli is computed and represented in these regions, and in that it is found that verbalised expressions of the reward (or punishment) value of stimuli are dampened by damage to these systems. (For example, damage to the orbitofrontal cortex renders painful input still identifiable as pain, but without the strong affective, "unpleasant", reaction to it.) This language system which enables long-term planning may be contrasted with the first system in which behaviour is directed at obtaining the stimulus (including the remembered stimulus) which is currently most rewarding, as computed by brain structures that include the orbitofrontal cortex and amygdala. There are outputs from this system, perhaps those directed at the basal ganglia, which do not pass through the language system, and behaviour produced in this way is described as implicit, and verbal declarations cannot be made directly about the reasons for the choice made. When verbal declarations are made about decisions made in this first system, those verbal declarations may be confabulations, reasonable explanations or fabrications, of reasons why the choice was made. These reasonable explanations would be generated to be consistent with the sense of continuity and self that is a characteristic of reasoning in the language system. These points are developed next.

2. A theory of consciousness

A starting point is that many actions can be performed relatively automatically, without apparent conscious intervention. An example sometimes given is driving a car. Such actions could involve control of behaviour by brain systems which are old in evolutionary terms such as the basal ganglia. It is of interest that the basal ganglia (and cerebellum) do not have backprojection systems to most of the parts of the cerebral cortex from which they receive inputs (Rolls 1994; Rolls & Johnstone 1992; Rolls et al. 1998). In contrast, parts of the brain such as the hippocampus and amygdala, involved in functions such as episodic memory and emotion respectively, about which we can make (verbal) declarations (hence declarative memory, Squire 1992) do have major backprojection systems to the high parts of the cerebral cortex from which they receive forward projections (Rolls & Deco 2002; Rolls & Treves 1998; Treves & Rolls 1994; Rolls 2000b, 1996). It may be that evolutionarily newer parts of the brain, such as the

language areas and parts of the prefrontal cortex, are involved in an alternative type of control of behaviour, in which actions can be planned with the use of a (language) system which allows relatively arbitrary (syntactic) manipulation of semantic entities (symbols).

The general view that there are many routes to behavioural output is supported by the evidence that there are many input systems to the basal ganglia (from almost all areas of the cerebral cortex), and that neuronal activity in each part of the striatum reflects the activity in the overlying cortical area (Rolls 1994; Rolls & Johnstone 1992; Rolls & Treves 1998). The evidence is consistent with the possibility that different cortical areas, each specialised for a different type of computation, have their outputs directed to the basal ganglia, which then select the strongest input, and map this into action (via outputs directed for example to the premotor cortex) (Rolls & Johnstone 1992; Rolls & Treves 1998). Within this scheme, the language areas would offer one of many routes to action, but a route particularly suited to planning actions, because of the syntactic manipulation of semantic entities which may make long-term planning possible. A schematic diagram of this suggestion is provided in Figure 1.

Some of the evidence that supports the hypothesis of multiple routes to action, only some of which utilize language, is the evidence that split-brain patients may not be aware of actions being performed by the "non-dominant" hemisphere (Gazzaniga 1988, 1995; Gazzaniga & LeDoux 1978). Another important line of evidence for multiple, including non-verbal, routes to action, is that patients with focal brain damage, for example to the orbitofrontal cortex, may perform actions, yet comment verbally that they should not be performing those actions (Hornak 2003b, 1999b; Rolls et al. 1994). The actions which appear to be performed implicitly, with surprise expressed later by the explicit system, include making behavioral responses to a no-longer rewarded visual stimulus in a visual discrimination reversal (Hornak 2003b; Rolls 1994). In both these types of patient, confabulation may occur, in that a verbal account of why the action was performed may be given, and this may not be related at all to the environmental event which actually triggered the action (Gazzaniga 1988, 1995; Gazzaniga & LeDoux 1978; Rolls 1994). It is possible that sometimes in normal humans when actions are initiated as a result of processing in a specialized brain region such as those involved in some types of rewarded behaviour, the language system may subsequently elaborate a coherent account of why that action was performed (i.e., confabulate). This would be consistent with a general view of brain evolution in which as areas of the cortex evolve, they are laid on top of existing circuitry connecting inputs to outputs, and

in which each level in this hierarchy of separate input-output pathways may control behaviour according to the specialised function it can perform (see schematic in Figure 1). (It is of interest that mathematicians may get a hunch that something is correct, yet not be able to verbalise why. They may then resort to formal, more serial and language-like, theorems to prove the case, and these seem to require conscious processing. This is a further indication of a close association between linguistic processing, and consciousness. The linguistic processing need not, as in reading, involve an inner articulatory loop.)

We may next examine some of the advantages and behavioural functions that language, present as the most recently added layer to the above system, would confer. One major advantage would be the ability to plan actions through many potential stages and to evaluate the consequences of those actions without having to perform the actions. For this, the ability to form propositional statements, and to perform syntactic operations on the semantic representations of states in the world, would be important. Also important in this system would be the ability to have second-order thoughts about the type of thought that I have just described (e.g., I think that he thinks that . . .), as this would allow much better modelling and prediction of others' behaviour, and therefore of planning, particularly planning when it involves others. (Second order thoughts are thoughts about thoughts. Higher order thoughts refer to second order, third order etc. thoughts about thoughts. . .) This capability for higher order thoughts would also enable reflection on past events, which would also be useful in planning, and in particular in correcting these multistep plans. In this sense, higher order thoughts I propose help to solve a credit assignment problem (see below). In contrast, non-linguistic behaviour would be driven by learned reinforcement associations, learned rules etc., but not by flexible planning for many steps ahead involving a model of the world including others' behaviour. (For an earlier view which is close to this part of the argument see Humphrey 1980.) (The examples of behaviour from non-humans that may reflect planning may reflect much more limited and inflexible planning. For example, the dance of the honey-bee to signal to other bees the location of food may be said to reflect planning, but the symbol manipulation is not arbitrary. There are likely to be interesting examples of non-human primate behaviour, perhaps in the great apes, that reflect the evolution of an arbitrary symbol-manipulation system that could be useful for flexible planning (cf. Cheney & Seyfarth 1990). It is important to state that the language ability referred to here is not necessarily human verbal language (though this would be an example). What it is suggested is important to planning is the syntactic manipulation of symbols, and it is this syntactic manipulation of symbols which is the sense

in which language is defined and used here. This functionality is termed by some philosophers mentalese. The functionality required is not as strong as that required for natural language, which implies a universal grammar.

It is next suggested that this arbitrary symbol-manipulation using important aspects of language processing and used for planning but not in initiating all types of behaviour is close to what consciousness is about. In particular, consciousness may *be* the state which arises in a system that can think about (or reflect on) its own (or other peoples') thoughts, that is in a system capable of second or higher order thoughts (cf. Dennett 1991, 1990, 1993; Rosenthal 1986). On this account, a mental state is non-introspectively (i.e., non-reflectively) conscious if one has a roughly simultaneous thought that one is in that mental state. Following from this, introspective consciousness (or reflexive consciousness, or self consciousness) is the attentive, deliberately focussed consciousness of one's mental states. It is noted that not all of the higher order thoughts need themselves be conscious (many mental states are not). However, according to the analysis, having a higher-order thought about a lower order thought is necessary for the lower order thought to be conscious. A slightly weaker position than Rosenthal's on this is that a conscious state corresponds to a first order thought that has the *capacity* to cause a second order thought or judgement about it (Carruthers 1996). [Another position that is close in some respects to that of Carruthers and the present position is that of Chalmers (1996), that awareness is something that has *direct availability for behavioral control*, which amounts effectively *for him* in humans to saying that consciousness is what we can report (verbally) about.] This analysis is consistent with the points made above that the brain systems that are required for consciousness and language are similar. In particular, a system which can have second or higher order thoughts about its own operation, including its planning and linguistic operation, must itself be a language processor, in that it must be able to bind correctly to the symbols and syntax in the first order system. According to this explanation, the feeling of anything is the state which is present when linguistic processing that involves second or higher order thoughts is being performed.

It might be objected that this captures some of the process aspects of consciousness, what it is good for in an information processing system, but does not capture the phenomenal aspect of consciousness. I agree that there is an element of "mystery" that is invoked at this step of the argument, when I say that it feels like something for a machine with higher order thoughts to be thinking about its own first or lower order thoughts. But the return point is the following: *if a human with second order thoughts is thinking about its own first order thoughts, surely it is very difficult for us to conceive that this would* NOT *feel like*

something? (Perhaps the higher order thoughts in thinking about the first order thoughts would need to have in doing this some sense of continuity or self, so that the first order thoughts would be related to the same system that had thought of something else a few minutes ago. But even this continuity aspect may not be a requirement for consciousness. Humans with anterograde amnesia cannot remember what they felt a few minutes ago; yet their current state does feel like something.)

It is suggested that part of the evolutionary adaptive significance of this type of higher order thought is that is enables correction of errors made in first order linguistic or in non-linguistic processing. Indeed, the ability to reflect on previous events is extremely important for learning from them, including setting up new long-term semantic structures. It was shown elsewhere (Rolls & Treves 1998) that the hippocampus may be a system for such "declarative" recall of recent memories. Its close relation to "conscious" processing in humans (Squire 1992, has classified it as a declarative memory system) may be simply that it enables the recall of recent memories, which can then be reflected upon in conscious, higher order, processing (Rolls 1996). Another part of the adaptive value of a higher order thought system may be that by thinking about its own thoughts in a given situation, it may be able to better understand the thoughts of another individual in a similar situation, and therefore predict that individual's behaviour better (cf. Barlow 1997, 1986; Humphrey 1980).

As a point of clarification, I note that according to this theory, a language processing system is not *sufficient* for consciousness. What defines a conscious system according to this analysis is the ability to have higher order thoughts, and a first order language processor (that might be perfectly competent at language) would not be conscious, in that it could not think about its own or others' thoughts. One can perfectly well conceive of a system which obeyed the rules of language (which is the aim of much connectionist modelling), and implemented a first-order linguistic system, that would not be conscious. [Possible examples of language processing that might be performed non-consciously include computer programs implementing aspects of language, or ritualized human conversations, e.g., about the weather. These might require syntax and correctly grounded semantics, and yet be performed non-consciously. A more complex example, illustrating that syntax could be used, might be "If A does X, then B will probably do Y, and then C would be able to do Z." A first order language system could process this statement. Moreover, the first order language system could apply the rule usefully in the world, provided that the symbols in the language system (A, B, X, Y etc.) are grounded (have meaning) in the world.] In line with the argument on the adaptive value of higher

order thoughts and thus consciousness given above, that they are useful for correcting lower order thoughts, I now suggest that correction using higher order thoughts of lower order thoughts would have adaptive value primarily if the lower order thoughts are sufficiently complex to benefit from correction in this way. The nature of the complexity is specific: that it should involve syntactic manipulation of symbols, probably with several steps in the chain, and that the chain of steps should be a one-off (or in American, "one-time", meaning used once) set of steps, as in a sentence or in a particular plan used just once, rather than a set of well learned rules. The first or lower order thoughts might involve a linked chain of "if" ... "then" statements that would be involved in planning, an example of which has been given above. It is partly because complex lower order thoughts such as these which involve syntax and language would benefit from correction by higher order thoughts, that I suggest that there is a close link between this reflective consciousness and language. The hypothesis is that by thinking about lower order thoughts, the higher order thoughts can discover what may be weak links in the chain of reasoning at the lower order level, and having detected the weak link, might alter the plan, to see if this gives better success. In our example above, if it transpired that C could not do Z, how might the plan have failed? Instead of having to go through endless random changes to the plan to see if by trial and error some combination does happen to produce results, what I am suggesting is that by thinking about the previous plan, one might for example using knowledge of the situation and the probabilities that operate in it, guess that the step where the plan failed was that B did not in fact do Y. So by thinking about the plan (the first or lower order thought), one might correct the original plan, in such a way that the weak link in that chain, that "B will probably do Y", is circumvented. To draw a parallel with neural networks: there is a "*credit assignment*" problem in such multi-step syntactic plans, in that if the whole plan fails, how does the system assign credit or blame to particular steps of the plan? The suggestion is that this is the function of higher order thoughts and is why systems with higher order thoughts evolved. The suggestion I then make is that if a system were doing this type of processing (thinking about its own thoughts), it would then be very plausible that it should feel like something to be doing this. I even suggest to the reader that it is not plausible to suggest that it would not feel like anything to a system if it were doing this.

Two other points in the argument should be emphasised for clarity. One is that the system that is having syntactic thoughts about its own syntactic thoughts would have to have its symbols grounded in the real world for it to feel like something to be having higher order thoughts. The intention of this

clarification is to exclude systems such as a computer running a program when there is in addition some sort of control or even overseeing program checking the operation of the first program. We would want to say that in such a situation it would feel like something to be running the higher level control program only if the first order program was symbolically performing operations on the world and receiving input about the results of those operations, and if the higher order system understood what the first order system was trying to do in the world. The issue of symbol grounding is considered further by Rolls (Rolls 1999a: Section 10.4, 2000e). The symbols (or symbolic representations) are symbols in the sense that they can take part in syntactic processing. The symbolic representations are grounded in the world in that they refer to events in the world. The symbolic representations must have a great deal of information about what is referred to in the world, including the quality and intensity of sensory events, emotional states, etc. The need for this is that the reasoning in the symbolic system must be about stimuli, events, and states, and remembered stimuli, events and states, and for the reasoning to be correct, all the information that can affect the reasoning must be represented in the symbolic system, including for example just how light or strong the touch was, etc. (This suggestion may be close to the view that thoughts may be grounded by the way they function in "belief-desire psychology" which considers the functions of intentional states and attitudinal states such as beliefs, desires etc. as discussed by Fodor (1994, 1987, 1990). The notion of what constitutes a thought is itself a major issue. Animals can be said to have "expectations"; but if they are based on functionality that implements Stimulus-Response habits or stimulus-reinforcement associations, they would not I believe constitute thoughts. For the purposes of this chapter, "thought" can be read as "human thought", and would normally involve symbols and syntactic operations. The issue of the extent to which animals have thoughts which operate in this way remains to be fully examined and assessed, and is in principle a matter that can be resolved empirically. I am thus open-minded about the operation in animals of the type of processing described in this chapter as higher order syntactic thought.) Indeed, it is pointed out in *The Brain and Emotion* (Rolls 1999a: 252–253) that it is no accident that the shape of the multidimensional phenomenal (sensory etc.) space does map so clearly onto the space defined by neuronal activity in sensory systems, for if this were not the case, reasoning about the state of affairs in the world would not map onto the world, and would not be useful. Good examples of this close correspondence are found in the taste system, in which subjective space (Schiffman & Erikson 1971) maps simply onto the multidimensional space represented by neuronal firing in primate cortical taste areas.

In particular, if a three-dimensional space reflecting the distances between the representations of different tastes provided by macaque neurons in the cortical taste areas is constructed, then the distances between the subjective ratings by humans of different tastes is very similar (Plata-Salaman et al. 1996; Smith-Swintosky et al. 1991; Yaxley et al. 1990). Similarly, the changes in human subjective ratings of the pleasantness of the taste, smell and sight of food parallel very closely the responses of neurons in the macaque orbitofrontal cortex (see *The Brain and Emotion*, Chapter 2). The representations in the first order linguistic processor that the HOSTs process include beliefs (for example "Food is available", or at least representations of this), and the HOST system would then have available to it the concept of a thought (so that it could represent "I believe [or there is a belief] that food is available"). However, as summarized by Rolls (Rolls 2000e), representations of sensory processes and emotional states must be processed by the first order linguistic system, and HOSTs may be about these representations of sensory processes and emotional states capable of taking part in the syntactic operations of the first order linguistic processor. Such sensory and emotional information may reach the first order linguistic system from many parts of the brain, including those such as the orbitofrontal cortex and amygdala implicated in emotional states (see *The Brain and Emotion*, Figure 9.3 and p. 253). When the sensory information is about the identity of the taste, the inputs to the first order linguistic system must come from the primary taste cortex, in that the identity of taste, independent of its pleasantness (in that the representation is independent of hunger) must come from the primary taste cortex. In contrast, when the information that reaches the first order linguistic system is about the pleasantness of taste, it must come from the secondary taste cortex, in that there the representation of taste depends on hunger.

The second clarification is that the plan would have to be a unique string of steps, in much the same way as a sentence can be a unique and one-off (or one-time) string of words. The point here is that it is helpful to be able to think about particular one-off plans, and to correct them; and that this type of operation is very different from the slow learning of fixed rules by trial and error, or the application of fixed rules by a supervisory part of a computer program.

Sensory qualia

This analysis does not yet give an account for sensory qualia ("raw sensory feels", for example why "red" feels red), for emotional qualia (e.g., why a rewarding touch produces an emotional feeling of pleasure), or for motivational

qualia (e.g., why food deprivation makes us *feel* hungry). The view I suggest on such qualia is as follows. Information processing in and from our sensory systems (e.g., the sight of the colour red) may be relevant to planning actions using language and the conscious processing thereby implied. Given that these inputs must be represented in the system that plans, we may ask whether it is more likely that we would be conscious of them or that we would not. I suggest that it would be a very special-purpose system that would allow such sensory inputs, and emotional and motivational states, to be part of (linguistically based) planning, and yet remain unconscious. It seems to be much more parsimonious to hold that we would be conscious of such sensory, emotional and motivational qualia because they would be being used (or are available to be used) in this type of (linguistically based) higher order thought processing, and this is what I propose.

The explanation for emotional and motivational subjective feelings or qualia that this discussion has led towards is thus that they should be felt as conscious because they enter into a specialised linguistic symbol-manipulation system that is part of a higher order thought system that is capable of reflecting on and correcting its lower order thoughts involved for example in the flexible planning of actions. It would require a very special machine to enable this higher-order linguistically-based thought processing, which is conscious by its nature, to occur without the sensory, emotional and motivational states (which must be taken into account by the higher order thought system) becoming felt qualia. The qualia are thus accounted for by the evolution of the linguistic system that can reflect on and correct its own lower order processes, and thus has adaptive value. To expand on this, qualia of low-level sensory details may not themselves involve higher order thoughts (thoughts about thoughts), and may involve thoughts about sensory processes, but nevertheless become conscious because they enter the processing system that implements HOSTs. The HOST processing system may sometimes have to reason (perform multistep planning and evaluation) about low-level sensory features of objects, e.g. to establish whether an antique has the correct coloring and surface features for it to be genuine.

This account implies that it may be especially animals with a higher order belief and thought system and with linguistic symbol manipulation that have qualia. It may be that much non-human animal behaviour, provided that it does not require flexible linguistic planning and correction by reflection, could take place according to reinforcement-guidance (using e.g., stimulus-reinforcement association learning in the amygdala and orbitofrontal cortex, Rolls 1990), and rule-following (implemented e.g., using habit or stimulus-

response learning in the basal ganglia, Rolls 1994; Rolls & Johnstone 1992). Such behaviours might appear very similar to human behaviour performed in similar circumstances, but would not imply qualia. It would be primarily by virtue of a system for reflecting on flexible, linguistic, planning behaviour that humans (and animals close to humans, with demonstrable syntactic manipulation of symbols (termed mentalese), and the ability to think about these linguistic processes) would be different from other animals, and would have evolved qualia. (The hypothesis described here implies that consciousness in animals could be related to the extent to which mentalese is possible in each given type of animal, and does not require natural language.)

In order for processing in a part of our brain to be able to reach consciousness, appropriate pathways must be present. Certain constraints arise here. For example, in the sensory pathways, the nature of the representation may change as it passes through a hierarchy of processing levels, and in order to be conscious of the information in the form in which it is represented in early processing stages, the early processing stages must have access to the part of the brain necessary for consciousness. An example is provided by processing in the taste system. In the primate primary taste cortex, neurons respond to taste independently of hunger, yet in the secondary taste cortex, food-related taste neurons (e.g., responding to sweet taste) only respond to food if hunger is present, and gradually stop responding to that taste during feeding to satiety (see Rolls 1997b). Now the quality of the tastant (sweet, salt etc.) and its intensity are not affected by hunger, but the pleasantness of its taste is decreased to zero (neutral) (or even becomes unpleasant) after we have eaten it to satiety. The implication of this is that for quality and intensity information about taste, we must be conscious of what is represented in the primary taste cortex (or perhaps in another area connected to it which bypasses the secondary taste cortex), and not of what is represented in the secondary taste cortex. In contrast, for the pleasantness of a taste, consciousness of this could not reflect what is represented in the primary taste cortex, but instead what is represented in the secondary taste cortex (or in an area beyond it). The same argument arises for reward in general, and therefore for emotion, which in primates is not represented early on in processing in the sensory pathways (nor in or before the inferior temporal cortex for vision), but in the areas to which these object analysis systems project, such as the orbitofrontal cortex, where the reward value of visual stimuli is reflected in the responses of neurons to visual stimuli (see Rolls 1999a, 1995a, 1995b, 1990). It is also of interest that reward signals (e.g., the taste of food when we are hungry) are associated with subjective feelings of pleasure (see Rolls 1999a, 1997a, 1995a, 1997b, 1990). I sug-

gest that this correspondence arises because pleasure is the subjective state that represents in the conscious system a signal that is positively reinforcing (rewarding), and that inconsistent behaviour would result if the representations did not correspond to a signal for positive reinforcement in both the conscious and the non-conscious processing systems.

Do these arguments mean that the conscious sensation of e.g., taste quality (i.e., identity and intensity) is represented or occurs in the primary taste cortex, and of the pleasantness of taste in the secondary taste cortex, and that activity in these areas is sufficient for conscious sensations (qualia) to occur? I do not suggest this at all. Instead the arguments I have put forward above suggest that we are only conscious of representations when we have higher order thoughts about them. The implication then is that pathways must connect from each

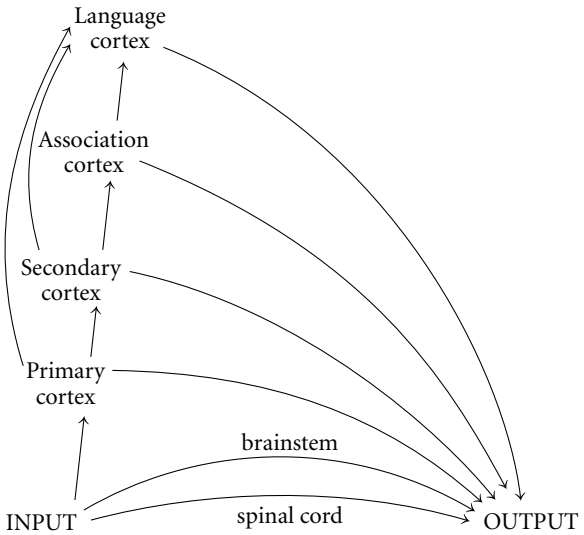

Figure 2. Schematic illustration indicating that early cortical stages in information processing may need access to language areas that bypass subsequent levels in the hierarchy, so that consciousness of what is represented in early cortical stages, and which may not be represented in later cortical stages, can occur. Higher-order syntactic thoughts could be implemented in the language cortex itself, which for the theory described here needs to implement syntactic processing on symbols, but not natural language. (Natural language implies a universal grammar.) Backprojections, a notable feature of cortical connectivity, with many probable functions including recall, attention, and influencing the categories formed in earlier cortical areas during learning , probably reciprocate all the connections shown. (After Rolls 1999a: Figure 9.3).

of the brain areas in which information is represented about which we can be conscious, to the system which has the higher order thoughts, which as I have argued above, requires language. Thus, in the example given, there must be connections to the language areas from the primary taste cortex, which need not be direct, but which must bypass the secondary taste cortex, in which the information is represented differently (see Figure 2 and Rolls 1995a). There must also be pathways from the secondary taste cortex, not necessarily direct, to the language areas so that we can have higher order thoughts about the pleasantness of the representation in the secondary taste cortex (see Figure 2). There would also need to be pathways from the hippocampus, implicated in the recall of declarative memories, back to the language areas of the cerebral cortex (at least via the cortical areas which receive backprojections from the amygdala, orbitofrontal cortex, and hippocampus, see Figure 1, which would in turn need connections to the language areas). I note that given the distributed nature of neuronal representations (see Rolls & Deco 2002), there need be no loss of information given the large numbers of neurons in any representation for information from early vs late cortical areas, provided that reasonable numbers of connections are present.

One of the arguments that Rosenthal (1993) uses to support a role for higher order thoughts in consciousness is that learning a repertoire of HOTs makes the two rather similar sensory inputs available to consciousness, and consciously discriminable. An example is with wine. The blackcurrant or peppermint notes in a wine may not be conscious at first; but after you have been trained with language, using different words to describe the different notes, then you become conscious of the different notes. So having a (higher order) thought may make some quality enter consciousness. However, words are inherently orthogonal representations, and these representations could by the top-down backprojections to earlier cortical areas make the representations of two rather similar qualities become more different in the early cortical areas by influencing the categories being formed in competitive networks during learning, in ways described by Rolls and Treves (Rolls & Treves 1998: Section 4.5) and by Rolls and Deco (Rolls & Deco 2002: Section 7.4.5). If this is the mechanism, then, at least after the learning, the different notes may be brought into consciousness because of a better sensory representation, rather than by a special operation of higher order thoughts on an unchanged sensory input.

A causal role for consciousness

One question that has been discussed is whether there is a causal role for consciousness (e.g., Armstrong & Malcolm 1984). The position to which the above arguments lead is that indeed conscious processing does have a causal role in the elicitation of behaviour, but only under the set of circumstances when higher order thoughts play a role in correcting or influencing lower order thoughts. The sense in which the consciousness is causal is then, it is suggested, that the higher order thought is causally involved in correcting the lower order thought; and that it is a property of the higher order thought system that it feels like something when it is operating. As we have seen, some behavioural responses can be elicited when there is not this type of reflective control of lower order processing, nor indeed any contribution of language (see further Rolls 2003 for relations between implicit and explicit processing). There are many brain processing routes to output regions, and only one of these involves conscious, verbally represented processing which can later be recalled (see Figure 1).

It is of interest to comment on how the evolution of a system for flexible planning might affect emotions. Consider grief which may occur when a reward is terminated and no immediate action is possible (see Rolls 1990). It may be adaptive by leading to a cessation of the formerly rewarded behaviour and thus facilitating the possible identification of other positive reinforcers in the environment. In humans, grief may be particularly potent because it becomes represented in a system which can plan ahead, and understand the enduring implications of the loss. (Thinking about or verbally discussing emotional states may also in these circumstances help, because this can lead towards the identification of new or alternative reinforcers, and of the realization that for example negative consequences may not be as bad as feared.)

Free will

This account of consciousness also leads to a suggestion about the processing that underlies the feeling of free will. Free will would in this scheme involve the use of language to check many moves ahead on a number of possible series of actions and their outcomes, and then with this information to make a choice from the likely outcomes of different possible series of actions. (If in contrast choices were made only on the basis of the reinforcement value of immediately available stimuli, without the arbitrary syntactic symbol manipulation made possible by language, then the choice strategy would be much more lim-

ited, and we might not want to use the term free will, as all the consequences of those actions would not have been computed.) It is suggested that when this type of reflective, conscious, information processing is occurring and leading to action, the system performing this processing and producing the action would have to believe that it (the system) could cause the action, for otherwise inconsistencies would arise, and the system might no longer try to initiate action. This belief held by the system may partly underlie the feeling of free will. At other times, when other brain modules are initiating actions (in the implicit systems), the conscious processor (the explicit system) may confabulate and believe that it caused the action, or at least give an account (possibly wrong) of why the action was initiated. The fact that the conscious processor may have the belief even in these circumstances that it initiated the action may arise as a property of it being inconsistent for a system which can take overall control using conscious verbal processing to believe that it was overridden by another system. This may be the reason why confabulation occurs, and one of the reasons for the feeling of the *unity of consciousness*. The feeling of the unity of consciousness may also be related to the suggested involvement of syntactic processing in consciousness, which appears to be inherently serial and with a limited binding capacity. Indeed, these properties of the implementation of syntax in the brain may provide some of the underlying computational reasons why consciousness feels unitary.

In the operation of such a free will system, the uncertainties introduced by the limited information possible about the likely outcomes of series of actions, and the inability to use optimal algorithms when combining conditional probabilities, would be much more important factors than whether the brain operates deterministically or not. (The operation of brain machinery must be relatively deterministic, for it has evolved to provide reliable outputs for given inputs. What I am suggesting here though is that an interesting question about free will is not whether it reflects the operation of deterministic machinery or not, but instead is what information processing and computations are taking place when we feel that we have free will, and the confabulations that may arise when we operate using implicit computations, i.e. computations which are not available to consciousness, such as for example the reward reversal implemented by the orbitofrontal cortex, see Rolls et al. 1994)

Self-identity and the unity of consciousness

Before leaving these thoughts, it may be worth commenting on the feeling of continuing self-identity that is characteristic of humans, and the unity of con-

sciousness. Why might these arise? One suggestion is that if one is an organism that can think about its own long-term multi-step plans, then for those plans to be consistently and thus adaptively executed, the goals of the plans would need to remain stable, as would memories of how far one had proceeded along the execution path of each plan. If one felt each time one came to execute, perhaps on another day, the next step of a plan, that the goals were different; or if one did not remember which steps had already been taken in a multi-step plan, the plan would never be usefully executed. So, given that it does feel like something to be doing this type of planning using higher order thoughts, it would have to feel as if one were the same agent, acting towards the same goals, from day to day. Thus it is suggested that the feeling of continuing self-identity falls out of a situation in which there is an actor with consistent long-term goals, and long-term recall. If it feels like anything to be the actor, according to the suggestions of the higher order thought theory, then it should feel like the same thing from occasion to occasion to be the actor, and no special further construct is needed to account for self-identity. Humans without such a feeling of being the same person from day to day might be expected to have for example inconsistent goals from day to day, or a poor recall memory. It may be noted that the ability to recall previous steps in a plan, and bring them into the conscious, higher-order thought system, is an important prerequisite for long-term planning which involves checking each step in a multi-step process.

These are my initial thoughts on why we have consciousness, and are conscious of sensory, emotional and motivational qualia, as well as qualia associated with first-order linguistic thoughts. However, as stated above, one does not feel that there are straightforward criteria in this philosophical field of enquiry for knowing whether the suggested theory is correct; so it is likely that theories of consciousness will continue to undergo rapid development; and current theories should not be taken to have practical implications.

3. Dual routes to action, and decisions

The question arises of how decisions are made in animals such as humans that have both the implicit, direct reward-based, and the explicit, rational, planning systems (see Figure 1). One particular situation in which the first, implicit, system may be especially important is when rapid reactions to stimuli with reward or punishment value must be made, for then the direct connections from structures such as the orbitofrontal cortex to the basal ganglia may allow rapid actions (e.g. Rolls 1994). Another is when there may be too many fac-

tors to be taken into account easily by the explicit, rational, planning, system, when the implicit system may be used to guide action. In contrast, when the implicit system continually makes errors, it would then be beneficial for the organism to switch from automatic, direct, action, based on obtaining what the orbitofrontal cortex system decodes as being the most positively reinforcing choice currently available, to the explicit conscious control system which can evaluate with its long-term planning algorithms what action should be performed next. Indeed, it would be adaptive for the explicit system to regularly be assessing performance by the more automatic system, and to switch itself in to control behaviour quite frequently, as otherwise the adaptive value of having the explicit system would be less than optimal. Another factor which may influence the balance between control by the implicit and explicit systems is the presence of pharmacological agents such as alcohol, which may alter the balance towards control by the implicit system, may allow the implicit system to influence more the explanations made by the explicit system, and may within the explicit system alter the relative value it places on caution and restraint versus commitment to a risky action or plan.

There may also be a flow of influence from the explicit, verbal system to the implicit system, in that the explicit system may decide on a plan of action or strategy, and exert an influence on the implicit system which will alter the reinforcement evaluations made by and the signals produced by the implicit system. An example of this might be that if a pregnant woman feels that she would like to escape a cruel mate, but is aware that she may not survive in the jungle, then it would be adaptive if the explicit system could suppress some aspects of her implicit behaviour towards her mate, so that she does not give signals that she is displeased with her situation. [In the literature on self-deception, it has been suggested that unconscious desires may not be made explicit in consciousness (or actually repressed), so as not to compromise the explicit system in what it produces (see e.g., Alexander 1975, 1979; and the review by Nesse & Lloyd 1992; Trivers 1976, 1985)]. Another example might be that the explicit system might because of its long-term plans influence the implicit system to increase its response to for example a positive reinforcer. One way in which the explicit system might influence the implicit system is by setting up the conditions in which for example when a given stimulus (e.g., person) is present, positive reinforcers are given, to facilitate stimulus-reinforcement association learning by the implicit system of the person receiving the positive reinforcers. Conversely, the implicit system may influence the explicit system, for example by highlighting certain stimuli in the environment that are currently associated with reward, to guide the attention of the explicit system to such stimuli.

However, it may be expected that there is often a conflict between these systems, in that the first, implicit, system is able to guide behaviour particularly to obtain the greatest immediate reinforcement, whereas the explicit system can potentially enable immediate rewards to be deferred, and longer-term, multistep, plans to be formed. This type of conflict will occur in animals with a syntactic planning ability, that is in humans and any other animals that have the ability to process a series of "if. . .then" stages of planning. This is a property of the human language system, and the extent to which it is a property of non-human primates is not yet fully clear. In any case, such conflict may be an important aspect of the operation of at least the human mind, because it is so essential for humans to correctly decide, at every moment, whether to invest in a relationship or a group that may offer long-term benefits, or whether to directly pursue immediate benefits (Nesse & Lloyd 1992). As Nesse and Lloyd (Nesse & Lloyd 1992) describe, analysts have come to a somewhat similar position, for they hold that intrapsychic conflicts usually seem to have two sides, with impulses on one side and inhibitions on the other. Analysts describe the source of the impulses as the *id*, and the modules that inhibit the expression of impulses, because of external and internal constraints, the *ego* and *superego* respectively (Leak & Christopher 1982; see Nesse & Lloyd 1992: 613). The superego can be thought of as the conscience, while the ego is the locus of executive functions that balance satisfaction of impulses with anticipated internal and external costs. A difference of the present position is that it is based on identification of dual routes to action implemented by different systems in the brain, each with its own selective advantage.

Some investigations in non-human primates on deception have been interpreted as showing that animals can plan to deceive others (see, e.g., Trivers 1985), that is to utilize "Machiavellian intelligence". For example, a baboon may "deliberately" mislead another animal in order to obtain a resource such as food (e.g., by screaming to summon assistance in order to have a competing animal chased from a food patch) or sex (e.g., a female baboon who very gradually moved into a position from which the dominant male could not see her grooming a subadult baboon) (see Dawkins 1993). The attraction of the Machiavellian argument is that the behaviour for which it accounts seems to imply that there is a concept of another animal's mind, and that one animal is trying occasionally to mislead another, which implies some planning. However, such observations tend by their nature to be field-based, and may have an anecdotal character, in that the previous experience of the animals in this type of behaviour, and the reinforcements obtained, are not known (Dawkins 1993). It is possible for example that some behavioural responses that appear to

be Machiavellian may have been the result of previous instrumental learning in which reinforcement was obtained for particular types of response, or of observational learning, with again learning from the outcome observed. However, in any case, most examples of Machiavellian intelligence in non-human primates do not involve multiple stages of "if. . .then" planning requiring syntax to keep the symbols apart (but may involve learning of the type "if the dominant male sees me grooming a subadult male, I will be punished") (see Dawkins 1993). Nevertheless, the possible advantage of such Machiavellian *planning* could be one of the adaptive guiding factors in evolution which provided advantage to a multi-step, syntactic system which enables long-term planning, the best example of such a system being human language. However, another, not necessarily exclusive, advantage for the evolution of a linguistic multi-step planning system could well be not Machiavellian planning, but planning for social co-operation and advantage. Perhaps in general an "if. . .then" multi-step syntactic planning ability is useful primarily in evolution in social situations of the type: "if X does this, then Y does that; then I would / should do that, and the outcome would be . . .". It is not yet at all clear whether such planning is required in order to explain the social behaviour of social animals such as hunting dogs; or socialising monkeys (Dawkins 1993). However, in humans there is evidence that members of "primitive" hunting tribes spend hours recounting tales of recent events (perhaps who did what, when; who then did what, etc.), perhaps to help learn from experience about good strategies, necessary for example when physically weak men take on large animals (see Pinker & Bloom 1992). Thus, social co-operation may be as powerful a driving force in the evolution of syntactical planning systems as Machiavellian intelligence. What is common to both is that they involve social situations. However, such a syntactic planning system would have advantages not only in social systems, for such planning may be useful in obtaining resources purely in a physical (non-social) world. An example might be planning how to cross terrain given current environmental constraints in order to reach a particular place.

The thrust of this argument thus is that much complex animal including human behaviour can take place using the implicit, non-conscious, route to action. We should be very careful not to postulate intentional states (i.e., states with intentions, beliefs and desires) unless the evidence for them is strong, and it seems to me that a flexible, one-off, linguistic processing system that can handle propositions is needed for intentional states. What the explicit, linguistic, system does allow is exactly this flexible, one-off, multi-step planning ahead type of computation, which allows us to defer immediate rewards based on such a plan.

This consideration of dual routes to action has been with respect to the behaviour produced. There is of course in addition a third output of brain regions such as the orbitofrontal cortex and amygdala involved in emotion, that is directed to producing autonomic and endocrine responses. Although it has been argued by Rolls (Rolls 1999a: Ch. 3) that the autonomic system is not normally in a circuit through which behavioural responses are produced (i.e., against the James-Lange and related somatic theories), there may be some influence from effects produced through the endocrine system (and possibly the autonomic system, through which some endocrine responses are controlled) on behaviour, or on the dual systems just discussed which control behaviour. For example, during female orgasm the hormone oxytocin may be released, and this may influence the implicit system to help develop positive reinforcement associations and thus attachment.

4. Discussion

Some ways in which the current theory may be different from other related theories follow. The current theory holds that it is higher order *syntactic* thoughts (HOSTs) that are closely associated with consciousness, and this may differ from Rosenthal's higher order thoughts (HOTs) theory (1990, 1993; Rosenthal 1986), in the emphasis in the current theory on language. Language in the current theory is defined by syntactic manipulation of symbols, and does not necessarily imply verbal or natural language. The type of language required in the theory described here is sometimes termed "mentalese" by philosophers. The reason that strong emphasis is placed on language is that it is as a result of having a multi-step flexible "on the fly" reasoning procedure that errors which cannot be easily corrected by reward or punishment received at the end of the reasoning, need 'thoughts about thoughts', that is some type of supervisory and monitoring process, to detect where errors in the reasoning have occurred. This suggestion on the adaptive value in evolution of such a higher order linguistic thought process for multi-step planning ahead, and correcting such plans, may also be different from earlier work. Put another way, this point is that *credit assignment* when reward or punishment are received is straightforward in a one layer network (in which the reinforcement can be used directly to correct nodes in error, or responses); but is very difficult in a multi-step linguistic process executed once "on the fly". Very complex mappings in a multilayer network can be learned if hundreds of learning trials are provided. But once these complex mappings are learned, their success or failure in a new situation on

a given trial cannot be evaluated and corrected by the network. Indeed, the complex mappings achieved by such networks (e.g., backpropagation nets, see Rolls & Deco 2002) mean that after training they operate according to fixed rules, and are often quite impenetrable and inflexible. In contrast, to correct a multi-step, single occasion, linguistically based plan or procedure, recall of the steps just made in the reasoning or planning, and perhaps related episodic material, needs to occur, so that the link in the chain which is most likely to be in error can be identified. This may be part of the reason why there is a close relation between declarative memory systems, which can explicitly recall memories, and consciousness.

Some computer programs may have supervisory processes. Should these count as higher order linguistic thought processes? My current response to this is that they should not, to the extent that they operate with fixed rules to correct the operation of a system which does not itself involve linguistic thoughts about symbols grounded semantically in the external world. If on the other hand it were possible to implement on a computer such a high order linguistic thought supervisory correction process to correct first order one-off linguistic thoughts with symbols grounded in the real world, then this process would *prima facie* be conscious. If it were possible in a thought experiment to reproduce the neural connectivity and operation of a human brain on a computer, then *prima facie* it would also have the attributes of consciousness. It might continue to have those attributes for as long as power was applied to the system.

Another possible difference from earlier theories is that raw sensory feels are suggested to arise as a consequence of having a system that can think about its own thoughts. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution.

A property often attributed to consciousness is that it is *unitary*. The current theory would account for this by the limited syntactic capability of neuronal networks in the brain, which render it difficult to implement more than a few syntactic bindings of symbols simultaneously (McLeod et al. 1998; see Rolls & Treves 1998). This limitation makes it difficult to run several "streams of consciousness" simultaneously. In addition, given that a linguistic system can control behavioural output, several parallel streams might produce maladaptive behaviour (apparent as e.g., indecision), and might be selected against. The close relation between, and the limited capacity of, both the stream of consciousness, and auditory-verbal short term memory, may be that both implement the capacity for syntax in neural networks. Whether syntax in real neuronal networks is implemented by temporal binding is still very much an unresolved issue (Rolls & Deco 2002; Rolls & Treves 1998). However, the hypothesis

that syntactic binding is necessary for consciousness is one of the postulates of the theory I am describing (for the system I describe must be capable of correcting its own syntactic thoughts); and the fact that the binding must be implemented in neuronal networks may well place limitations on consciousness, which lead to some of its properties, such as its unitary nature. The postulate of Crick and Koch (Crick & Koch 1990) that oscillations and synchronization are necessary bases of consciousness could thus be related to the present theory if it turns out that oscillations or neuronal synchronization are the way the brain implements syntactic binding. However, the fact that oscillations and neuronal synchronization are especially evident in anaesthetized cats does not impress as strong evidence that oscillations and synchronization are critical features of consciousness, for most people would hold that anaesthetized cats are *not* conscious. The fact that oscillations and synchronization are much more difficult to demonstrate in the temporal cortical visual areas of awake behaving monkeys might just mean that during evolution to primates the cortex has become better able to avoid parasitic oscillations, as a result of developing better feedforward and feedback inhibitory circuits (see Rolls & Deco 2002; Rolls & Treves 1998).

The current theory holds that consciousness arises by virtue of a system that can think linguistically about its own linguistic thoughts. The advantages for a system of being able to do this have been described, and this has been suggested as the reason why consciousness evolved. The evidence that consciousness arises by virtue of having a system that can perform higher order linguistic processing is however, and I think may remain, circumstantial. (Why must it feel like something when we are performing a certain type of information processing? The evidence described here suggests that it does feel like something when we are performing a certain type of information processing, but does not produce a strong reason for why it has to feel like something. It just does, when we are using this linguistic processing system capable of higher order thoughts.) The evidence, summarized above, includes the points that we think of ourselves as conscious when for example we recall earlier events, compare them with current events, and plan many steps ahead. Evidence also comes from neurological cases, from for example split brain patients (who may confabulate conscious stories about what is happening in their other, non-language, hemisphere; and from cases such as frontal lobe patients who can tell one consciously what they should be doing, but nevertheless may be doing the opposite. (The force of this type of case is that much of our behaviour may normally be produced by routes about which we cannot verbalize, and are not conscious about.) This raises the issue of the causal role of consciousness. Does

consciousness cause our behaviour?[1] The view that I currently hold is that the information processing which is related to consciousness (activity in a linguistic system capable of higher order thoughts, and used for planning and correcting the operation of lower order linguistic systems) can play a causal role in producing our behaviour (see Figure 1). It is, I postulate, a *property* of processing in this system (capable of higher order thoughts) that it feels like something to be performing that type of processing. It is in this sense that I suggest that consciousness can act causally to influence our behaviour: consciousness is the property that occurs when a linguistic system is thinking about its lower order thoughts. The hypothesis that it does feel like something when this processing is taking place is at least to some extent testable: humans performing this type of higher order linguistic processing, for example recalling episodic memories and comparing them with current circumstances, who denied being conscious, would *prima facie* constitute evidence against the theory. Most humans would find it very implausible though to posit that they could be thinking about their own thoughts, and reflecting on their own thoughts, without being conscious. This type of processing does appear to be for most humans to be necessarily conscious.

Finally, I provide a short specification of what might have to be implemented in a neural network to implement conscious processing. First, a linguistic system, not necessarily verbal, but implementing syntax between symbols implemented in the environment would be needed (i.e., a "mentalese" language capability). Then a higher order thought system also implementing syntax and able to think about the representations in the first order language system, and able to correct the reasoning in the first order linguistic system in a flexible manner, would be needed. So my view is that consciousness can be implemented in neural networks, (and that this is a topic worth discussing), but that the neural networks would have to implement the type of higher order linguistic processing described in this Chapter.

5. Conclusion

It is suggested that it feels like something to be an organism or machine that can think about its own (linguistic, and semantically based) thoughts. It is suggested that qualia, raw sensory and emotional feels, arise secondary to having evolved such a higher order thought system, and that sensory and emotional processing feels like something because it would be unparsimonious for it to enter the planning, higher order thought, system and *not* feel like something.

The adaptive value of having sensory and emotional feelings, or qualia, is thus suggested to be that such inputs are important to the long-term planning, explicit, processing system. Raw sensory feels, and subjective states associated with emotional and motivational states, may not necessarily arise first in evolution. Some issues that arise in relation to this theory are discussed by Rolls (Rolls 2000e); reasons why the ventral visual system is more closely related to explicit than implicit processing are considered by Rolls and Deco (2002) and by Rolls (2003): and reasons why explicit, conscious, processing may have a higher threshold in sensory processing than implicit processing are considered by Rolls (2003).

It may be useful to comment that this theory, while sharing much with Rosenthal's HOT theory of consciousness (1990, 1993, 1986), may be different in a number of respects. I take an information processing / computational approach, and try to define the computations that appear to be taking place when we are conscious. I take a brain design approach, and try to link implicit vs conscious computations with processes in different brain regions, as a way of constraining the theory, and providing useful links to medical concerns and issues. I propose that it is syntactic thoughts implementing multistep planning operations that raise a credit assignment problem, and that the *adaptive value* of higher order thoughts is to help solve this credit assignment problem. My HOST theory is based therefore on the adaptive utility of higher order thoughts for correcting lower order syntactic operations such as those involved in planning, and it is because the higher order system must be able to correct a lower order syntactic processing system that the higher order system needs to be a Higher Order Syntactic Thought (HOST) system. When I use the term "thoughts", I can thus be read as meaning "human thoughts", which imply the ability to perform syntactic processing, that is to implement the correct and flexible binding of symbols in relational operations. (This does not imply a need for human language, and I am open-minded about the extent to which this type of information processing may be possible in animals.) When Rosenthal uses the term "thought" he may be using it in a wider sense, and I hope that this is an issue dealt with by Rosenthal (2004).

Acknowledgements

I am very grateful to David Rosenthal (City University of New York), Marian Dawkins (Oxford University), and Martin Davies (Australian National

University) for many fascinating and inspiring discussions on the topics considered here.

Note

1. This raises the issue of the causal relation between mental events and neurophysiological events, part of the mind-body problem. My view is that the relation between mental events and neurophysiological events is similar (apart from the problem of consciousness) to the relation between the program running in a computer and the hardware on the computer. In a sense, the program causes the logic gates to move to the next state. This move causes the program to move to its next state. Effectively, we are looking at different levels of what is overall the operation of a *system*, and causal explanations can usefully be understood as operating both within levels (causing one step of the program to move to the next), as well as between levels (e.g., software to hardware and vice versa). This is the solution I propose to this aspect of the mind-body (or mind-brain) problem.

References

Alexander, R. D. (1975). The search for a general theory of behavior. *Behav. Sci., 20*, 77–100.

Alexander, R. D. (1979). *Darwinism and Human Affairs*. Seattle: University of Washington Press.

Allport, A. (1988). What concept of consciousness? In Marcel A. J. & Bisiach E. (Eds.), *Conscousness in Contemporary Science* (pp. 159–182). Oxford: Oxford University Press.

Armstrong, D. M. & Malcolm, N. (1984). *Consciousness and Causality*. Oxford: Blackwell.

Barlow, H. B. (1997). Single neurons, communal goals, and consciousness. In Ito M., Miyashita Y., & Rolls E. T. *Cognition, Computation, and Consciousness* (pp. 121–136). Oxford: Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behav. Brain. Sci., 18*, 227–247.

Booth, D. A. (1985). Food-conditioned eating preferences and aversions with interoceptive elements: Learned appetites and satieties. *Ann. N.Y. Acad. Sci., 443*, 22–37.

Carruthers, P. (1996). *Language, Thought and Consciosness*. Cambrudge: Cambridge University Press.

Chalmers, D. J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Cheney, D. L. & Seyfarth, R. M. (1990). *How Monkeys See the World*. Chicago: University of Chicago Press.

Crick, F. H. C. & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Sem. Neurosci., 2*, 263–275.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals*. Chicago: University of Chicago Press.

Dawkins, M. S. (1993). *Through Our Eyes Only? The Search for Animal Consciousness*. Oxford: Freeman.

Dennett, D. C. (1991). *Consciousness Explained*. London: Penguin.
Ekman, P. (1982). *Emotion in the Human Face*. Cambridge: Cambridge University Press.
Ekman, P. (1993). Facial expression and emotion. *Am. Psych., 48*, 384–392.
Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.
Fodor, J. A. (1990). *A Theory of Content and other Essays*. Cambridge, MA: MIT Press.
Fodor, J. A. (1994). *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, MA: MIT Press.
Gazzaniga, M. S. & LeDoux, J. (1978). *The Integrated Mind*. New York: Plenum.
Gazzaniga, M. S. (1988). Brain modularity: towards a phlosophy of conscious experience. In Marcel A. J. & Bisiach E. (Eds.), *Consciousness in Contempoary Science* (pp. 218–238). Oxford: Oxford University Press.
Gazzaniga, M. S. (1995). Consciousness and the cerebral hemispheres. In Gazzaniga M. S. (Ed.), *The Cognitive Neurosciences* (pp. 1392–1400). Cambridge, Mass.: MIT Press.
Goldman-Rakic, P. S. (1996). The prefrontal landscape: Implications of functional architecture for understanding human mentation and the central executive. *Phil. Trans. R. Soc. Lond. B, 351*, 1445–1453.
Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., & Polkey, C.E. (2003a). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain, 126*, 1691–1712.
Hornak, J., O'Doherty, J., Bramham, J., Rolls, E. T., Morris, R. G., Bullock, P. R., & Polkey, C. E. (2003b). Reward-related reversal learning after surgical excisions in orbitofrontal and dorsolateral prefrontal cortex in humans. *J. Cogn. Neurosci.*.
Humphrey, N. K. (1980). Nature's psychologists. In Josephson B. D. & Ramachandran V. S. (Eds.), *Consciousness and the Physical World* (pp. 57–80). Oxford: Pergamon.
Humphrey, N. K. (1986). *The Inner Eye*. London: Faber.
Krebs, J. R. & Kacelnik, A. (1991). Decision making. In Krebs J. R. & Davies N. B. (Eds.), *Behavioural Ecology* (3rd ed.), (pp. 105–136). Oxford: Blackwell.
Leak, G. K. & Christopher, S. B. (1982). Freudian psychoanalysis and sociobiology: a synthesis. *Am. Psych., 37*, 313–322.
McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.
Nesse, R. M. & Lloyd, A. T. (1992). The evolution of psychodynamic mechanisms. In Barkow J. H., Cosmides L., & Tooby J. (Eds.), *The Adapted Mind* (pp. 601–624). New York: Oxford University Press.
Oatley, K. & Jenkins, J. M. (1996). *Understanding Emotions*. Oxford: Backwell.
Petrides, M. (1996). Specialized systems for the processing of mnemonic information within the primate frontal cortex. *Philosophical Transactions of the Royal Society B, 351*, 1455–1462.
Pinker, S. & Bloom, P. (1992). Natural language and natural selection. In Barkow J. H., Cosmides L., & Tooby J. (Eds.), *The Adapted Mind* (pp. 451–493). New York: Oxford University Press.
Plata-Salaman, C. R., Smith-Swintosky, V. L., & Scott, T. R. (1996). Gustatory neural coding in the monkey cortex: Mixtures. *J. Neurophysiol.*, 2369–2379.

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cog. Emot., 4*, 161–190.

Rolls, E. T. (1994). Neurophysiology and cognitive functions of the striatum. *Rev. Neurol. (Paris), 150*, 648–660.

Rolls, E. T. (1995a). Central taste anatomy and neurophysiology. In Doty R. L. (Ed.), *Handbook of Olfaction and Gustation* (pp. 549–573). New York.: Dekker.

Rolls, E. T. (1995b). A theory of emotion and consciousness, and its application to understanding the neural basis of emotion. In Gazzaniga M. S. (Ed.), *The Cognitive Neurosciences* (pp. 1091–1106). Cambridge, Mass.: MIT Press.

Rolls, E. T. (1996). A theory of hippocampal function in memory. *Hippocampus, 6*, 601–620.

Rolls, E. T. (1997a).Brain mechanisms of vision, memory, and consciousness. In Ito M., Miyashita Y., & Rolls E. T. (Eds.), *Cognition, Computation, and Consciousness* (pp. 81–120). Oxford: Oxford University Press.

Rolls, E. T. (1997b). Taste and olfactory processing in the brain and its relation to the control of eating. *Critical Reviews in Neurobiology, 11*, 263–287.

Rolls, E. T. (1999a). *The Brain and Emotion*. Oxford: Oxford University Press.

Rolls, E. T. (1999b). The functions of the orbitofrontal cortex. *Neurocase, 5*, 301–312.

Rolls, E. T. (2000a). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. *Neuron, 27*, 205–218.

Rolls, E. T. (2000b). Hippocampo-cortical and cortico-cortical backprojections. *Hippocampus, 10*, 380–388.

Rolls, E. T. (2000c). Neurophysiology and functions of the primate amygdala, and the neural basis of emotion. In Aggleton J. P. (Ed.), *The Amygdala: A Functional Analysis* (Second Edition ed.). Oxford: Oxford University Press.

Rolls, E. T. (2000d). The orbitofrontal cortex and reward. *Cereb Cortex, 10*, 284–294.

Rolls, E. T. (2000e). Précis of the brain and emotion. *Behav. Brain Sci., 23*, 177–233.

Rolls, E. T. (2002). The functions of the orbitofrontal cortex. In Stuss D. T. & Knight, R. T. (Eds.), *Principles of Frontal Lobe Function* (pp. 354–375). New York: Oxford University Press.

Rolls, E. T. (2003). Consciousness absent and present: a neurophysiological exploration. *Progress in Brain Research, 144*, 95–106.

Rolls, E. T. & Johnstone, S. (1992). Neurophysiological analysis of striatal function. In Vallar G. & Wallesch C. W. (Eds.), *Neuropsychological Disorders Associated with Subcortical Lesions* (pp. 61–97). Oxford: Oxford University Press.

Rolls, E. T., Hornak, J., Wade, D., & McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. *J. Neurol. Neurosurg. and Psychiat., 57*, 1518–1524.

Rolls, E. T. & Treves, A. (1998). *Neural Networks and Brain Function*. Oxford, UK: Oxford University Press.

Rolls, E. T., Treves, A., Robertson, R. G., Georges-Francois, P., & Panzeri, S. (1998). Information about spatial view in an ensemble of primate hippocampal cells. *J. Neurophysiol., 79*, 1797–1813.

Rolls, E. T. & Stringer, S. M. (2001). A model of the interaction between mood and memory. *Network: Computation in Neural Systems, 12*, 111–129.

Rolls, E. T. & Deco, G. (2002). *Computational Neuroscience of Vision*. Oxford: Oxford University Press.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Phil. Stud., 49*, 329–359.

Rosenthal, D. M. (1990). A theory of consciousness. In *ZIF. Bielefeld*. Germany: Zentrum für Interdisziplinaire Forschung.

Rosenthal, D. M. (1993). Thinking that one thinks. In Davies M. & Humphreys, G. W. (Eds.), *Consciousness* (pp. 197–223). Oxford: Blackwell.

Rosenthal, D. M. (2004). Varieties of Higher-Order Theory. In Gennaro R. J. (Ed.), *Higher Order Theories of Consciousness*. Amsterdam: John Benjamins.

Schiffman, S. S. & Erikson, R. P. (1971). A psychophysical model for gustatory quality. *Physiol. Behav., 7*, 617–633.

Shallice, T. & Burgess, P. (1996). The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society B, 351*, 1405–1411.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, 49–65.

Smith-Swintosky, V. L., Plata-Salaman, C. R., & Scott, T. R. (1991). Gustatory neural encoding in the monkey cortex: stimulus quality. *J. Neurophysiol., 66*, 1156–1165.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys and humans. *Psych. Rev., 99*, 195–231.

Treves, A. & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus, 4*, 374–391.

Trivers, R. L. (1976). Foreword. In Dawkins R. (Ed.), *The Selfish Gene*. Oxford: Oxford Univerisity Press.

Trivers, R. L. (1985). *Social Evolution*. California: Benjamin Cummings.

von der Malsburg, C. (1990). A neural architecture for the representation of scenes. In McGaugh J. L., Weinberger N. M., & Lynch G. (Eds.), *Brain Organisation and Memory: Cells, Systems and Circuits* (pp. 356–372). New York, NY: Oxford University Press.

Yaxley, S. T. RE, & Sienkiewicz, Z. J. (1990). Gustatory responses of single neurons in the insula of the macaque monkey. *J. Neurophysiol., 63*, 689–700.

Chapter 8

Assumptions of a subjective measure of consciousness

Three mappings

Zoltán Dienes and Josef Perner

Consider some event in the world, for example there being an object moving up. There are two distinct ways of knowing this event: being merely aware of it, or being consciously aware of it. We are aware of the event when the event influences cognitive processing or behaviour, we are sensitive to the event in some way. But just because we are sensitive to the event, that does not mean we are consciously aware of it. Similarly, we might know some regularity in the world, e.g. we may be behaviourally sensitive to the perceptual cues that indicate whether a chick is male or female. But that does not mean we consciously know the regularity. Experimental psychologists have debated for over a century how to determine whether we can be aware of stimuli or regularities (sensitive to them in some way) without being consciously aware of them (e.g. Peirce & Jastrow 1884). The debate has been heated (e.g. Holender 1986; Shanks & St. John 1994; Dienes & Perner 1999; Merikle & Daneman 1998), and we believe higher order theories can help provide clarity in evaluating useful criteria for deciding when people consciously know rather than simply know. Rosenthal (e.g. 1986, 2000, forthcoming, this volume) suggests we adopt a common intuition that we consciously know (that something is so) when we are conscious of the mental state by which we know. It is difficult to regard a person as consciously seeing if, despite 100% correct discrimination performance on a task, the person vigorously and adamantly insists that they see nothing, they are just guessing, they are not conscious of seeing at all (Weiskrantz 1988, 1997). Thus, Weiskrantz argued that conscious seeing required a separate commentary by the person on the visual processing of the stimulus, i.e. on the mental state of seeing itself. Similarly, Rosenthal suggested that a mental state is conscious when we have a roughly contemporaneous thought (a higher order thought) to the effect that

we are in that state. Carruthers (1992, 2000, this volume) likewise argued that to understand phenomena such as blindsight we need to be able to distinguish worldly subjectivity (representing the world, a first order representation) and experiential subjectivity (representing one's mental states), the latter providing conscious awareness.

For adult humans, seeing is understood as a means of knowing; merely guessing that something is so cannot be a case of seeing that it is so. Thus, on higher order theory, when a mental state of e.g. seeing (that something is so) is a case of consciously seeing (that something is so), we (adults) would be able to distinguish between whether we merely guessed that something was the case, or that we knew it to some extent. Similarly, for any other occurrent knowledge state: To know consciously entails we can distinguish between knowing (to some degree) and guessing.

Higher order theories lead naturally to criteria for assessing conscious knowledge, namely criteria based on people's ability to determine the mental state that they are in, and that provides the knowledge (so called "subjective measures" of consciousness). Criteria based on merely the person's ability to determine the stimulus that was shown ("objective measures" of consciousness) could easily lead to false positives: First-order mental states allow the discrimination of stimuli (indeed, that is generally their function) and second order states are not needed at all for discrimination. Objective discrimination of stimuli entails that you are aware of the stimuli, but not that you are consciously aware of them. Subjective measures, which directly test for the existence of second order states, thus (according to the theory) directly test for the existence of conscious awareness.

One such subjective criterion is the "zero confidence-accuracy relationship criterion", normally called the "zero correlation criterion" for short (Dienes & Berry 1997). When the subject makes a judgement, ask the subject to distinguish between guessing and different degrees of knowing. If the judgment expresses conscious knowledge – on those cases when it is knowledge and not guessing – then the subject should give a higher confidence rating when she actually knows the answer and a lower confidence rating when she is just guessing. In other words, conscious knowledge would prima facie be revealed by a correlation between confidence and accuracy, and unconscious knowledge by no correlation (the person does not know when she is guessing and when she is applying knowledge).

Another criterion based on a person's ability to determine the mental state that they are in is the "guessing criterion" (Dienes et al. 1995). Take all the cases where a person says they are guessing, and see if they are actually demonstrat-

ing the use of knowledge. This is the criterion that is satisfied in cases of blindsight (Weiskrantz 1988, 1997). The person insists they are just guessing, but they can be discriminating 90% correctly. So, based on the guessing criterion, the knowledge in blindsight is unconscious.

While the zero correlation and guessing criteria have face validity as measures of conscious awareness, are there conditions in which the criteria would give the wrong answer? What do we need to assume to ensure their validity? Dienes (in press) used Rosenthal's higher order thought theory to flesh out some assumptions behind both the zero correlation and guessing criteria, regarding the extent to which the criteria could be biased.[1] In this paper, we will use higher order thought theory to consider some further assumptions required for using the zero correlation criterion.[2]

We will first summarize the contexts in which the zero-correlation criterion has been previously applied, and then consider in detail its application to, first, implicit learning, and then subliminal perception.

Previous use of the zero correlation criterion

The zero-correlation criterion has now been applied extensively in the implicit learning literature. Implicit learning occurs when people learn by acquiring unconscious knowledge (for reviews see Shanks & St. John 1994; Dienes & Berry 1997; Cleeremans, Destrebecqz, & Boyer 1998). The term "implicit learning"was introduced by Reber (1967). Reber asked subjects to memorize strings of letters, where, unbeknownst to subjects, the order of letters within the string was constrained by a complex set of rules (i.e. an artificial grammar). After a few minutes of memorizing strings, the subjects were told about the existence of the rules (but not what they were) and asked to classify new strings as obeying the rules or not. Reber (see Reber 1993, for a review of his work) found that subjects could classify new strings 60–70% correctly on average, while finding it difficult to say what the rules were that guided their performance. He argued the knowledge was unconscious. But, starting with Dulany, Carlson, and Dewey (1984), critics have been unhappy with free report as an indicator of unconscious knowledge. Free report gives the subject the option of not stating some knowledge if they choose not to (by virtue of not being certain enough of it); and if the free report is requested some time after the decision, the subject might momentarily forget some of the bits of knowledge they brought to bear on the task.

Chan (1992) elicited a confidence rating in each classification decision, and showed subjects were no more confident in correct than incorrect decisions. Dienes et al. (1995), Dienes and Altmann (1997), Allwood, Granhag, and Johansson (2000), Channon et al. (2002), Tunney and Altmann (2001) and Dienes and Perner (2003) replicated these results, finding some conditions under which there was no within-subject relationship between confidence and accuracy. We argued this indicated subjects could not discriminate between mental states providing knowledge and e.g. those just corresponding to guessing; hence, the mental states were unconscious (see also Kelley, Burton, Kato, & Akamatsu 2001; Newell & Bright 2002, who used the same lack of relationship between confidence and accuracy to argue for the use of unconscious knowledge in other learning paradigms). The method has an advantage over free report in that low confidence is no longer a means by which relevant conscious knowledge is excluded from measurement; rather the confidence itself becomes the object of study and can be directly assessed on every trial.

Kolb and Braun (1995) and Kunimoto, Miller, and Pashler (2001) applied a similar methodology to perception. Kolb and Braun investigated texture discrimination and Kunimoto et al. word perception. They both found conditions under which confidence was not related to accuracy, and argued this demonstrated the existence of unconscious perception (for a failure to replicate Kolb and Braun, see Morgan, Mason, & Solomon 1997).

We will now consider some conditions that should be satisfied for the zero-correlation criterion to provide valid answers about the conscious status of mental states. After some preliminary considerations, we will consider implicit learning in detail, and then subliminal perception.

Preliminary considerations: Attitudes and higher order thoughts

Human beings meet the world with a highly evolved learning system. Over evolutionary history this learning system has faced environments with specific statistical and other kinds of structures. When it meets a new environment or structured domain it can implicitly presume the structure will belong to a certain class of structure types and the need is to determine the parameters that specify that class of structure. In a first encounter with a domain recognized as novel in important respects, there will be uncertainty as to the right parameter values. If we are asked to decide whether a chick is male or female by looking at it, we may initially be just guessing. After much practice, we may make judgements with certainty. The system in making a judgement ("the chick is

male"), has an attitude toward that judgement (guessing, knowing) that expresses itself in how the content is used, for example, the consistency with which it is used, the amount of contrary evidence that would overthrow it, etc. In between the attitudes of guessing and knowing there can be degrees of confidence, or thinking with some conviction. The English language does not express these attitudes very well; if one says there is an attitude of knowing or guessing, in everyday usage it may imply there must be second or even third order thoughts about guessing or knowing. But for the purposes of this paper consider guessing, thinking with some conviction etc. as being first order mental states, regardless of any higher order thoughts that may or may not be present. There can be an attitude (knowing, guessing, etc.) independently of being aware of that attitude.

Consider, for example, one possible way in which an attitude may show itself as a particular attitude (independently of any awareness of the attitude), namely in the consistency with which it is used. If I am merely guessing that a given chick is male, on repeated presentations of the same chick (without knowing that it is the same chick) I may respond "male" as often as "female". The fact that I say "male" 50% of the time is an expression of my complete uncertainty as to whether the chick is a male. However, if I was certain it was male, I should say "male" 100% of the time. Similarly, an attitude in between guessing and certainty may express itself by my saying "male" e.g. 60% of the time. In this case, I would be engaging in a type of probability matching (like animals and people do in many but not all situations, Reber 1989; Shanks, Tunney, & McCarthy 2002) – matching relative frequency of response to a subjective probability. My tendency to say "male" 60% of the time would reflect the extent of my uncertainty for the judgement: I am somewhat sure that the chick is male, but I am not certain. In sum, there may be situations in which the consistency with which a judgment is made indicates how well the system has normally learnt that the judgment accurately represents the world.

Note that if we are using consistency to measure attitude, we cannot tell from any one trial what the attitude is. We need an ensemble of trials. On any one trial, there will be a certain attitude, but we won't know the attitude from just one trial. Over a set of trials we can estimate the attitude. In order to use consistency as a measure of attitude, we need to assume representational stability over the time span we are investigating. For example, one could be highly certain that Fluffy is male one day (e.g. because we are told); and be highly certain that Fluffy is female the next day (because that is what we are told that day). In this case, certainty co-exists with inconsistency, but that's because the state of knowledge has changed from trial to trial. Only by presenting the ob-

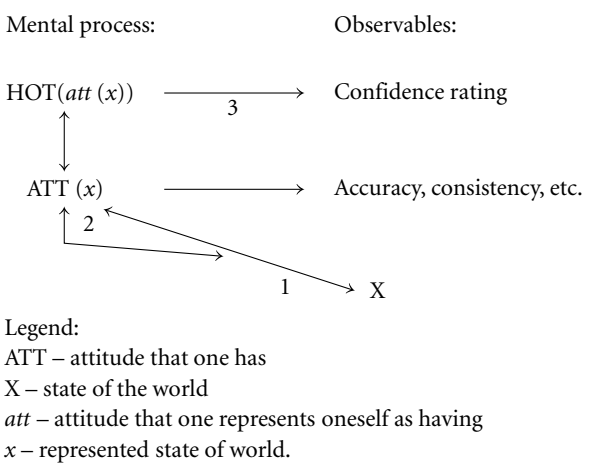

Figure 1. Relationship between the world, first order mental states, and higher order thoughts.

ject of the judgement repeatedly, while the person remains in the *same* state of knowledge, can consistency be used to measure attitude. We need to assume representational stability over the test phase.

To facilitate the arguments that follow, consider a state of affairs in the world, X. For example, X could be that "this chick is male". One forms a judgement (a mental state) about X with content *x*. For example, *x* could be "this chick is male". Call the attitude of the mental state ATT. So the full judgement can be written as ATT(*x*). For example, ATT(*x*) could be: I am fairly sure that "this chick is male", where ATT is fairly sure and *x* is "this chick is male". If the judgement is correct than *x* will correspond to X; if the judgement is incorrect, *x* will not correspond to X. This state of affairs is shown in Figure 1.

Consider a series of judgements, involving a number of different attitudes because I have different attitudes regarding the sex of different chicks; I may be certain about Donald, almost certain about Daffy, etc. Mapping (1) in the figure refers to the extent to which *x* corresponded to X over the series of test trials: What percentage of times over all judgments did I get it right? How often when I formed the judgement that a chick was a particular gender, was the chick that gender?

There is another relationship to consider (labelled 2 in the figure): Between the different attitudes and the percentage correct for each attitude. If the learning system is well adapted to the domain, then an attitude of certainty should be associated with a higher percentage of correct judgments than an attitude of

lesser certainty would be. For example, if I am certain about Donald and only fairly sure about Daisy, am I correct more often for Donald than for Daisy?

Finally consider when the person has a higher order thought (HOT) which represents the person as having a certain attitude (*att*). For example, the HOT could represent that "I am guessing that this chick is male". One might represent oneself as guessing (*att*) even though one's actual attitude on that trial was one of being fairly sure (ATT). Mapping 3 refers to the extent to which the attitude, *att*, one represents oneself as having corresponds to the attitude ATT that one actually has.

When using the zero-correlation criterion, we wish to assess mapping 3: Are people aware of the attitude of knowing when they know something and guessing when they are just guessing? However, neither one's actual attitude (ATT) nor the higher order thought about one's attitude (*att*) can be strictly directly observed by the experimenter. Instead, as shown on the right hand side of Figure 1, the experimenter determines the subject's higher order thought by the subject's confidence ratings, and the subject's attitude by how correct the subject is. In the latter case, the assumption is when the attitude involves more certainty, the subject will make more decisions correctly: the attitude of guessing should lead to chance performance and the attitude of knowing should lead to high performance. When is it valid to use these observables (confidence ratings and percent correct) to infer whether the quality of mapping 3 is better than zero? In other words, when can the zero-correlation criterion be used to infer the presence of conscious or unconscious mental states? That is the question we will consider in the rest of this paper.

Now we will use the mappings 1 to 3 illustrated in Figure 1 to consider the application of the zero-correlation criterion to implicit learning, and then to subliminal perception. In what follows we will need to consider a series of judgments. The zero-correlation criterion (and the guessing criterion) cannot be applied to a single judgment to determine if it was conscious or unconscious. It can only be applied to an ensemble of judgments to determine whether all of them were unconscious or whether at least some were conscious.

Applying the zero correlation criterion to implicit learning

Consider a subject learning an artificial grammar. After a training phase consisting of looking at grammatical strings, the subject is informed of the existence of a set of rules and asked to classify test strings. The subjects first order mental state for each judgment about a test string consists of two components:

The content of the judgment (*x* in Figure 1, e.g. "this string is grammatical") and the subject's attitude to that content (ATT). Now we ask the subject for their second order thought (their confidence rating) which is their report (*att*) of their attitude used in the first order mental state. In assessing the conscious or unconscious status of the first order state, we wish to establish whether the content of the higher order thought (*att*) represents the actual attitude (ATT) used in the first order state (when subjects know, do they know that they are knowing?). But there are in fact three mappings to consider:

First, there is the mapping (1) of the content of the first order mental state onto the world (specifically, in this case, the world is the grammaticality of each test string as determined by the experimenter's grammar): $x \leftrightarrow X$ (mapping 1 in Figure 1). Researchers into the content of learning (e.g., in implicit learning paradigms, Tunney & Altmann 1999; Cleeremans 1993; Dienes & Fahey 1995, 1998) are interested in investigating this mapping.

Consider a plot of the relationship between confidence and accuracy for a particular subject. Let us say it shows no relationship. Does this mean we have established the unconscious status of the subjects' knowledge? Not necessarily. Maybe the subject was mapping the attitude used in the first order state onto the content of their second order thoughts perfectly; but the content of the first order state was completely uncorrelated with the world. For example, the subject may have induced a grammar that would generate all training strings, but also half of the grammatical test strings and half of the non-grammatical test strings. On the basis of this grammar the subject would judge half the grammatical and half the non-grammatical test strings correctly the other half of each type of string incorrectly. As a result the subject would appear to be objectively guessing when in fact subjectively the subject was making 100% knowledgeable judgments. The experimenter had not come up with a psychologically plausible grammar; so the subject's highly evolved learning system presumed a different type of structure than the one the experimenter arbitrarily dreamt up. The learning system presupposed the "wrong" kind of model of the world; but under normal conditions it would have learnt to some degree successfully at this juncture. So the subject has a first order attitude of "knowing" or "thinking with 100% conviction" on some trials, and this is reflected completely in the confidence rating. Nonetheless, there would be no relation between confidence and accuracy because the first order content-world mapping is not reliable. The first order state is a conscious mental state, but the zero-correlation criterion would indicate it is unconscious.

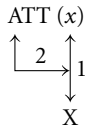

Figure 2. Mapping (2).

Second, there is the mapping (2) between the first order attitude (ATT), on the one hand, and the reliability of the first-order-content to world mapping (1), on the other, as shown simply in Figure 2.

Presumably, the first order attitude will normally map onto the reliability of the first-order-content to world mapping. Attitudes of greater certainty will generally be associated with contents that map the world reliably, when the learning system is adapted to the environment it is learning about. But of course this in no way guarantees that attitudes of certainty will always be towards contents that reliably map the world (even though they would under conditions to which the learning system is adapted). It does not even guarantee that in some restricted domain, variations in attitude correlate with variations in the reliability of content-world mapping. Researchers into metacognition are effectively interested in this mapping (e.g Herrmann, Grubs, Sigmundi, & Grueneich 1986; Reder & Ritter 1992; Gruneberg & Sykes 1993; Koriat 1993, 1997; Gigerenzer 2000): How does a person determine whether he or she has an appropriate attitude of knowing?

In an artificial grammar learning task a subject may have different attitudes towards the grammaticality of different test strings, and the subject may have accurate second order thoughts about those attitudes. But if the subject has as many percent correct choices for first order attitudes of high certainty as for attitudes of low certainty, there will be no correlation between confidence and accuracy despite appropriate higher order thoughts making all mental states conscious.[3]

Third, there is the first order attitude – second order content mapping, which is the one we are actually interested in to assess the conscious or unconscious status of the first order states, as shown simply in Figure 3.

Mapping (3) is also of interest to people working in metacognition (e.g. Reder & Ritter 1992; Koriat 1993, 1997; Gigerenzer 2000, as referenced for mapping (2)), who often are effectively interested in the joint effect of mappings (2) and (3), without distinguishing them.

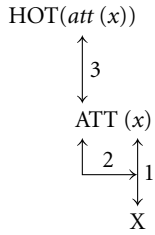

Figure 3. Mapping (3).

Our central question now is: How can we make sure that we are assessing mapping (3) with the zero correlation criterion, when mappings (1) and (2) also strongly affect the confidence-accuracy relationship?

First, we have to make sure the subject has a minimal (non-zero) first order content-world mapping over the set of test items as a whole before applying the criterion. We can determine this simply by looking at subjects' percent correct classification. There would be no point looking at the strength of the confidence-accuracy relationship unless subjects have demonstrably induced knowledge that correlates with the experimenter's grammar; i.e. their percent correct classification is above baseline. This shows mapping (1) over all items as a whole to be above chance.

Next, in order to establish mapping (2), we need some way of measuring attitude. Let us presume that consistency of response directly reflects attitude. Since Reber (1967), people in the artificial grammar learning literature have often tested subjects twice on the same test strings to measure consistency: The proportion of strings classified twice correctly (CC), once correctly and once in error (CE), and twice in error (EE). Note that having a reliable mapping (2) requires variability in attitude (as observably indexed by e.g. response consistency). Typically, CC, EE, and CE are all non-negligible (e.g. CC is about 0.60, EE about .15, and CE about .25: Reber 1989). This is important in showing variability in consistency. If subjects responded deterministically with a single rule that classified some items correctly and some incorrectly, being correct or incorrect to different items would not be related to different attitudes, because there may only be one attitude, i.e. of knowing (as pointed out by Dienes & Perner 1996). In this case, CE would be zero and mapping (2) would be zero or undefined.

Given the minimal requirement of CE being non-negligible, how are we to establish whether there is a positive relationship between attitude and percent correct (i.e. mapping (2))? To recap, it appears subjects have different attitudes

towards the judgments they make about the grammaticality of strings. For some strings, a subject may respond "grammatical" and "non-grammatical" about equally often, reflecting the attitude of guessing. For other strings, subjects respond "grammatical" with some probability above 0.5, indicating a stronger attitude, one of greater certainty that the string is grammatical, or with a probability below 0.5, indicating greater certainty that the string is non-grammatical. How can we determine whether stronger attitudes are associated with greater accuracy?

Answering this question depends on how the zero-correlation criterion is to be measured. For example, let us say the zero-correlation criterion is to be measured by comparing the average percent correct for high confidence responses with average per cent correct for low confidence responses (as used by Tunney and Shanks, in press). Consider four test items, and the subject is tested on each 100 times (with an ideal subject who never gets bored and has no memory of previous test trials). The subject gets item 1 correct 100 times, the subject gets item 2 correct 0 times, and he gets items 3 and 4 correct 80 times each. The overall percent correct is 65% (mapping 1 is fine). The subject has given 400 responses. If we take the 200 responses with the highest consistency (hence the stronger attitudes), this would be items 1 and 2, where the probability of emitting a given response is 1.0. The average percent correct for these two items is 50%. For the remaining 200 responses (i.e. to items 3 and 4), 80% of the responses are correct. Thus, average accuracy is higher for the weaker attitudes (80%) compared to the stronger attitudes (50%), as shown in Figure 4. Mapping (2) is negative. So even if subjects had completely accurate higher order thoughts (mapping 3), and hence completely conscious knowledge, the zero-correlation criterion would not show a positive relation between confidence and accuracy (as measured by: %correct given high confidence minus %correct given low confidence).

We will now consider another measure of the quality of mapping (2). Chan (1992) measured the zero-correlation criterion by subtracting the average confidence for all incorrect responses from the average confidence for all correct responses (The "Chan difference score"). We could use the Chan difference score, but apply it to actual attitudes (ATT) rather than confidence ratings (as estimates of our HOTs (*att*) about our attitudes) in order to investigate mapping (2). That is, we subtract the average strength of attitude for correct responses from the average strength of attitude for incorrect responses. In the above example, the average attitude strength is 0.88 for correct responses and 0.77 for incorrect responses.[4] The difference is positive, mapping (2) is satisfied. So if HOTs could directly represent attitude strength, the zero-correlation

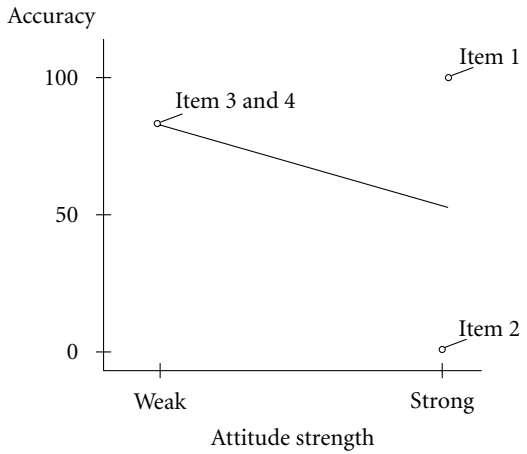

Figure 4. Example of mapping (2) being negative.

criterion would be positive (appropriately) by Chan's difference score applied to confidence ratings in this example.[5]

The Appendix considers in detail the measured strength of mapping (2) assuming different response models and measures of mapping (2) (including the Chan difference score). It concludes that the strength of mapping (2) is normally positive as measured by the Chan difference score, and inversely related to EE (Reber's notation for being consistently in error). Mapping (2) needs to be carefully considered depending on the way the zero-correlation criterion is to be measured. In general, the lower the value of EE, the better mapping (2), at least for the Chan difference score (and – so we speculate – more generally).

Having established positive mappings (1) and (2), the zero-correlation criterion can be investigated. A zero correlation provides evidence for unconscious mental states; a positive correlation indicates the presence of at least some conscious mental states (without ruling out the possibility of the existence of at least some unconscious mental states) (see also Dienes & Perner 2001, for one further proviso, namely that the confidence ratings should not arise from any inferences of which the subject is conscious).

Under most experimental conditions, subjects are likely to have at least some conscious knowledge. In this case, the zero-correlation criterion can still be used to compare the relative amount of conscious and unconscious knowledge across two conditions (compare Jacoby's process dissociation framework). First, one would establish the two conditions had the same percent correct classification (mapping (1) is the same) and the same EE value (mapping (2) is

the same). Under these conditions, differences in the confidence-accuracy relationship then indicate different mixes of conscious and unconscious knowledge across the two conditions. For example, Chan (1992) compared a group of subjects in training who just memorized strings with a group who searched for rules. At test they had the same (non-significantly different) percent correct classification - mapping (1) was the same. They also had the same average level of confidence. Reber (e.g. 1989) had previously shown that rule search compared to memorization instructions changes consistency by increasing EE; i.e. mapping (2) is worse for rule search than memorization subjects (see Appendix for justification of this logic).[6] Nonetheless, Chan found that the rule search group had a stronger confidence-accuracy correlation than the memorization group (according to the Chan difference score amongst other measures); this must have been due to a better mapping (3) for rule search than memorization. Hence there is evidence for a greater use of unconscious knowledge in the memorization group than the rule search group.

Applying the zero correlation criterion to subliminal perception

Now consider the application of these issues to subliminal perception. Kunimoto et al. (2001) found conditions under which subjects were just as accurate in identifying stimuli for high confidence as low confidence decisions. They concluded that this indicated the presence of unconscious perceptual states (after carefully considering, and then rejecting, possible psychometric artifactual explanations of the dissociation between confidence and accuracy). What mapping assumptions does their conclusion rely on?

The task of the subject in Kunimoto et al. (2001) experiments one and two was first to say what word was there and then to give a confidence rating. Perception presents a somewhat different situation from implicit learning. Arguably, in the first instance perception always involves the attitude of knowing towards whatever its content is. Even when we know what we see is an illusion, our vision still seems to present it to us as 100% fact. Nonetheless, the content of one's perception may not clearly be one of the contents the experimenter is forcing us to respond with (e.g. "the word is green" etc. is allowed by the experimenter, but "A funny bunch of lines" is not a response option). So one considers the content of the perception as evidence for the different words, and converts that into an attitude towards an allowable content e.g. being 60% sure that the word is green. The logic of the zero-correlation criterion runs like this: (i) The more (conscious or unconscious) perceptual evidence one has, the

greater one's accuracy; and (ii) the more conscious perceptual evidence one has, the greater one's confidence. So a positive relationship between confidence and accuracy indicates conscious perceptual evidence. Now let's consider the mapping assumptions, as shown in Figure 1, required for this logic to actually go through.

Mapping (1) is the same as in Figure 1: To what extent does the perceptual system produce contents that accurately represent the stimuli? If the perceptual system sometimes misrepresents one word as another, this content could produce confident inaccurate answers, reducing the confidence accuracy correlation, even when perception is quite conscious. Kunimoto et al. (2001) established that their subjects classified above chance, so mapping (1) is not problematic for their experiments.

There is also a version of mapping (2) to consider. The perceptual representation used to judge what word was presented will have various conceptual and perceptual contents (for example, that "the word is red", or how bright the word is, etc.), and other properties like fine-grainedness, time taken to form, or clarity, and so on. Mapping (2) is the extent to which those representational properties taken to indicate the extent to which the representation is accurate (vs misrepresenting) actually map onto to the extent to which the representation is accurate or misrepresenting.

We have probably evolved so that in normal conditions mapping (2) is very good; at least, we tend to think so: We usually feel we are pretty accurate in judging whether we have seen or failed to have seen something. But the system is not infallible; sometimes even in everyday life we have what seem to us to be clear percepts but in fact we have seen incorrectly. After all, the system cannot have a God's eye view of when it gets things right and when it gets things wrong. Mapping (2) might be compromised in two ways.

First, the first order discrimination and the assessment of representational accuracy may be based on at least some different representational properties. For example, a person may be conscious, somewhat dimly, of seeing the word green, and this perception may have certain qualities the person is also aware of. The first order discrimination is made on the basis of the relevant content provided by the perceptual system, namely that the word is green. Now how is the person to provide a confidence rating? They may rely on vividness to guide attitude even when it is irrelevant to discriminating what word was presented. Then different confidence ratings can be produced for the same level of first order accuracy. Subjects presumably use properties for assessing accuracy that are correlated with accuracy under normal conditions. But so long as the correlation is not perfect, restricting the range of variability of the properties (as

would be done in e.g. a subliminal perception experiment where accuracy is forced to vary over a limited range) could reduce the correlation to zero.[7]

Second, consider the case where the first order discrimination and the assessment of representational accuracy are based on exactly the same representational properties (e.g. the mechanism considered by Kunimoto et al. 2001, Appendix B, "the ideal observer"; Björkman, Juslin, & Winman 1993). Where there is a finite number of options (e.g. red vs green vs blue vs yellow), a simple decision procedure is to consider the activation of representations having each of the specified contents; whichever representation has the greatest activation is chosen, or is more likely to be chosen (first order discrimination) and the amount of differential activation can be used to determine the accuracy of the representation. Let's assume that mapping (1) is satisfied so that the first order discrimination is above chance. For mapping (2) to fail, the activation of e.g. the "red" representation must on average not be stronger when the word is "red" than when the red is not "red". Mapping (2) could fail in this way if there were inhibitory links between the representational choices (or, more generally, cleaning up processes) that force the system into four attractors: red, yellow, green, or blue. It will go into the right attractor more often than the wrong one (mapping (1) satisfied), but when it is in the wrong attractor, the final activation level of the chosen representation will on average be the same, regardless of whether it is right or wrong (mapping (2) fails). In this case, the inevitable variation of activation above and below the mean level for an attractor state would provide the basis for assigning high and low confidence ratings. The same idea could apply to any of the representations that could serve as evidence for which word was presented when there is no clear percept of a particular word being presented. When the level of perceptual signal is low, the cleaned up percept based purely on noise (or a different stimulus from that perceived as being there) may be just as good a percept as that produced by signal and noise. Treisman and Schmidt (1982) and Treisman and Souther (1986) found that with rapid presentation of visual stimuli, subjects often reported seeing stimuli that were illusory conjunctions of the features presented. Such illusory conjunctions were reported with high confidence and (based on subjects' verbal reports) had the character of perceptual experiences. In such an experiment, confidence and accuracy could fail to correlate even when the subject sincerely reports consciously seeing.

The postulated top-down cleaning up of percepts is consistent with the fact that people easily confuse imagination and real stimuli of weak intensity (e.g. Perky 1910, cited in Kelley & Rhodes 2002; Wickless & Kirsch 1989). Top down influences are likely to be particularly important the less time the subject has

to appraise the stimulus. Lewicki & Czyzewska (2001) found that about 30% of undergraduates had split-second hallucinations "very often", e.g. they might have the illusion of seeing an animal moving off the road, only to find out a moment later it was a piece of newspaper. A further 45% of students had this sort of experience "sometimes". In a subliminal perception experiment, thinking about or expecting a particular word, based on partial evidence provided by a different stimulus, might induce an active visual representation of that word just as good as that produced by the actual word itself. The activation of such errors could range to just as high a level as that of accurate representations of very impoverished stimuli. Mapping (2) could then be at chance, and the zero-correlation criterion would suggest perception had been subliminal when, in fact, whatever the subject saw (correctly or incorrectly) was seen quite consciously. The possible influence of expectation interfering with mapping (2) could be tested by relating individual differences in top down influences (e.g. as measured by Lewicki & Czyzewska 2001) to the confidence-accuracy correlation.

This is an issue that deserves further investigation. Maybe when subjects give the wrong answer it was because the first order visual information systematically pointed in the wrong direction, but the system had no way of knowing this, no way of calibrating mapping (2). In this case, a zero-confidence accuracy correlation would not indicate the presence of unconscious mental states. Or maybe people were simply not aware of the perceptual evidence used in making the discriminations, i.e. perception was unconscious. This will be a matter for future research to determine. One approach is to derive predicted dissociations that more plausibly derive from the difference between conscious and unconscious mental states rather than between mental states that are systematically wrong rather than systematically right. In general, any measure of the conscious status of mental states shows its usefulness in participating in such theory driven research (Merikle 1992). In fact, it is hard to see how there could be any theory independent measure of the conscious or unconscious status of mental states.

Conclusion

Higher order theory, like Rosenthal's (1986) higher order thought theory, provides a tool by which we can see clearly the relevance of subjective rather than objective measures of being consciously aware of the world, and also analyse the appropriate use of various subjective measures such as the zero correlation cri-

terion. Certain preconditions must be met before the zero correlation criterion can be applied in either subliminal perception or implicit learning paradigms as the confidence accuracy relationship depends not just on the strength of the mapping between the properties of first order mental states and the content of second order thoughts. Fortunately, one can often get a handle on the other relevant mappings, and hence plausibly use the zero-correlation criterion in implicit learning and perception research.

In this chapter we have considered the use of the zero correlation criterion where first order attitude is assessed by accuracy. Figure 1 indicates another possible measure of first order attitude, namely consistency. That is, another version of the zero correlation criterion – not considered in this chapter – is to measure the relationship between consistency and confidence ratings to determine if higher order thoughts (assessed by confidence ratings) are related to attitudes (assessed by consistency), and hence determine the conscious status of knowledge states. Lau (2002) showed that sometimes training on a dynamic control task led to dissociations between confidence and consistency. He argued that this reflected the use of implicit knowledge. But ultimately the worth of the zero correlation criterion, in whatever guise it is used, is shown by participating in theory driven research: The criterion should separate knowledge types that have the different properties predicted by a psychological theory of the functioning of conscious and unconscious knowledge.

We have assumed in this chapter that the process of verbally reporting a second order thought is not problematic. In fact, as well as the three mappings considered in this chapter, there is a fourth mapping to consider, that is more or less fallible, and that is the mapping between the second order thought and verbal report. Dienes (in press) considers this mapping in detail. These four mappings indicate four ways in which verbal report can fail: It could fail because the content of the first order state is incorrect about the world (mapping 1, so a verbal report would fail to describe the world correctly); or because the first order attitude is inappropriate for the degree of correctness of the first order state (mapping 2, so reports of confidence would fail to relate to accuracy); or because a second order thought misrepresents the first order state (mapping 3, so verbal reports would fail to reflect actual first-order attitudes held); or because the verbal reports misrepresent the second order thought (so verbal reports would fail to reflect conscious experience). This chapter, together with Dienes (in press), indicates how nonetheless verbal reports can be essential, if fallible, guides to determining whether a subject is consciously aware or merely aware of a stimulus or regularity.

Notes

1. Essentially, bias in the way higher order thoughts are formed from first order states does not lead to subjective measures being biased measures of the conscious status of the first order states; but bias in measuring a second order thought does lead to biased measurement of the conscious status of first order states.

2. We will assume an actualist higher order theory, like Rosenthal's, in which the first order state and the higher order state are different representations. Thus, the Carruthers (1992, 2000, this volume) potentialist theory will be inconsistent with some of the arguments that follow, because it postulates the first order state does not require a separate second order representation for the first order state to be conscious. While we will refer to higher order *thoughts*, consistent with Rosenthal's theory, regarding the higher order states as thoughts rather than perceptions (e.g., Armstrong 1980) is irrelevant for the arguments that follow.

3. In terms of Gigerenzer's (e.g. 2000) theory, if the subject had induced cues and represented cue validities from the training set, but these cue validities were uncorrelated with the "ecological validities" of those same cues in the test set, then confidence would be unrelated to accuracy regardless of whether the subject was conscious of their first order attitude (as determined by the cue validities, on Gigerenzer's theory).

4. The relevant calculations may be best understood after reading the Appendix. We assume that the proportion of times a stimulus elicits the correct response is the attitude towards the correct response; i.e. the attitude for the correct response is the proportion correct for that item. Conversely, the attitude towards the incorrect response is (1 – proportion correct). The calculations are: For correct responses, there are 100 responses of strength 1.0 from item 1, 80 of strength 0.8 from item 3 and another 80 of strength 0.8 from item 4. This weighted average is 0.88. For incorrect responses, there are 100 responses with strength 1.0 from item 2, 20 responses of strength 0.2 from item 3, and 20 responses of strength 0.2 from item 4, giving a weighted average of 0.77.

5. The following may be best understood after reading the Appendix: In the example in the text of mapping (2) being negative, different attitudes were collapsed together into one category. Specifically, items 3 and 4 are correct 80% of the time, and so those correct judgments have an attitude strength of .80; they are incorrect 20% of the time, and so those incorrect judgments have an attitude strength of .20. In Figure 2, an attitude strength of .80 for the correct responses was lumped together with an attitude strength of .20 for the incorrect responses, both attitudes being put in the "weak attitude" category. Maybe collapsing over attitude strengths in this way was the problem with this measure of mapping (2). Maybe the Chan difference score showed mapping (2) to be positive because it respects all the distinctions between attitudes strengths, making full use of the data. However, for reasons that are unclear, Tunney and Shanks (2003) found binary confidence ratings resulted in a more sensitive measure of the zero-correlation criterion than more fine-grained confidence ratings. As this is opposite to expected, we are running a study to explore the finding.

6. Another possibility is that subjects told to search for rules simply believed they should be extra consistent in responding to the same item, as compared to memorization subjects. However, in his 1989 review (p. 228), Reber found that the total amount of consistency (CC + EE) in rule search and memorization conditions is identical.

7. Whittlesea, Brooks, and Westcott (1994), present a situation where a categorization decision was based on different information than the confidence in the same decision. Subjects could be biased to base one decision (e.g. categorization) on typicality of individual features and the other (e.g. confidence) on exemplar similarity, or vice versa.

References

Allwood, C. M., P. A. Granhag, & H. Johansson (2000). Realism in confidence judgements of performance based on implicit learning. *European Journal of Cognitive Psychology, 12*, 165–188.

Armstrong, D. (1980). *The nature of mind and other essays*. Cornell University Press.

Bjőrkman, M., P. Juslin, & A. Winman (1993). Realism of confidence in sensory discrimination: The under-confidence phenomenon. *Perception & Psychophysics, 54*, 75–81.

Block, N. (2001). Paradox and cross purposes in recent work on consciousness. *Cognition, 79*, 197–219.

Carruthers, P. (1992). Consciousness and concepts. *Proceedings of the Aristotelian Society, Supplementary, Vol. LXVI,* 42–59.

Carruthers, P. (2000). *Phenomenal consciousness naturally*. Cambridge: Cambridge University Press.

Chan, C. (1992). Implicit cognitive processes: theoretical issues and applications in computer systems design. Unpublished D.Phil thesis, University of Oxford.

Channon, S., D. Shanks, T. Johnstone, K. Vakili, J. Chin, & E. Sinclair (2002). Is implicit learning spared in amnesia? Rule abstraction and item familiarity in artificial grammar learning. *Neuropsychologia, 40*, 2185–2197.

Cleeremans, A. (1993). *Mechanisms of Implicit Learning: Connectionist Models of Sequence Processing.* Cambridge, MA: MIT Press

Cleeremans, A., A. Destrebecqz., & M. Boyer (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences, 2*, 406 – 415.

Dienes, Z. (in press). Assumptions of subjective measures of unconscious mental states: Higher order thoughts and bias. *Journal of Consciousness Studies.*

Dienes, Z. & G. Altmann (1997). Transfer of implicit knowledge across domains? How implicit and how abstract? In D. Berry (Ed.), *How implicit is implicit learning?* (pp. 107–123). Oxford: Oxford University Press.

Dienes, Z., G. Altmann, L. Kwan, & A. Goode (1995) Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1322–1338.

Dienes, Z. & D. C. Berry (1997). Implicit learning: below the subjective threshold. *Psychonomic Bulletin and Review, 4*, 3–23.

Dienes, Z. & R. Fahey (1995). The role of specific instances in controlling a dynamic system. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 848–862.

Dienes, Z. & R. Fahey (1998) The role of implicit memory in controlling a dynamic system. *Quarterly Journal of Experimental Psychology, 51A*, 593–614.

Dienes, Z., A. Kurz, R. Bernhaupt, & J. Perner (1997). Application of implicit knowledge: deterministic or probabilistic? *Psychologica Belgica, 37*, 89–112.

Dienes, Z. & J. Perner (1999) A theory of implicit and explicit knowledge. *Behavioural and Brain Sciences, 22*, 735–755.

Dienes, Z. & J. Perner (2001). When knowledge is unconscious because of conscious knowledge and vice versa. *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society, 1–4 August, Edinburgh, Scotland* (pp. 255–260). Lawrence Erlbaum Associates: Mahwah, NJ.

Dienes, Z. & J. Perner (2002a). A theory of the implicit nature of implicit learning. In French R. M. & A. Cleeremans (Eds.), *Implicit Learning and Consciousness: An Empirical, Philosophical, and Computational Consensus in the Making?* (pp. 68–92). Psychology Press.

Dienes, Z. & J. Perner (2002b). The metacognitive implications of the implicit-explicit distinction. In Chambres, P., M. Izaute, & P.-J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 241–268). Dordrecht, Netherlands: Kluwer Academic Publishers.

Dienes, Z. & J. Perner (2003). Unifying consciousness with explicit knowledge. In Cleeremans, A. (Ed.), *The unity of consciousness: binding, integration, and dissociation* (pp. 214–232). Oxford University Press.

Dulany, D. E., R. Carlson, & G. Dewey (1984). A case of syntactical learning and judgement: How conscious and how abstract? *Journal of Experimental Psychology: General, 113,* 541–555.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.

Gruneberg M. M. & R. Sykes (1993) The generalizability of confidence-accuracy studies in eyewitnessing. *Memory, 1*, 185–190.

Herrmann, D. J., L. Grubs, R. Sigmundi, & R. Grueneich (1986). Awareness of memory ability before and after relevant memory experience. *Human Learning, 5*, 91–108.

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision, and visual masking: A survey and appraisal. *Behavioural and Brain Sciences, 9*, 1–66.

Kelley, C. M., & M. G. Rhodes (2002). Making sense and nonsense of experience: Attributions in memory and judgment. *The Psychology of Learning and Motivation, 41*, 293–320.

Kelley, S. W., A. M. Burton, T. Kato, & S. Akamatsu (2001). Incidental learning of real world regularities in Britain and Japan. *Psychological Science, 12*, 86–89.

Kolb, F. C. & J. Braun (1995). Blindsight in normal observers. *Nature, 377*, 336–338.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609–639.

Koriat, A. (1997). Monitoring one's knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126,* 349–370.

Kunimoto, C., J. Miller, & H. Pashler. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Consciousness and Cognition, 10,* 294–340.

Lau, K. K. (2002). Metacognitive measures of implicit learning in a dynamic control task. Unpublished D.Phil thesis, University of Sussex.

Lewicki, P., & M. Czyzewska (2001). Styles of nonconscious intelligence. In Lewicki, P. & A. Cleeremans (Ed.), *Proceedings of the AISB'01 Symposium on Nonconscious Intelligence: From Natural to Artificial, 21st–24th March 2001, University of York* (pp. 43–50). York, UK: University of York.

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist, 47*, 792–795.

Merikle, P. M. & M. Daneman (1998). Psychological investigations of unconscious perception. *Journal of Consciousness Studies, 5,* 5–18.

Millikan, R. G. (1984). *Language, thought, and other biological categories*. Cambridge, MA: MIT Press.

Morgan, M. J., A. J. S., Mason, & J. A. Solomon (1997). *Nature (London) 385*, 401–402.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing judgments. *Psychological Bulletin, 95,* 109–133.

Newell, B. R. & J. E. H. Bright (2002). Well past midnight: Calling time on implicit invariant learning? *European Journal of Cognitive psychology, 14,* 185–205.

Peirce, C. S. & J. Jastrow (1884). On small differences in sensation. *Memoires of the National Academy of Sciences, 3,* 73–83.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behaviour, 6,* 855–863.

Reber, A. S. (1989). Implicit learning and tactic knowledge. *Journal of Experimental Psychology: General, 118,* 219–235.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious.* New York: Oxford University Press.

Reder, L. M. & F. Ritter (1992) What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 435–451.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359.

Rosenthal, D. M. (2000).Consciousness, Content, and Metacognitive Judgments, *Consciousness and Cognition, 9,* 203–214.

Rosenthal, D. M. (2003). Consciousness and higher order thought. In *Encyclopedia of Cognitive Science* (pp. 717–726). Macmillan.

Shanks, D. R. & M. F. St. John (1994). Characteristics of dissociable human learning systems. *Behavioural and Brain Sciences, 17*, 367–448.

Siegal, S. & N. J. Castellan (1988). *Nonparametric statistics for the behavioural sciences*, 2nd edition. McGraw Hill: London.

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

Treisman, A. & J. Souther (1986). Illusory words: The roles of attention and of top-down constraints in conjoining letters to form words. *Journal of Experimental Psychology: Human Perception and Performance, 12*, 3–17.

Tunney, R. J. & G. T. M. Altmann (2001). Two modes of transfer in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 1322–1333.

Tunney, R. J. & G. T. M. Altmann (1999). The transfer effect in artificial grammar learning: Re-appraising the evidence on the transfer of sequential dependencies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1322–1333.

Tunney, R. J. & D. R. Shanks (2003). Subjective measures of awareness and implicit cognitition. *Memory & Cognition, 31*, 1060–1071.

Twyman, M. (2001). Metacognitive measures of implicit knowledge. Unpublished D.Phil thesis, University of Sussex.

Weiskrantz, L. (1988). Some contributions of neuropsychology of vision and memory to the problem of consciousness. In A. J. Marcel & E. Bisiach (Eds.), *Consciousness in contemporary science* (pp. 183–199). Oxford: Clarendon Press.

Weiskrantz, L. (1997). *Consciousness lost and found*. Oxford University Press.

Whittlesea, B. W. A., L. R. Brooks, & C. Westcott (1994). After the learning is over: Factors controlling the selective application of general and particular knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 259–274.

Wickless, C. & I. Kirsch (1989). The effect of verbal and experiential expectancy manipulations on hypnotic susceptibility. *Journal of Personality and Social Psychology, 57*, 762–768.

Appendix 1

Factors influencing mapping (2)

The aim of this appendix is to consider factors affecting the strength of mapping (2), the relation between accuracy and attitude strength. We will consider two measures of mapping (2), namely, Chan's difference score and the Goodman-Kruskal Gamma statistic. This is for purely illustrative reasons; the Chan difference score has often been used as a measure of the confidence-accuracy relationship in previous studies using the logic of the zero-correlation criterion, reviewed in the chapter; and Gamma, though rarely used in the implicit learning literature (for an exception, see Twyman 2001), is the statistic of choice in the metacognition literature (Nelson 1984).

As a measure of the attitude-accuracy relationship, the Chan difference score is the average attitude strength when a correct decision has been made minus the average attitude strength when an incorrect decision has been made. The higher score, the more positive the relation between attitude strength and accuracy. Gamma is a measure of association between two ordinally-scaled variables, each with two or more values (see Siegal & Castellan 1988: 291–298).

Consider a set of strings, and towards each string the subject has a certain attitude that the string is grammatical or non-grammatical. We will assume the attitude is reflected perfectly in the consistency with which a subject gives a

response. For example, if a subject responds "grammatical" to the string 80% of the time and "non-grammatical" 20% of the time, then the subject has an attitude strength of 0.8 that the string is grammatical. We can equivalently say the subject has an attitude strength of 0.2 that the string is non-grammatical. We will first consider the Chan difference score, and then Gamma.

Let us say the accuracy over all the strings is 60%. Such an overall score could come about in various ways. For example, one way is by the subject having an attitude strength of 0.60 towards the correct response for each string individually. Thus, if there were 10 strings in total, we would expect there to be six correct decisions and four incorrect ones. For the correct decisions, the attitude strength would be 0.60 (the response is emitted with a probability of 0.60 for that item). For the incorrect decisions the attitude strength would be 0.40 (the response is emitted with a probability of 0.40 for that item). So the Chan difference score would be 0.60 – 0.40 = 0.20. On this model, if the overall percent correct is labelled PC, then the probability of giving the right response for each item comes from a set with only one member, {PC}, e.g. {0.6}.

On another model, the subject gives each response with complete certainty, i.e. there is deterministic responding. Thus, the probability of giving the right response to each item comes from the set {0, 1}. For 10 items and an overall percent correct of 0.6, the subject would respond correctly to six items with an attitude strength of 1.0, and incorrectly to four items with an attitude strength of 1.0. The Chan difference score would be zero.

According to Reber (1989), subjects either guess the response to an item randomly, or they know the response to that item and respond perfectly consistently. Thus, on the Reber model, the probability of giving the right response to each item comes from the set {0.5, 1}. For example, for 10 items, if the subject guessed randomly for 8 of them, one would expect four to be correct (with an attitude strength of 0.5) and four incorrect (with an attitude strength of 0.5). By knowing the response to the other two items (attitude strength of 1.0), there would be six items correct in total out of 10. The average attitude strength for the correct items would be (4*0.5 + 1+ 1)/6 = 0.67. The average attitude strength for the incorrect items is 0.5. The Chan difference score is 0.17.

Finally, one can consider models with the probability correct for each item drawn from some other set. We consider all the models described above, plus a model in which the probabilities are drawn from the set {0.1, 0.9}.

Figure 5 below shows the Chan difference score for mapping 2 for different probabilities correct. There are four curves, each created by assuming the probability, p, of a correct response to any item could only be selected from a given specified set. For example, on one extreme, the probability of correct response

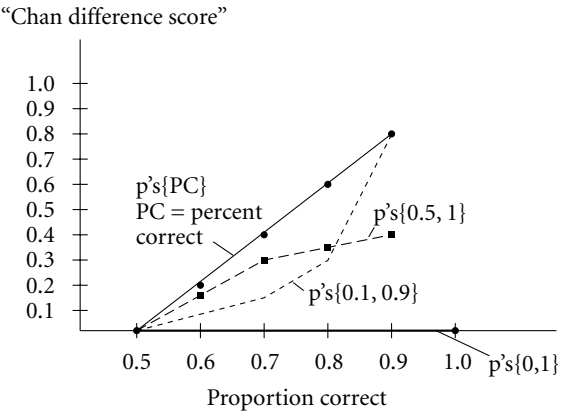

Figure 5. The Chan difference score measure of mapping (2) for different distributions of attitudes over items.

to each item was just the overall percent correct over all items. This is the top line. Here is no variance in p's between items for a given percent correct. At the other extreme, the probability of correct response to each item was either 1 or 0, there is maximum variance between the p's across items. This is the bottom thick line.

Note that the Chan difference score is not defined for an overall proportion correct of 1.0 because there are no incorrect answers. For a proportion correct of 0.9, the {0.1, 0.9} model becomes a degenerate case, because all strings must have a proportion of being correct of 0.9. Thus, the {0.9, 0.1} model is the same as the {PC} model in this case.

For the bottom curve, the Chan difference score is always zero. Mapping 2 is zero and so the assumptions of the zero-correlation criterion are not satisfied. (This is the case mentioned in the text, page *15*, where EC=0.) In no case is the Chan difference score negative. Mapping (2) is always positive or else zero. Is there any indicator of the strength of the mapping (2)? Consider a test where subjects are tested on each item twice. Figure 6 below shows how the proportion of times an item would be classified twice in error (EE) for each model and percentage correct.

For a given percent correct, the ranking of the Chan difference score, from highest to lowest, is always the same as the ranking of EE scores from lowest to highest. In other words, for a given percent correct, the lower the EE proportion, the better is mapping (2), as assessed by the Chan difference score. This is useful in comparing two conditions with the same percent correct; if

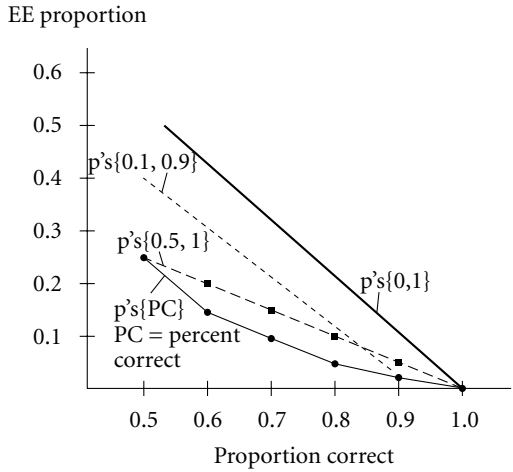

Figure 6. A measure of proportion of items classified in error twice in a row (EE) for different distributions of attitudes over items.

Table 1. Data for calculating gamma in the 2X2 case; the entries in the table are counts

	attitude 1	attitude 2
correct	a	b
incorrect	c	d

one has a higher EE than the other, it is reasonable to conclude mapping (2) is worse. This fact is used in the text to draw conclusions about the relative amount of unconscious knowledge in two conditions, as explained in the text (pp. 184–185).

Now we will consider another measure of mapping 2, Gamma. All the models we are considering here generate just two attitudes (e.g. the {0.5, 1.0} model involves just attitude strengths of either 0.5 or 1.0), so the relevant data for calculating Gamma can be represented by a 2 X 2 table, where a, b, c, and d are counts, as shown in Table 1.

The formula for Gamma reduces to (ad–bc)/(ad+bc) for the 2 X 2 case. Gamma varies between +1 and –1, where 0 indicates no association.

Nelson (1984) argued that Gamma had various properties desirable in a measure of metacognition; for example, if a subject has perfect metaknowledge this can be reflected in a Gamma of 1, regardless of the overall level of performance (whereas the Chan difference score depends on this level); it does not assume an interval scale for either variable (in contrast, the Chan difference score assumes an interval scale); and it is not sensitive to ties (in the general

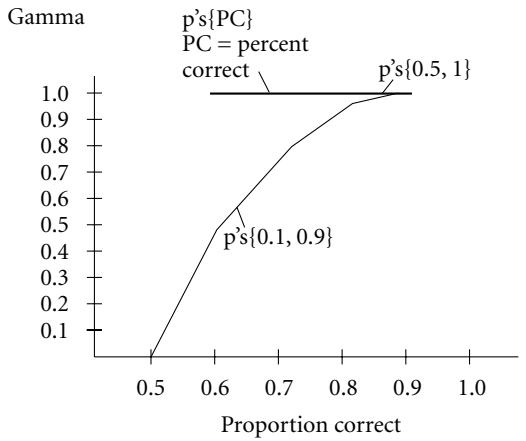

Figure 7. The Gamma measure of mapping 2 for different distributions of attitudes over items.

n X n case, in contrast to other nonparametric measures of associations, like Kendall's Tau).

Figure 7 shows the Gamma for each model considered above. Note that for the {0,1} model, Gamma is undefined or zero, as there is only one attitude. The other models are displayed on the figure. For the {PC} and {1, 0.5} models, Gamma is uniformly one (except for a percent correct of 0.5 and 1, where there is only one attitude). This is unlike the Chan difference score, which increased with percent correct for these models. For the {0.1, 0.9} model, Gamma does increase with percent correct, like the Chan difference score. Dienes, Kurz, Bernhaupt, and Perner (1997) argued that the pattern of subjects' responding over repeated testing of the same items was most consistent with a model in which each subject responded with a range of probabilities correct for different items (and which are different for different subjects), not just {1, 0.5, PC}; that is, one might expect in general the underlying Gamma measure of attitude-accuracy association to increase with percent correct in the artificial grammar learning paradigm.

Where the Gamma measure of mapping (2) does vary for the same percent correct between different models, the model with the higher EE has the lower measured strength of mapping (2) (as was the case for the Chan difference score).

In summary, we have considered two measures of the strength of mapping (2), the Chan difference score and Gamma. Where the subject's first-order attitude is not certainty for every item, and overall classification performance is be-

tween 0.5 and 1.0, both produce positive measures of mapping (2) for the models considered. For both measures, EE can be used as an index of the strength of mapping (2) for the same overall classification performances. Gamma has some nicer properties than the Chan difference score, but the Chan difference score still produces interpretable results (i.e. for the same percent correct, the measured strength of mapping (2) varies inversely with EE, a property used in the text, pp. 184–185).

If mapping (3) was perfect, the Chan difference score and Gamma calculated on confidence ratings, rather than first-order attitudes, would vary according to overall percent correct in the way that mapping (2) expresses itself in these measures, as explored in this appendix.

Finally, we note that subjects' apparent use of probability matching in expressing first-order attitudes is essentially irrational. If I believe a string is grammatical with 80% certainty, rationally I should say "grammatical" on every occasion. But subjects do not; sometimes they say "non-grammatical", an answer about which they are fairly certain is wrong ("fairly certain" in the sense of a first-order attitude). Presumably, on those trials it does not seem to the subject that they believe they are wrong, otherwise they would not give that answer; that is, presumably their HOTs do not match their attitudes. If subjects were perfectly aware of their attitudes, either they would sometimes give a response about which they have a confidence of less than 50%, or more likely, they would not probability match at all. If subjects have partial awareness of their attitudes, their HOTs may always lie in the range 50–100 even while the subject probability matches, but nonetheless, stronger attitudes may be associated with higher confidence ratings on average. If subjects have no awareness of their attitudes (the knowledge is unconscious), confidence ratings will bear no relation to attitudes.

Part II

Critics of the higher-order approach

Chapter 9

What phenomenal consciousness is like*

Alex Byrne

> . . .the phrase 'what it's like' is now worse than useless
>
> W. G. Lycan[1]

> Baby You don't know What It's like to be me
>
> Britney Spears

The terminology surrounding the dispute between higher-order and first-order theories of consciousness is piled so high that it sometimes obscures the view. When the debris is cleared away, there is a real prospect – to borrow a memorable phrase from Mark Johnston – that this is one of those genial areas of enquiry in which the main competing positions are each in their own way perfectly true.[2]

This paper has the following three goals, in ascending order of ambition. First, to show that there is *a* genuine dispute in the vicinity. Second, to rebut arguments against first-order theories based on a distinction (to be explained shortly) between "experiences" and "conscious experiences". Third, to show that higher-order theories are mistaken. Even if none of these goals is met, at least the paper can serve as a terminological guide for the perplexed.

The dispute (as understood here) concerns theories of *phenomenal* consciousness – Section 1 is a quick refresher on that topic. Section 2 describes the competing positions, and the crucial distinction between experiences and conscious experiences. Section 3 recounts the crucial distinction as explained by, respectively, Carruthers, Lycan, and Rosenthal. These three higher-order theorists use the distinction in somewhat different ways to argue against first-order theories; Section 4 examines these arguments. Section 5 presents a brief argument against higher-order theories.

1. Phenomenal consciousness

Here is how Block introduces the notion of phenomenal consciousness:

> P-consciousness [phenomenal consciousness] is experience. P-conscious properties are experiential properties. P-conscious states are experiential, that is, a state is P-conscious if it has experiential properties. The totality of the experiential properties of a state are "what it is like" to have it. Moving from synonyms to examples, we have P-conscious states when we see, hear, smell, taste and have pains. (Block 1995: 230)

As a simplification, this paper restricts attention to *perceptual* experiences (hereafter simply called 'experiences'), which will be taken to be *events*. The issue of which *states* are phenomenally conscious will accordingly be sidestepped, as will the issue of whether other events – e.g. episodes of thinking – are phenomenally conscious.

For what follows, an example of a Blockean "experiential property" will be useful. Consider visual experiences (in good light) of cucumbers, limes, green peppers, grass, and so on. These experiences saliently resemble each other in respect of "what it's like" to undergo them; they accordingly share an experiential property, call it 'G'. So-called spectrum inversion is – at least – a case where experiential properties are systematically permuted. When an inverted subject looks at a tomato, his experience has G. Sometimes a property like G is called a 'phenomenal character', a 'qualitative character', or a 'quale'. The last two expressions are used in this way in Levine 2001; in the additional terminology of that book, G is "greenishness".

2. Three distinctions: Intentionalism vs. phenomenism; FO vs. HO; experiences vs. conscious experiences

There is a pair of distinctions that mark important differences between accounts of phenomenal consciousness. The first distinction is between *intentionalism* (or representationalism), and *phenomenism*. According to intentionalism, phenomenal consciousness is entirely intentional or representational. Less imprecisely, and restricting attention to perceptual experiences, intentionalism implies that facts about the representational content of an experience (together with facts about the representational content of the subject's other mental events or states[3]) fix or determine the facts about its phenomenal character. In other words, intentionalism implies that phenomenal character supervenes

on representational content. Phenomenism rejects this supervenience thesis. One classic argument against intentionalism is based on an inverted spectrum thought experiment which is claimed to be a case of same representational content, yet different phenomenal character (Shoemaker 1981; Block 1990).

The second distinction is our topic: *first-order* (FO) vs. *higher-order* (HO) theories of consciousness. According to FO theories, an event may be phenomenally conscious even though it is not represented by one of the subject's mental states/events. In other words, higher-order representations are not *necessary* for phenomenal consciousness. In particular, we may take the first-order theorist to hold that a certain first-order condition – to be explained in a moment – is *sufficient* for having phenomenal character G. Take a phenomenally conscious perceptual experience e of a cucumber in daylight. If e is in fact represented by one of the subject's mental states/events, then consider e*, an event as similar to e as possible, except that e* is not represented by one of the subject's mental states/events. (Hence, if e is not in fact the target of a higher-order representation, then e*=e.) Repeat this process for other phenomenally conscious experiences of cucumbers, limes, green peppers,. . ., resulting in other (possible) experiences e**, e***,. . . Then the first-order theorist's sufficient condition for G is this: being mentally exactly like e* (or like e**, or like e***,. . .).

According to HO theories, on the other hand, higher-order representations *are* necessary for phenomenal consciousness. On this view, e* is *not* phenomenally conscious (and hence does not have G). That is, the higher-order theorist claims that there can be events that are mentally exactly like phenomenally conscious experiences, except that they are not targeted by higher-order representations, and thus are not phenomenally conscious. (The higher-order theorist also gives a sufficient condition; this will be explained in a moment.)

We now face a terminological decision. Suppose – for the rest of this paragraph – that the higher-order theorist is correct, and that e* is not phenomenally conscious. e* is not e, but presumably it's pretty much like it. Like e, e* represents the subject's environment as containing an elongated green thing; like e, e* causes the subject to believe that an elongated green thing is before her, and so forth. e*'s similarity to e counts in favor of labeling it a perceptual experience. On the other hand, given that phenomenal consciousness is often equated with experience (cf. Block's slogan "P-consciousness is experience"), calling e* 'an experience' has something to be said against it. Since one of our chief protagonists – Carruthers – comes down on the side of using 'experience' broadly, to include events like e*, it will be convenient to follow his usage here. (See Carruthers 2000: 13–14, 18–20.) So e and e* are both experiences; the for-

mer is conscious and the latter isn't. Block's slogan, recast in this terminology, is 'P-consciousness is conscious experience'.

These two distinctions – intentionalism vs. phenomenism, FO theories vs. HO theories – are independent. FO theory may be combined with intentionalism or phenomenism; likewise for HO theory. For example, Block and Levine are FO phenomenists; Dretske (1995) and Tye (1995) are FO intentionalists; and Carruthers and Lycan are HO intentionalists – albeit with a qualification in Lycan's case.[4] Rosenthal's position has some hidden complications. This will be explained more fully later (Section 3.3); for now the following enigmatic remark must suffice: Rosenthal is an HO intentionalist, but if he *were* to be converted to FO theory, he would be a phenomenist.[5]

To avoid getting bogged down with too many qualifications, it will be helpful to assume (in agreement with Carruthers, Lycan, and Rosenthal) that intentionalism is correct. Nothing of substance will turn on this assumption. Thus (in the usual terminology) the dispute concerns first-order representational (FOR) and higher-order representational (HOR) theories of phenomenal consciousness.

We may think of the FOR theorist as maintaining that the following condition is sufficient for phenomenal consciousness (in particular, sufficient for G): being exactly like e^* (or e^{**}, or e^{***},...) in all representational or intentional respects. This unwieldy formulation will abbreviated as follows: the FOR theorist thinks that *being an experience of green* is sufficient for G.[6] The HOR theorist disagrees, holding that some higher-order representation is necessary.

What about the HOR theorist's sufficient condition for G? This will depend on the type of HOR theory. HOR theories come in two main flavors, depending on whether the relevant higher-order representations are (a) thoughts or (b) perceptions (or, at any rate, importantly similar to perceptions and importantly unlike thoughts). Rosenthal holds the HOT (higher-order thought) version;[7] Carruthers and Lycan hold the HOP (higher-order perception) version.[8]

Take the HOT version first. Since the HOR theorist is an intentionalist, she holds that there is an intentional sufficient condition for having G. Since she is a HOT theorist, she holds that this intentional sufficient condition involves the presence of a thought about the first-order experience: in particular the higher-order thought whose content is that one is having an experience of green.[9]

Now take the HOP version. The HOP theorist holds that the intentional sufficient condition for G involves the presence of a (quasi-) perceptual experience directed on the first-order experience: in particular the higher-order perceptual experience whose content is that one is having an experience of green.

So, we have three candidate sufficient conditions for G, as follows:

FOR: being an experience of green
HOT: being an experience of green, accompanied by the thought that one is having an experience of green.
HOP: being an experience of green, accompanied by an experience that one is having an experience of green.[10,11]

FOR theories and HOR theories have been explained here in their weakest versions. The FOR theorist gives a sufficient condition for having G, and accordingly denies that a higher-order representation is necessary. Against this, the HOR theorist claims that a higher-order representation is necessary, and that the FOR theorist's sufficient condition must be strengthened by the presence of a certain higher-order representation. Neither theorist need offer necessary *and* sufficient conditions for phenomenal consciousness, or the having of particular phenomenal characters. Further, neither theorist need offer naturalistic, physicalistic, or reductive conditions; or conditions that explain phenomenal consciousness.

As it happens, FOR theorists like Dretske and Tye, and HOR theorists like Carruthers, Lycan and Rosenthal, are more ambitious. They offer necessary and sufficient conditions (or implicitly give recipes for specifying such conditions). Second, these conditions are broadly physicalistic (at least, suggestions are made for how the conditions as officially stated could be given a physicalist reduction). Third, and connectedly, these conditions are intended to give some sort of explanation of phenomenal consciousness.

According to the FOR theorist, necessarily any experience of green is (phenomenally) conscious; according to the HOR theorist, there could be experiences of green that are not conscious. Of course, the FOR theorist will add other sufficient conditions for phenomenal consciousness (in our abbreviated formulation: being an experience of blue, being an experience of triangularity, and so on); again, the HOR theorist will claim that there could be such experiences that are not conscious. It is important to note that the FOR theorist is not committed to denying a certain consequence of the HOR theorist's view, namely that there could be non-conscious experiences. For example, the FOR theorist could grant that we frequently have perceptual experiences of the disposition of our limbs (proprioceptive experiences) that are not phenomenally conscious.[12]

The distinction that figures prominently in certain arguments against FOR theories offered by Carruthers, Lycan, and Rosenthal is that between experiences and conscious experiences. They draw the distinction in somewhat dif-

ferent ways, and their arguments based on it are quite different – as will be explained later, Lycan's "argument" is, by his own lights, just a relatively uncontroversial observation. To complicate the picture further, they each coin their own terminology: "worldly subjectivity" and "experiential subjectivity" (Carruthers); "lower-order" and "higher-order" readings of 'what it's like' (Lycan); "thin phenomenality" and "thick phenomenality" (Rosenthal).

The first order of business is to show that these three distinctions are basically the same. Along the way, important differences between Carruthers, Lycan, and Rosenthal will emerge. The next item (Section 4) is to examine why the distinction is supposed to be real, as opposed to merely notional, and the subsequent arguments against FOR theories.

3. Experiences vs. conscious experiences: Three accounts of this distinction

3.1 Carruthers: Worldly vs. experiential subjectivity

Carruthers' distinction between *worldly* subjectivity (what the *world* is like for the subject) and *experiential* subjectivity (what the *subject's experience* is like for the subject) is one between properties:

> . . .what the world (or state of the organism's own body) is like for an organism. . .is a property of the world (or of a world-perceiver pair, perhaps).

And:

> . . .what the organism's *experience of the world* (or of its own body) is like for the organism. . .is a property of the organism's experience of the world (or of an experience-experience pair) (2000: 127–128)

Explaining this distinction in an earlier paper, Carruthers asks which of these two sorts of subjectivity "deserves the title 'phenomenal consciousness' ", and replies that he is "happy whichever reply is given" (1998: 209). In his later book, the question gets a more opinionated answer: "The subjectivity of experience, surely" (2000: 129, Note 7). Since 'Phenomenal Consciousness' is the title of Carruthers' book, we may safely presume that Carruthers takes himself to be using the term in the same way as Block – its inventor. It is also clear that Carruthers takes the FOR theorists Dretske and Tye to be offering accounts of phenomenal consciousness, as Carruthers understands the term. (These points might seem too trivial to mention. But as will transpire shortly, not every HO theorist agrees with Carruthers on these exegetical issues.)

The distinction between worldly and experiential subjectivity may be further clarified as follows. Suppose an organism has color vision. Then what the *world* is like for the organism is, inter alia, colored. In other words, the organism's perceptual experience *represents* the world as colored. So we may say that a *worldly subjective property* (relative to this organism) is simply a property *represented by* the organism's perceptual experience: green, for example. This is a property of objects in the organism's environment (at least if its color experiences are veridical). *Experientially subjective properties*, on the other hand, are just Blockean experiential properties, like our old friend G. (See Carruthers 2000: 13.) G is a *property of* experiences, and hence is not the property green – worldly subjective properties are plainly different from experientially subjective properties.

An experience of green that there is nothing it's like for the subject to undergo is said by Carruthers to "possess worldly subjectivity but [to] lack experiential subjectivity" (2000: 147); an experience of green that there is something it's like for the subject to undergo possesses both forms of subjectivity. So the distinction between experiences and conscious experiences (as explained in the previous section) is the same as the distinction between experiences that only have worldly subjectivity, and those that have both worldly and experiential subjectivity.

3.2 Lycan: What it's like – lower-order vs. higher-order

According to Lycan (this volume), a distinction that he has made in a number of places is the same as, or at any rate very similar to, Carruthers' distinction. Lycan's distinction, he explains, corresponds to an "ambiguity" in philosophers' talk of "what it's like" (1996: 77).

Lycan claims that on one ("lower-order") reading of 'what it's like', the phrase picks out what Lycan – following C. I. Lewis – calls a *quale*.[13] A quale is an "introspectible monadic qualitative property of what seems to be a phenomenal individual, such as the color of what Russell called a visual sense datum" (1996: 69). Lycan holds a "representational theory of qualia", which basically amounts to the view that qualia are properties that visual experiences represent objects as having, and which in some cases may not in fact be properties of such objects. One may seem to see a green thing, according to Lycan, even though there is nothing green to be seen – not even a sense datum. (We may take this view to be common ground among all our protagonists.) Notice that Lycan's qualia are the same as the worldly subjective properties of the previous section.

According to Lycan, on the lower-order reading of 'what it's like', there is something it's like to experience green and, in general, something it's like to undergo a perceptual experience (Lycan 1999a). So talk of what it's like in Lycan's lower-order sense is straightforwardly translatable by Carruthers' terminology of 'worldly subjectivity'.

On the other ("higher-order") reading of 'what it's like', Lycan claims that this expression picks out "a higher-order aspect of the quale, namely, what it's like *to experience* that quale" (1999b: 128; see also Lycan, this volume); so talk of what it's like in Lycan's higher-order sense is straightforwardly translatable by Carruthers' terminology of 'experiential subjectivity'. Lycan takes these higher-order aspects or properties to be properties *of* qualia (i.e. properties of colors, shapes, odors, etc.). The higher-order "what it's like" properties are accordingly not the same as the experientially subjective properties of the previous section (which are properties of experiences), but there is a simple correspondence between them. Let g be the Lycanian higher-order "what it's like" aspect of "green qualia". Then experience e has G iff it is an experience of a quale that has g.

Since the expressions 'qualitative character', 'phenomenal character', and 'phenomenal consciousness' are often explained using 'what it's like', it is no surprise that Lycan finds an ambiguity here too (Lycan 1999a). This allows us to avoid clumsy locutions such as 'higher-order "what it's like" properties', and to rephrase Lycan's distinction in Lycanian terminology thus: lower-order qualitative character vs. higher-order qualitative character.

Lycan himself prefers to use 'what it's like' with its higher-order reading (this volume). He avoids 'phenomenal consciousness', but if forced would presumably have a similar preference. So in this respect we may take him to be in terminological agreement with Carruthers. However, unlike Carruthers, Lycan thinks Block's use of 'phenomenal consciousness' picks out qualitative character of the *lower*-order kind (Lycan 1999a: Note 1). Further, Lycan takes FOR theorists like Dretske and Tye to be merely offering representationalist theories of lower-order qualitative character (1999a; this volume). That is, according to Lycan, Dretske and Tye are simply giving representationalist accounts of *perceptual experience* – for example, "perceiving a [green] object as such", or (in the 'qualia' terminology), "registering" a "green quale" (Lycan 1996: 76).

Carruthers thinks that the dispute between Block, Dretske and Tye, on the one hand, and Carruthers, Lycan and Rosenthal, on the other, is genuine; *Lycan*, however, thinks it is largely terminological. According to Lycan, when Dretske, for example, inveighs against HOR theories, he is correctly claiming that higher-order representations are not needed in accounts of *lower*-order qualitative character. And when Carruthers touts the virtues of HOR theories,

he is correctly claiming that such representations are needed in accounts of *higher*-order qualitative character. For Lycan, the philosophical action is to be found further down the road, in the dispute between HOP and HOT.

3.3 Rosenthal: Thin and thick phenomenality

Rosenthal makes a distinction between two kinds of phenomenal consciousness, or, in Block's (2001) currently preferred terminology, "phenomenality":

> One kind consists in the subjective occurrence of mental qualities, while the other kind consists just in the occurrence of qualitative character without there also being anything it's like for one to have that qualitative character. Let's call the first kind *thick phenomenality* and the second *thin phenomenality*. Thick phenomenality is just thin phenomenality together with there being something it's like for one to have that thin phenomenality. (2002a: 657)

In order to explain what "the occurrence of qualitative character" is supposed to be, more details of Rosenthal's view are needed. Rosenthal thinks that experiences have introspectible properties that he elsewhere calls *sensory qualities*, or *mental qualities* (Rosenthal 1999). One such mental quality is characteristic of perceptions as of green objects, and Rosenthal calls this quality *mental green*. An "occurrence of qualitative character" is an experience that has some sensory or mental quality.

Mental green is the same as the Lycanian property of being a registering of a green quale and, as noted, Lycan identifies this with the intentional property of being an experience of green.[14] But Rosenthal would refuse to make this identification, because he thinks that mental green is a *non*-intentional property; in this, he is in agreement with Block and Levine. ('Mental green' is used here is a neutral way, so that it is an open question whether mental green is non-intentional.) If Rosenthal converted to FO theories, he would think that mental green, and not any intentional property, was sufficient for G. Section 2 claimed that if Rosenthal were an FO theorist, he would be a phenomenist, and this is the reason why.

Solely for convenience, let us assume that Rosenthal is wrong about mental green, in which case it may be identified with the property of being an experience of green. Hence, an experience of green is an example of the "occurrence of qualitative character".

Since thin phenomenality is "the occurrence of qualitative character *without* there also being anything it is like for one to have that qualitative character", it appears that thin phenomenality is the same as what Section 2 called

'non-conscious experience'. However, the last sentence of the quotation implies that thin phenomenality does *not* preclude there being something it's like (see also the quotation below). It is clear that Rosenthal's intent is best served by ignoring the stipulation that there is nothing it is like to undergo thin phenomenality. Thin phenomenality, then, will be taken simply to be "the occurrence of qualitative character", in which case it may be identified with worldly subjectivity and lower-order qualitative character.[15]

Rosenthal's explanation of thick phenomenality is in essence the same as Lycan's explanation of the higher-order reading of 'what it's like'; so thick phenomenality may be identified with higher-order qualitative character, and experiential subjectivity.

According to Rosenthal, "Block describes phenomenality in different ways that arguably pick out distinct types of mental occurrence" (655), these "distinct types" being thin and thick phenomenality. But Rosenthal thinks that, charitably interpreted, Block has *thin* phenomenality in mind:

> If we bracket the issue about how to understand the admittedly vexed phrase 'what it's like', Block's view seems to be that phenomenality is simply thin phenomenality, and what I'm calling thick phenomenality is phenomenality plus reflexivity [i.e. plus an appropriate higher-order representation]...
>
> Terminology aside, this fits neatly with my own view of things.
>
> (2002a: 657; see also Rosenthal 1997a)

Rosenthal, then, shows some sympathy with Lycan's ecumenicism. Like Lycan, and unlike Carruthers, Rosenthal takes Block's use of 'phenomenality' and 'phenomenal consciousness' – at any rate when charitably interpreted – to refer to *thin* phenomenality (i.e. worldly subjectivity or lower-order qualitative character). Thus, according to Rosenthal, when Block claims that Rosenthal's higher-order thought theory is inadequate as an account of "phenomenal consciousness" (i.e. thin phenomenality), Block is *right*: thin phenomenality does not require higher-order thoughts. But, Rosenthal protests, the higher-thought theory was never intended as an account of thin phenomenality. Rather, it is an account of *thick* phenomenality – the higher-order kind of qualitative character, or experiential subjectivity.

4. Against FOR theories

So far, we have seen three accounts of what is – at least after some shoehorning – the same distinction: worldly vs. experiential subjectivity, lower-order vs.

higher-order qualitative character, and thin vs. thick phenomenality. This explosion of terminology will be all sound and fury if the distinction turns out to be merely notional; if, that is, there can't be non-conscious experiences. Now the task is to examine why our three HOR theorists think that the distinction is real, and their related reasons for thinking that FOR theories are false. To anticipate: it will be argued that the reality of the distinction has not been demonstrated, and that the associated case against FOR theories fails.

Recall an earlier complication: the reality of the distinction between experiences and conscious experiences does not entail the falsity of FOR theories (at least as explained here). According to the FOR theorist, subtracting higher-order representations from a phenomenally conscious experience does not subtract phenomenal consciousness. That is consistent with the claim that there could be experiences that are not phenomenally conscious.

In Carruthers' presentation, the argument for the reality of the distinction is only the first stage in a much longer argument against FOR theories and for HOR theories. In contrast, it is an immediate consequence of Rosenthal's and Lycan's explanations of why the distinction is real that FOR theories are false.

Given the Lycan/Rosenthal interpretation of Block (Sections 3.2 and 3.3), ambiguity threatens to break out in this paper. Temporarily setting aside the question of what Block actually means, 'phenomenal consciousness' is used in this paper *in Carruthers' sense*: phenomenal consciousness, experiential subjectivity, higher-order qualitative character, and thick phenomenality are all the same.

4.1 Thick phenomenality and higher-order representations: Rosenthal

One place where Rosenthal argues at length for the reality of the distinction between thick and thin phenomenality is in a paper responding to Block. According to Block (2001), in cases of visual extinction (where the subject reports he cannot see the stimulus, yet by other measures perceives it), it may be that the subject is undergoing a phenomenally conscious experience of the stimulus. That is – as Block would be happy to rephrase it – it may be that there is something it's like to perceive the stimulus. Rosenthal objects:

> As Nagel stressed in the article that launched that phrase, what it's like to have an experience is what it's like *for* the individual that has the experience. When a person enjoys the taste of wine, thereby enjoying gustatory phenomenality, there is something it's like *for that person* to experience the taste of the wine.
>
> Not so in cases of visual extinction; there is nothing it's like *for* an extinction subject to have a qualitative experience of the extinguished stimuli. That's why

> seeing visual extinction as the having of phenomenality without knowing it does not fit comfortably with the explanation of phenomenality in terms of what it's like to have an experience. (Rosenthal 2002a: 656)

This passage claims that an extinction subject is undergoing a visual experience that is *not* phenomenally conscious (because there is nothing it's like for the subject to undergo the experience). If correct, this would immediately establish the reality of the distinction between experiences and conscious experiences. (As just noted, more work is required to establish that FO theories are false.) Since this is basically Carruthers' style of argument for the reality of the distinction, discussion of this will be postponed to the section after next.

However, the second paragraph makes another point, one that is potentially more powerful. It suggests that if there is something it's like *for the subject* to undergo an experience, the subject is aware of having the experience, a suggestion that is repeated in the surrounding paragraphs. ('Knowing it' in the last quoted sentence is evidently interchangeable with 'being conscious of it'; see p. 658.) If that is correct, then FOR theories are false: being aware that one is having an experience is a necessary condition for the experience to be phenomenally conscious.

Rosenthal's argument for this necessary condition turns on the observation that there *is* something it's like to be a rock, or – his example – a table. ("What it's like to be a table, for example, is roughly something's having the characteristic features of tables" (656).) This "more general, nonmental" use of 'what it's like' is distinguished from the "special use to describe subjectivity" by the addition of the prepositional phrase 'for so-and-so'. There *isn't* something it's like *for* the table to be kicked in the leg; there *is* something it's like *for* Mr. N.N. to be kicked in the leg.[16] The crucial step is this:

> . . .conscious[ness] of oneself. . .must in any case occur if there is something it's like to have the experience. We're not interested in there being something it's like *for somebody else* to have the experience; there must be something it's like *for one* to have it, oneself. Without specifying that, what it's like would be on a par with what it's like to be a table. (656, my italics)

This passage only purports to explain why awareness *of oneself* is a necessary condition for having a phenomenally conscious experience (as opposed to awareness that one is having the experience), so some additional steps are needed to show that FOR theories are false. But let us waive this point.

There is certainly a distinction between there being (a) something it's like *for Mr. N.N.* to have an experience; (b) something it's like *for Ms. M.M.* to have an experience; (c) something it's like *for a person* (no particular person)

to have an experience. Rosenthal is claiming that the correct account of these distinctions will imply that if there is something it's like *for so-and-so* to have an experience then so-and-so is aware of himself (and/or his experience). However, there is also a distinction between there being (a) something it's like *to be Tim-the-table*; (b) something it's like *to be Tom-the-table*; (c) something it's like *to be a table*. What's more, Rosenthal evidently agrees: "there is something it's like. . .even to be this very table" (656). Clearly the correct account of the latter distinctions will not imply that if there's something it's like to be Tim-the-table, then Tim is aware of itself. So why think that the correct account of the former distinctions will have a similar implication?

In fact, the quoted passage doesn't really offer a reason: it seems to be simply repeating the claim that if there is something it's like *for so-and-so* to have an experience then so-and-so is aware of himself. And this is far from obvious.

Consider a specific example:

(*) There is something it's like for Mr. N.N. to see a cucumber.

(*) is equivalent to 'For Mr. N.N. to see a cucumber is like something' which in turn is equivalent to 'For a cucumber to be seen by Mr. N.N. is like something'. This illustrates the fact that in (*) the preposition 'for' has no particular attachment to 'Mr. N.N.'; it is instead the complementizer of the infinitival clause 'Mr. N.N. to see a cucumber'. (Unlike, for example, 'for' in 'The police are looking for Mr. N.N.'.) Hence there is no syntactic reason to think that (*) will have some exciting entailment solely about Mr. N.N. – say, that he is aware of himself. Perhaps more strongly, it is doubtful that 'There is something it's like for so-and-so to φ' has some "special use to describe subjectivity" (dialects of analytic philosophy aside). 'What was it like for the car to be driven in the desert?' appears to be an intelligible and literal question, that one might ask a driver returning from the Paris-Dakar Rally.

In any case, even if we grant for the sake of the argument that if there is something it's like for Mr. N.N. to see a cucumber, then Mr. N.N. is aware of *something*, it is unclear why this needs to be Mr. N.N. himself, or his experience. An analogy: there is something it's like for Mr. N.N. to read *Wittgenstein's Poker*. That might be true, even though Mr. N.N. is so captivated by the book that he is not aware of his reading; instead, he is aware of the dramatic encounter between Wittgenstein and Popper at the Moral Sciences Club, and other events the book describes. Similarly, there might be something it's like for Mr. N.N. to see a cucumber, simply because Mr. N.N. is aware of the cucumber – not of his seeing of the cucumber.

In short: there seems little prospect of deriving the falsity of FOR theories (and the consequent reality of the distinction between experiences and conscious experiences), from the semantics of 'what it's like for so-and-so to φ', considered as a phrase of ordinary English.

4.2 Higher-order qualitative character and higher-order representations: Lycan

As Section 3.2 explained, Lycan thinks that 'what it's like', in its *philosophical* usage, is ambiguous, and consequently that the participants in the debate are mostly talking past each other. Carruthers, remember, thinks that Tye is offering a first-order account of phenomenal consciousness, and accordingly expends much energy on the attempt to demonstrate its falsity. Lycan thinks that Carruthers is wrong about the exegetical claim, and that there is no disagreement once the ambiguity in 'what it's like' is pointed out:

> Now, "what it's like": that phrase certainly is usually used in Carruthers' rich way, but it has also been used to mean just a bare phenomenal quale, such as the phenomenal property of redness that Dretske, Tye and I explicate representationally and that figures in what Carruthers calls a percept. What Tye means, presumably, is that there are "phenomenally-conscious" states in the weak [i.e. lower-order] sense, states having qualitative character (so there is "something it is like" to be in them, in the weak [i.e. lower-order] sense of that expression), but of which the subject is unaware (and so there is nothing "it is like" for the subject to have them, in the strong [i.e. higher-order] sense of that expression). (Lycan 1999a)

Notice that the last sentence moves from 'unaware of the experience' to 'nothing it is like [in one sense]' as if the transition is entirely unproblematic. But the transition is not at all obvious.

Earlier, Lycan's distinction between lower-order and higher-order qualitative character was identified with the distinction between experiences and conscious experiences – that is, with the distinction between experiences, and experiences that there is something it's like for the subject to undergo. Now we may take Carruthers to hold that 'something it's like', as it appears in the preceding sentence, is not ambiguous (2000: 13); likewise for Rosenthal. So both Carruthers and Rosenthal should be comfortable with the explanation of the distinction between experiences and conscious experiences in terms of 'what it's like'.

On the other hand, by Lycan's lights, this explanation of the distinction is seriously ambiguous (as is much else in this paper). Resolving the ambiguity

one way, the distinction is merely notional: in the lower-order sense, for *any* experience, there is something it's like for the subject to undergo the experience. Resolving it another way, and assuming that there are experiences of which the subject is unaware, the reality of the distinction is shown by actual cases: in the higher-order sense, there is *nothing* it is like to undergo some experiences.[17]

FOR theories (as explained in Section 2) claim that there could be phenomenally conscious experiences (i.e. experiences with higher-order qualitative character) of which the subject is unaware. According to Lycan, from the fact that an experience has higher-order qualitative character it immediately follows that the subject is aware of the experience; hence, on Lycan's view, FOR theories are trivially false. (This is of course why Lycan thinks that Tye and Dretske are *not* FOR theorists in the sense of this paper.)

Let us examine Lycan's claim of ambiguity first. The quoted passage highlights a point noted in Section 3.2, namely that Lycan thinks that 'what it's like' in its lower-order sense picks out a quale – greenness or squareness, for example. Taking this at face value, Lycan is saying that when some philosophers (notably Tye and Dretske) ask 'What is it like to see cucumbers?', or 'What are experiences of cucumbers like?', they are using these expressions in such a way so that the correct answer is 'Green'. Now – as Lycan would agree – there is clearly no *natural* way of understanding these questions on which this is the right answer. 'Green' is a right answer to the question 'What is the *cucumber* like?'; extraordinary contexts aside, it is never a right answer to the question 'What is *seeing* the cucumber like?'. But, further, 'what it's like' is not taken by Tye or Dretske to be a *technical* expression, whose meaning is a matter of explicit or implicit stipulation. Unlike, say, 'intentionality', 'what it's like' is typically used in the literature without any special explanation or definition. Therefore there is no reason to postulate an ambiguity.[18]

And if there is no ambiguity, then Lycan's position reduces to Rosenthal's. Namely, that if there is something it's like for the subject to undergo an experience (in the univocal sense of 'something it's like'), then the subject is aware of the experience.[19] But the previous section found no persuasive argument for that conclusion.

4.3 Experiential subjectivity and higher-order representations: Carruthers

Progress so far can be summarized in four points. First – from the previous section – there is no ambiguity of the relevant sort in 'what it's like'. Second – from Section 4.1 – there is no easy route from the semantics of 'what it's like for so-and-so to φ' to the falsity of FO theory. The first two points support the

third and fourth. *A* version of the FOR/HOR dispute was explained successfully in Section 2, and has no obvious resolution. Lastly, because of the lack of ambiguity, and the fact that FO theories are live options, Block, Tye and Dretske *are* FO theorists, and hence are genuinely opposed to HOR theories.

Section 2 confidently proclaimed that Lycan and Rosenthal endorsed HOR theories; some qualifications and reservations are now needed. Because of Lycan's view that he and his opponents are mostly separated by terminology, one might question whether Lycan is an HOR theorist of phenomenal consciousness, given that this position really is controversial. And in fact, Lycan's usual way of explaining HOR theories is not remotely equivalent to the characterization given in Section 2: "HOR theories are. . .theories of 'conscious states', in the sense of states whose subjects are aware of being in them. . .HOR theories. . . [suppose] that what makes a state a conscious state *in this sense* is that it itself is represented by another of the subject's mental states" (2001a: 3; my italics).[20] In Rosenthal's case, although there is little doubt that he endorses an HOR theory of phenomenal consciousness, a significant amount of what he says under the rubric of 'a theory of consciousness' does not depend on this commitment.[21]

Carruthers, on the other hand, can be labeled an HOR theorist with little reservation or qualification. He rests his case for non-conscious experiences on a series of examples, of which the following are representative:

A. Armstrong's absent-minded driver
Armstrong's driver, his mind on other matters, "comes to" and can't recall seeing anything on the road he has been driving on for the last half an hour. Yet the driver navigated the road safely, so he must have seen the traffic. (Armstrong 1981; Carruthers 2000: 148–149.)

B. Two visual systems
Milner and Goodale's patient D.F. (suffering from visual form agnosia) can't recognize the orientation of a slot in a disk – whether the slot is vertical, horizontal, etc. – but can use visual information to post a letter through it (Milner & Goodale 1995: 128–129; Carruthers 2001: 161). Similar dissociations (between the so-called "what" and "how" pathways) can be found in normal subjects. In the Tichener illusion (illustrated in Milner and Goodale: 168; Carruthers: 163), two equally sized discs look different in size; yet when normal subjects are asked to pick up the discs, their grip aperture is approximately the same for both. (See also Rosenthal's discussion of visual extinction quoted in Section 4.1 above.)

Carruthers takes many extra steps from the reality of the distinction between experiences and conscious experiences to the falsity of FOR theories, and many more from that to his own brand of the HOP theory. Here only the first step will be disputed: the conclusion that these examples show that the distinction is real.

Take Armstrong's driver first. It is not implausible that the driver is undergoing experiences of the road. (Absent-minded academics spend most of their waking lives in such a condition.) Was there something it was like for him to see a green traffic light? Well, why not? After all, variants of the example are familiar, for example when some time after "coming to" one recalls that the light was green, remembers that the experience was like something, and in particular what it was like. If one remembers this, then there *was* something it was like for one to undergo the experience, and so the experience was phenomenally conscious. Suppose one is not able to remember what it was like: that does not show that the experience was *not* phenomenally conscious. Armstrong's driver is not a convincing exhibit. (In fact, Carruthers himself places little weight on examples like A.[22])

What about cases involving a dissociation between the "two visual systems"? Here there seems little pressure to say that the subjects are having the relevant kinds of *experiences* in the first place, and a fortiori little pressure to admit that the subjects are undergoing non-conscious experiences. If D.F. can see (or at any rate perceive) the orientation of the slot, then *she* is aware of the orientation of the slot. In the Tichener illusion experiment, if the subject sees (or at any rate perceives) that the discs are the same size, *she* is aware that the discs are of the same size. Yet it does not seem very plausible that *D.F.* is aware of the orientation of the slot – she is unable to say what it is, or make plans that depend on this piece of information. Similarly with the normal subjects in the Tichener illusion experiment, and Rosenthal's example of visual extinction.

The above points do not presuppose that the ability to report that p is a necessary condition of perceiving that p, merely that the inability to report is sometimes good evidence that the subject is not perceiving that p. And, of course, it should be admitted that the crucial information (that the slot is horizontal, that the discs are the same size) is playing an important role in the subjects' cognitive economy, in particular in guiding the subjects' behavior. However, the information that p may play such a role without *the subject* being aware that p. If there are such things as subjects or selves (i.e. persons, in the case of adult humans), then this can hardly be denied. Finally, the above points do not presuppose that the subject or self is a primitive irreducible entity. Perhaps some neo-Humean reductionist account of the self is correct: as

Block puts it, we are "loose federations of centers of control and integration" (1997: 162). That is entirely consistent with the claim that (for example) D.F. does not perceive the orientation of the slot.

The topic of the self hardly deserves such cursory treatment, but will receive it due to space limitations. Still, we may conclude that – at least prima facie – Carruthers' examples do not add up to a convincing case for the reality of the distinction between experiences and conscious experiences.

In fact, this conclusion leaves most of Carruthers' case against *broadly reductive* FOR theories unaffected. In the chapter where the distinction is put to the most work, Carruthers is arguing against FOR accounts that may be approximately schematized as follows:

(†) Event e is phenomenally conscious iff e has (first-order) functional role R.

And, whether or not examples like those in B are cases of non-conscious experiences, one can easily see how they could pose problems for the theories that are instances of (†). Maybe one theory predicts (implausibly, or so we may grant) that the Tichener illusion subjects *do* have phenomenally conscious experiences of the discs being the same size – which of course they are unable to report. Maybe another only avoids this prediction by making an ad hoc and otherwise ill-motivated assumption. And so on.[23]

Still, although Carruthers may have shown that particular reductive theories have serious problems, this is not to impugn FOR theories across the board.

5. Against HOR theories

The previous sections have not argued that HOR theories are mistaken; this final section gives a short argument for that conclusion.

What is *sufficient* for one to know what one's experience is like, in the sense relevant to phenomenal consciousness?[24] Suppose that one knows – via "introspection", whatever that is exactly – that one's present experience has property P, and one knows that certain past experiences had property P. Suppose that one knows that yet other past experiences had other properties, Q, R,... Further suppose that experiences possessing these properties thereby saliently resemble each other, and that one knows these facts about similarity. So, for example, one may know that one's (present) experience is more similar (in this salient respect) to a Q-experience than to an R-experience, and so on. This appears to be sufficient for one to know what one's (present) experience is like. If it is not sufficient, it is unclear what else needs adding. (Insisting that P, Q,

R,. . . must be *intrinsic*, for example, would make it very hard for a typical intentionalist to maintain that one can know what one's experience is like.) Assume, then, that it is sufficient. If all these facts about one's experience obtain, let us say that one's experience is in condition C. Knowing that one's experience is in C, then, is sufficient for knowing what one's experience is like. Hence, if one's experience is in C, then one's experience *is* like something (in the relevant sense), and therefore one's experience is phenomenally conscious.

Now observe that such a condition may be constructed from purely first-order materials. The roles of P, Q, R,. . . can be played by the properties of being an experience of green, being an experience of turquoise, being an experience of blue, and so forth. Everyone can agree that one can know (at least sometimes) by introspection that one's experience is an experience of green. Recall that we are assuming that Rosenthal's "mental green", an especially striking and "qualitative" property of experiences of limes, grass, etc., is identical to the property of being an experience of green (Section 3.1 above). Under this assumption, at least *some* of the salient similarities and differences between experiences are due to what they represent about one's environment: because of their content, experiences of green are saliently more similar to experiences of turquoise than to experiences of blue, and so on. If, on the other hand, Rosenthal is correct, and mental green is non-intentional, then the desired condition may be constructed with the non-intentional but first-order properties mental green, mental turquoise, and so forth, playing the roles of P, Q, R,. . .

Now distinguish *knowing* what one's experience is like from one's experience *being* like something. Setting misguided Cartesian doctrines about the self-intimation of phenomenology aside, one's experience may be like something (in the sense relevant to phenomenal consciousness), even though one does not know that one's experience is like something. (This is just an instance of the general principle that any contingent proposition p may be true and not known.)

Take condition C as constructed in the paragraph before last. From the point just made, we can conclude that one's experience may be in C, even though one does not know that one's experience is in C. Indeed, we can go further, and conclude that one's experience can be in C even though it is not the target of any higher-order representation at all. Setting aside Rosenthal's view for simplicity, this last step basically amounts to saying that an experience of green that is not the target of a higher-order representation is more similar in salient respects to experiences of turquoise than it is to experiences of blue, which HOR theories can hardly dispute. But it was argued that any experience

in C is a phenomenally conscious experience. HOR theories deny that, and so they are in error.

This argument does not show that FOR theories, as explained in Section 2, are true. For all that has been said, maybe higher-order representations do have something to contribute to phenomenal consciousness. Perhaps knowing what one's experience is *completely* like (in the sense relevant to phenomenal consciousness) requires knowing that one's experience is in some higher-order condition (cf. Carruthers 2000: 183–184). In particular, perhaps knowledge that one's experience has our old friend G is not first-order knowledge. Further argument is needed to foreclose this possibility. But if we take the explanation of phenomenal consciousness in terms of 'what it's like' seriously, then higher-order representations are not necessary for phenomenal consciousness.

Notes

* Many thanks to Peter Carruthers, Rocco Gennaro, Alec Marantz, Sarah McGrath, and an audience at ASU. I am especially indebted to Bill Lycan and David Rosenthal who kindly supplied extensive comments on earlier drafts.

1. 1996: 77.

2. See Johnston (1992: 221).

3. This broadening of the supervenience base is needed to accommodate higher-order theories.

4. Lycan adds facts about "functional organization" (1996: 11) to the intentional supervenience base for phenomenal character. On a more careful statement of intentionalism than the one given here, this is an intentionalist view. See Byrne (2001: 204–206).

5. Intentionalism or representationalism comes in a variety of inequivalent formulations. On some of these, as Rosenthal (forthcoming, Section 4) points out, he is *not* an HO intentionalist.

6. Why not necessary and sufficient? Here is one reason: some episodes of mental imagery have G (at least on a natural way of understanding the introduction of 'G'), and these are not perceptual experiences. The complicated question of how the FOR theorist should state a necessary and sufficient condition for G need not be examined here.

7. As does Gennaro (1996).

8. Carruthers also holds a dispositional version of the HOT theory but, for reasons that need not concern us here, thinks this is equivalent to a HOP theory (see Carruthers 2000; this volume).

9. Two qualifications concerning Rosenthal's view should be noted. First, according to Rosenthal the higher-order thought specifies one's experience in partly *non*-intentional respects, as having a property Rosenthal calls *mental green* (see Section 3.3 below). Sec-

ond, to avoid certain counterexamples, Rosenthal adds that the higher-order thought must "[rely] on neither inference nor observation" (Rosenthal 1997b: 738). For our purposes these qualifications can be ignored.

10. Carruthers (2000: 248) takes the experience with the lower-order content to be the same as the experience with the higher-order content.

11. One might wonder what the HOR theorist should say if the higher-order representation misrepresents the lower-order one. See Byrne (1997: 119–124), Neander (1998), Lycan (1998), Levine (2001: 104–111), Rosenthal (2002a), Rosenthal forthcoming; Carruthers this volume; Gennaro this volume.

12. A first-order theorist might say that such proprioceptive experiences are phenomenally non-conscious because they lack a certain special sort of content, or lack a certain functional role. Smith (2002: 165) claims that experiences of resistance to one's bodily movements have a "unique non-sensory nature", which is arguably equivalent to the claim that they need not be phenomenally conscious.

13. Cf. the alternative use of 'quale' noted in Section 1 (for yet another use, see Carruthers 2000: 15).

14. More cautiously: a (perceptual) experience has mental green iff it is a registering of a green quale. It may be that, in Rosenthal's intended usage, there are some events that have mental green but which aren't perceptual experiences (and hence not "registerings"); in particular, some "sub-personal" events in the subject's cognitive system. See the following note.

15. In fact, Rosenthal seems to intend thin phenomenality to be somewhat broader than the category of worldly subjectivity/lower-order qualitative character, including occurrences of qualitative character at early stages of perceptual processing (see 2002a: 660). This is another complication that can be set aside.

16. Nagel, incidentally, seems to take 'what it's like to *be* X' to be equivalent to 'what it's like *for* X' (1974: 436). This is quite doubtful. Note that 'What is it like to be X?' is complete in a way that 'What is it like for X?' is not. In order for the latter to express a question, some activity or state needs to be specified: 'What it is like for X to eat olives/drink gin/be drunk?' A more plausible equivalence is between 'what it's like to be X' and 'what it's like for X to exist/to be'.

17. One of Lycan's reasons for thinking that there are experiences of which the subject is unaware is that we speak of "unfelt pains" (1996: 16). See also Rosenthal (1997b: 731–733, 2002b: 411–412).

18. Dretske does say (in effect) that anyone who knows what green is knows what it's like to see green (1995: 81–93), which might suggest a Lycanian "lower-order" interpretation. But Dretske's point is that knowing what green is suffices (given a certain conceptual sophistication) to know that seeings of green *represent* green – *that* is what such experiences are like. See also Dretske (1999).

19. It appears that, for Lycan, only the *lower*-order sense of 'what it's like' is a special technical one; the higher-order sense is just the ordinary sense. In support of this interpretation, notice that Lycan himself *uses* 'what it's like' in explaining the higher-order sense (see the fourth paragraph in Section 3.2 above).

20. See also Lycan this volume, and the following comment on Siewert (1998):

> My theory of. . .consciousness is, in brief: (1) Phenomenal/qualitative character is a combination of one or more (Lewis-) qualia and some other components. . . (2) . . . (3) . . . (4) . . .If sensory *experience* (and Siewert's Phi-consciousness) presuppose awareness, something further needs to be added. (5) I add. . .awareness!. . . (Lycan 2001b: Section 5)

21. See, e.g., Rosenthal (1986, 1997b, 2002b). I should emphasize that this paper simply ignores the usual arguments that Rosenthal gives for his "theory of consciousness", none of which turn on the phrase 'what it's like'.

22. For an excellent discussion of this example, see Lycan and Ryder (2003). See also Rosenthal (2002b: 407).

23. See Carruthers (2000: 168–179).

24. It is assumed here that to "know what one's experience is like" is to have propositional knowledge (see Lycan 1996: 92–94). For the contrary view, see Lewis (1990). The addition of 'in the sense relevant to phenomenal consciousness' is meant to accommodate the context-sensitivity of 'know what one's experience is like', not to suggest any ambiguity. One may use a cerebroscope to know what one's experience is like (in various neural respects) – that does not imply that one's experience is phenomenally conscious.

References

Armstrong, D. M. (1981). What is consciousness? In D. Armstrong (Ed.), *The Nature of Mind and Other Essays* (pp. 55–67). Ithaca, NY: Cornell University Press.

Block, N. (1990). Inverted earth. *Philosophical Perspectives, 4*, 53–80.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227–247.

Block, N. (1997). Biology versus computation in the study of consciousness. *Behavioral and Brain Sciences, 20*, 159–166.

Block, N. (2001). Paradox and cross purposes in recent work on consciousness. *Cognition, 79*, 197–219.

Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies, 86*, 103–129.

Byrne, A. (2001). Intentionalism defended. *Philosophical Review, 110*, 199–240.

Carruthers, P. (1998). Natural theories of consciousness. *European Journal of Philosophy, 6*, 203–222.

Carruthers, P. (2000). *Phenomenal Consciousness.* Cambridge: Cambridge University Press.

Carruthers, P. (this volume). HOP over FOR, HOT theory.

Dretske, F. (1995). *Naturalizing the Mind.* Cambridge, MA: MIT Press.

Dretske, F. (1999). The mind's awareness of itself. *Philosophical Studies, 95*, 103–124.

Gennaro, R. (1996). *Consciousness and Self-Consciousness.* Philadelphia, PA: John Benjamins.

Gennaro, R. (this volume.) Higher-order thoughts, animal consciousness, and misrepresentation: A reply to Carruthers and Levine.
Johnston, M. (1992). How to speak of the colors. *Philosophical Studies, 68*, 221–263.
Levine, J. (2001). *Purple Haze.* Oxford: Oxford University Press.
Lewis, D. (1990.) What experience teaches. In W. Lycan (Ed.), *Mind and Cognition* (pp. 499–519). Oxford: Blackwell.
Lycan, W. G. (1996). *Consciousness and Experience.* Cambridge, MA: MIT Press.
Lycan, W. G. (1998). In defense of the representational theory of qualia (replies to Neander, Rey and Tye). *Philosophical Perspectives, 12*, 479–487.
Lycan, W. G. (1999a). A response to Carruthers' 'Natural Theories of Consciousness'. *Psyche* 5 <http://psyche.cs.monash.edu.au/v5/psyche-5-11-lycan.html>.
Lycan, W. G. (1999b). Dretske on the mind's awareness of itself. *Philosophical Studies, 95*, 125–133.
Lycan, W. G. (2001a). A simple argument for a higher-order representation theory of consciousness. *Analysis, 61*, 3–4.
Lycan, W. G. (2001b). Have we neglected phenomenal consciousness? *Psyche* 7 <http://psyche.cs.monash.edu.au/v7/psyche-7-03-lycan.html>.
Lycan, W. G. & Z. Ryder (2003). The loneliness of the long-distance truck driver. *Analysis, 63*, 132–136.
Lycan, W. G. (this volume). The superiority of HOP to HOT.
Milner, A. D. & M. A. Goodale. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.
Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435–450.
Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. *Philosophical Perspectives, 12*, 411–434.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 49*, 329–359.
Rosenthal, D. (1997a). Phenomenal consciousness and what it's like. *Behavioral and Brain Sciences, 20*, 156–157.
Rosenthal, D. (1997b). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness* (pp. 729–753). Cambridge, MA: MIT Press.
Rosenthal, D. (1999). The colors and shapes of visual experiences. In D. Fisette (Ed.), *Consciousness and Intentionality* (pp. 137–169). Dordrecht: Kluwer.
Rosenthal, D. (2002a). How many kinds of consciousness? *Consciousness and Cognition, 11*, 653–665.
Rosenthal, D. (2002b). Explaining consciousness. In D. Chalmers (Ed.), *Philosophy of Mind* (pp. 406–421). Oxford: Oxford University Press.
Rosenthal, D. (forthcoming). Sensory qualities, consciousness, and perception. In D. Rosenthal (Ed.), *Consciousness and Mind*. Oxford: Oxford University Press.
Shoemaker, S. (1981). The inverted spectrum. *Journal of Philosophy, 74*, 357–381.
Siewert, C. (1998). *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.
Smith, A. D. (2002). *The Problem of Perception*. Cambridge, MA: Harvard University Press.
Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Chapter 10

Either FOR or HOR
A false dichotomy

Robert W. Lurz

1. Introduction

We have intentional states that are *conscious*: we have conscious perceptual states, beliefs, and desires, for instance. We also have intentional states that make us conscious *of* things: we have perceptual states and beliefs that make us conscious of objects, properties of objects, and facts, for instance. What, though, is the relation between intentional states that *are conscious* and intentional states that make us *conscious of* things? Representationalism is the view that intentional states of the former sort are dependent upon intentional states of the latter sort, that no subject that did not have intentional states that made him conscious *of* things could have intentional states that were themselves *conscious*. And this is so, according to representationalism, because what makes a subject's intentional state *conscious* is the fact that the subject is (or is disposed to be) in an intentional state that makes him conscious *of* something.[1]

The aim of representationalism, it should be noted, is often expressed in broader terms, as being that of explaining what makes *any* type of mental state conscious, not just intentional states. This is certainly a correct expression of the theory's aim, and I do not mean to deny it. However, it is also true that the theory has the narrower aim of explaining what makes *intentional* states conscious; and it is this narrower aim of the theory that I shall be examining in this paper.

Now, representationalists – whether they are pursuing the wider or narrower aims of their theory – typically recognize but two kinds of intentional states that make subjects conscious *of* things – *first-order* and *higher-order* – and end up endorsing one of two kinds of representational theories – *first-order representationalism* (FOR) or *higher-order representationalism* (HOR). I

shall argue that this dichotomy is false, and that there is an alternative variety of intentional state that makes subjects conscious of things – the *same-order* variety – and an alternative representational theory – *same-order representationalism* (SOR). Moreover, I shall argue that SOR avoids some of the notable shortcomings of the two dominant representational theories and is, therefore, preferable to each. (I have argued for SOR – or a version of it – in Lurz (2002) and (2003).)

2. The explanandum and the explanans of representational theories

To avoid confusion, a few words need to be said about the intended *explanandum* and *explanans* of representational theories. Intentional states are states of a subject that are said to be of, about, or (more generally) directed at things outside themselves. Beliefs, desires, and perceptual experiences are paradigm examples. To have a belief is always to have a belief about something; to have a desire is always to have a desire for something; and to have a perceptual experience is always to have a perceptual experience of something. The item that an intentional state is said to be of, about, or directed at is its *intentional object*. If a subject has a belief about tomorrow's weather, or a desire for a cup of coffee, for instance, then the intentional object of his belief is tomorrow's weather, and the intentional object of his desire is a cup of coffee.

Representational theories aim to explain the nature of a particular property of intentional states – namely, the property that we ascribe to them when we say that they are conscious. Since this property is a property of intentional *states* – as opposed to being a property of the *subjects* of those states – representationalists typically call this property *state-consciousness*. (Sometimes this property is called '*intransitive*-consciousness' or '*intransitive-state*-consciousness,' so as to contrast it with *transitive-creature*-consciousness (discussed below)). So, for instance, were we to say that John's auditory experience of the ticking clock is conscious (as opposed to saying that it is unconscious or subliminal), or that Mary has a conscious belief about her illness (as opposed to saying that she has an unconscious belief about her illness), we would be ascribing to John's auditory experience and Mary's belief the property of state-consciousness. It is the nature of this property that is the intended *explanandum* of representational theories.

State-consciousness needs to be distinguished from another important property of intentional states: *phenomenal-consciousness*. An intentional state is said to be phenomenally conscious just in case there is something that it is

like for the subject to be in the state. The distinction between phenomenal-consciousness and state-consciousness is sometimes drawn by pointing to putative cases of state-consciousness without phenomenal-consciousness (e.g., cases of conscious belief) or cases of phenomenal-consciousness without state-consciousness (e.g., cases of unconscious or subliminal perception) (see Block 1995; Carruthers 2000). But, this way of drawing the distinction is somewhat contentious and, I think, unnecessary for the purpose of distinguishing the two properties. It is enough to note that phenomenal-consciousness is a *determinable* property (like the property of being colored), whereas state-consciousness is not. An auditory experience and a visual experience, for instance, are phenomenally conscious: each has the determinable property of there being *something* that it is like to undergo it. And yet, each experience also has its own distinct form or variety of phenomenal consciousness: what it is like to hear something is distinctly different from what it is like to see something.

The same cannot be said for state-consciousness, though. It makes no sense to ask what distinct form or variety of consciousness a conscious intentional state possesses. Now, it makes sense, no doubt, to ask what distinct form or variety of *intentional state* it is that is conscious, but not what distinct form or variety of *consciousness* the state possesses. In this way, state-consciousness is more like the property of being *visible* than the property of being *colored*. Both a lime and a tomato, for instance, may be visible, but they do not possess distinct varieties of visible as they do distinct varieties of color. In a similar way, a conscious visual experience does not possess a distinct variety of consciousness that is different from that possessed of a conscious auditory experience. Each experience is conscious, *period* – even though each has its own distinct form of phenomenal consciousness. Therefore, even if all and only conscious intentional states are phenomenally conscious, state-consciousness is not the same property as phenomenal-consciousness – any more so than the property of being visible is the same as the property of being colored: the latter properties are determinables, whereas the former are not.[2]

Now, a few words need to be said about the *explanantia* of representational theories: those intentional states that make subjects conscious *of* things. Representationalists typically call such intentional states *transitive-conscious states*, since the verb ('conscious') that is used to describe them is transitive, taking a grammatical object as its complement. (Sometimes these state are given the fuller title of transitive-*creature*-conscious states, since they are states that make the subject (the creature) conscious of things.) Since transitive-conscious states are intentional states that make a subject conscious of things, one can

always ask two sorts of questions about them. First, one can ask what item in the world the subject's transitive-conscious state is said to be of. This sort of question, of course, is about the state's intentional object. Second, one can ask whether the intentional object of a subject's transitive-conscious state is an immediate or a mediate intentional object, whether the subject is conscious of it simpliciter or conscious of it by means of being conscious of something else. This sort of question is about the directness or immediacy of a subject's transitive-conscious state.

To illustrate, suppose that I am conscious of a mouse on my bedroom floor by seeing it. My transitive-conscious (visual) state, in this case, has as its immediate intentional object a particular mouse; and so, I can be said to be immediately (or directly) conscious of the mouse on my bedroom floor. If, on the other hand, I was conscious of the mouse by seeing its reflection in the bedroom mirror, or its shadow on the floor, then the immediate intentional object of my transitive-conscious state would have been the mirror reflection or the shadow of the mouse, and its mediate intentional object would have been the mouse. In such a case, I can be said to be mediately (or indirectly) conscious of the mouse in my bedroom.

Representationalists have found it convenient to divide transitive-conscious states into two general classes in virtue of their immediate intentional objects. In one class, there are those transitive-conscious states that take as their immediate intentional objects the current intentional states of their subject. At the moment, for instance, I am conscious of having a desire for a cup of coffee; and yet I am not conscious of my desire by being conscious of something else, such as my behavior or reflection in the mirror; rather, I am conscious of having this desire in a direct and immediate way. Following the (somewhat) standard nomenclature in the field, I shall call those transitive-conscious states that have as their immediate intentional objects the subject's current intentional states, *higher-order* (HO) transitive-conscious states.

In addition to being immediately conscious of our own current intentional states, we are often immediately conscious of concrete items in our external environment, as the mouse example above illustrates.[3] Again, following the (somewhat) standard nomenclature in the field, I shall call those transitive-conscious states that take as their immediate intentional objects concrete items in a subject's external environment, *first-order* (FO) transitive-conscious states. Paradigm cases of FO transitive-conscious states are cases of external veridical perception (e.g., cases of seeing, hearing, smelling, etc.) and the perceptual beliefs that result from such states of perception.

The classification of transitive-conscious states into HO and FO types is not meant to be exhaustive, only exclusive. In fact, in Section 4, I will argue that there is a type of transitive-conscious state that is neither HO nor FO. But before I do this, I would like to use the above classification to explicate the two dominant representational theories that are currently on the philosophical market.

3. FOR and HOR

As stated above, representational theories aim to explain the nature of state-consciousness in intentional states in terms of transitive-conscious states. Given what was said above about the nature of such states, the general aim of representational theories can be expressed more perspicuously by the following general explanatory schema:

> *General Explanatory Schema* (GES): Where M is a conscious intentional state that is of or about *x* and is possessed by a subject S, what makes M conscious is the fact that S is (or is disposed to be) in a transitive-conscious state M* that makes S immediately conscious of *y*.

Although all representationalists endorse GES, they do not all agree on how it is to be filled out. In particular, they are divided over whether the subject's conscious intentional state M and his transitive-conscious state M* should be identical or distinct states. And they are divided over whether the immediate intentional object of the subject's transitive-conscious state M* – the thing denoted by '*y*' – should be the intentional object of his conscious state M – the thing denoted by '*x*' – or whether it should be the subject's conscious state M itself.[4] We can express this division more clearly by seeing how the two dominant representational theories go about answering the following three questions pertaining to GES:

> Q1: Is M = M*?
> Q2: Is *x* = *y*?
> Q3: Is *y* = M?

Those representationalists that answer Q1 and Q2 in the affirmative and Q3 in the negative typically endorse *first-order representationalism* (FOR). Perhaps, the best-known defender of FOR is Fred Dretske, who offers the following formulation of the theory:

> What makes an internal state or process conscious is the role it plays in making one (transitively) conscious – normally, the role it plays in making one (transitively) conscious of some thing or fact. An experience of *x* is conscious not because one is aware of the experience, or aware that one is having it, but because, being a certain sort of representation, it makes one aware of the properties (of *x*) and objects (*x* itself) of which it is a (sensory) representation. My visual experience of a barn is conscious not because I am introspectively aware of it (or introspectively aware that I am having it), but because it (when brought about in the right way) makes me aware of the barn. It enables me to perceive that barn. For the same reason, a certain belief is conscious not because the believer is conscious of *it* (or conscious of having it), but because it is a representation that makes one conscious of the fact (that *P*) that it is a belief about. (1993:281)

There are three important points expressed in Dretske's formulation of FOR that are relevant to GES. First, the transitive-conscious state M* that is involved in making a subject's intentional state M conscious is the same state as the subject's intentional state M: what makes M conscious is that *M itself*, and not some other intentional state, makes the subject (or disposes or enables the subject to be) conscious of something.[5] Second, since the transitive-conscious state M* that is involved in making a subject's intentional state M conscious is the same as the subject's intentional state M, the immediate intentional object of M* must be the intentional object of M: what M* makes the subject (or enables the subject to be) immediately conscious of (*x*) is whatever M is of or about (*x*). And third, if the intentional object of the subject's conscious intentional state M is a concrete item in the external environment (e.g., a barn), then the transitive-conscious state M* that makes M conscious will have this concrete item as its immediate intentional object. On FOR, then, the transitive-conscious state M* that is involved in making a subject's intentional state M conscious need not be a HO transitive-conscious state, but may be a FO transitive-conscious states – which, after all, is why FOR is typically called '*FO*R.'

Putting these points together, we can give the following slightly more formal definition of FOR:

> *FOR*: Where M is a conscious intentional state that is of or about *x* and is possessed by a subject S, what makes M conscious is the fact that M makes S (or enables S to be) immediately conscious of *x*.

Its elegance notwithstanding, FOR faces two serious problems. The first results from its affirmative answer to Q1 above. The problem is that not all conscious

intentional states are of the right sort to make their subjects (or enable their subjects to be) conscious of things. For to be conscious of something is to be in a *cognitive* intentional state whose function is to inform its subject of the way the world (supposedly) is. But not all intentional states are cognitive. Some intentional states, such as desires, are *conative*. And their function is to motivate, not to inform. My desire for another beer, for instance, motivates me to get another beer, but it does not make me (or enable me to be) conscious of another beer or of my having another beer. Nevertheless, my desire may very well be a conscious state. It is difficult to see, however, how FOR could explain the conscious status of my desire (or any conative state, for that matter) given that my desire is not a cognitive state that makes me (or enable me to be) conscious of anything.[6]

FOR also faces a serious problem as a result of its affirmative answer to Q2 above. The problem is that not all conscious intentional states make us conscious of their intentional objects; for some of their intentional objects do not actually exist, and one cannot be said to be conscious of something that does not actually exists – the expression, 'S is conscious of *x*,' is factive, taking only words and phrases that purport to denote actual existing items as its complement. How, then, would a FOR theorist explain the conscious status of, say, a conscious visual hallucination of a pink elephant or a conscious belief that Santa Claus exists? He cannot say that such states make their subjects (or enable their subjects to be) conscious of the things that the states are said to be of or about; for no such individual (a pink elephant or Santa Claus) or fact (Santa Claus existing) actually exists. And I assume that the FOR theorist does not want to say that such states make their subjects (or enable their subjects to be) conscious of mental particulars, such as sense data or mental images; for such a view is extremely problematic and leads directly to a type of mind-body dualism, a position that no FOR theorist that I know endorses. But what else can the FOR theorist say?

At one point, Dretske (1999: 5–8) suggests that, in cases of visual hallucinations, subjects *are* conscious of things that actually exist: they are conscious of uninstantiated universals. When one visually hallucinates a pink elephant, on this view, one is visually conscious of certain universals – the universals pink and elephantine, for instance – that are not instantiated by anything (at least, not at the time or in the vicinity of one's hallucination). Does Dretske's view here solve the problem for FOR? I do not believe so. Uninstantiated universals are abstracta on par with numbers; and though we certainly can be conscious of them through acts of thought and reason, we certainly cannot be conscious of them through acts of sensory perception. Just as one cannot be said to see

(be visually conscious of) numbers, one cannot be said to see (be visually conscious of) uninstantiated universals; and, therefore, one cannot be said to see (be visually conscious of) uninstantiated universals when undergoing visual hallucinations.

No doubt, one can form beliefs about uninstantiated universals (e.g., the uninstantiated universal pink) while undergoing visual hallucinations, and in virtue of being in such belief states, one may be said to be conscious of such universals, but the visual states that are involved in visual hallucinations no more make one conscious of uninstantiated universals than the visual states that are involved in veridical perception make one conscious of uninstantiated universals, numbers, or the like. I do not see, then, that Dretske's view here solves the problem that FOR incurs as a result of its affirmative answer to Q2 above.

But if one cannot be said to be conscious of uninstantiated universals, or pink elephants, or Santa Claus, or sense data, while undergoing a conscious visual hallucination of a pink elephant or a conscious belief that Santa Claus exists, then what can one be said to be conscious of in such cases, according to FOR? I do not believe that the problem that this question constitutes, or the problem of conscious conative states outlined earlier, can be answered easily by the FOR theorist. And since problems should be either solved or avoided, the general advice to representationalists at this point would appear to be that these two problems are best avoided, if possible. Representationalists, it seems, are best advised to answer Q1 and Q2 in the negative, if they can.

One group of representationalists that takes this advice to heart is the group that endorses *higher-order representationalism* (HOR). David Rosenthal offers perhaps the clearest and most succinct expression of HOR when he writes, "a mental state is intransitively conscious just in case we are transitively conscious of it" (1997:737). To be transitively conscious of one's intentional state M, according HOR theorists, is to be in a transitive-conscious state M* that is distinct from but contemporaneous with one's intentional state M and has one's intentional state M as its immediate intentional object.[7] The transitive-conscious state M* that is involved in making a subject's mental state M conscious, therefore, is a HO transitive-conscious states, according to HOR. HOR theorists, therefore, not only answer Q1 and Q2 above in the negative, they also answer Q3 in the affirmative. According to HOR, then, a subject's conscious desire for or belief about a cup of coffee, for instance, is conscious in virtue of the fact that the subject is (or is disposed to be) immediately conscious of having the desire or belief, which, in turn, involves the subject being in (or being disposed to be in) a HO transitive-conscious state (M*) that is distinct from but con-

temporaneous with his state of desire or belief (M) and takes the subject's state of desire or belief (M) as its immediate intentional object.

Bringing these points together, we can give the following more formal definition of HOR:

> *HOR*: Where M is a conscious intentional state that is of or about *x* and is possessed by a subject S, what makes M conscious is the fact that S is (or is disposed to be) in a contemporaneous transitive-conscious state M* that makes S immediately conscious of M.[8]

As a result of giving negative answers to Q1 and Q2, HOR avoids those problems that appear endemic to FOR. However, HOR faces its own unique problem as a result of its affirmative answer to Q3. The fact that HOR requires a subject to be (or be disposed to be) conscious of having intentional states in order for the subject's intentional states to be conscious makes HOR's explanation of conscious beliefs and desires in young children problematic. For children begin to *say* what they want and believe at around 1 or 2 years of age. It is quite plausible to suppose, for instance, that a normal child of this age who, pointing at his father coming through the front door, says, "Daddy home," is saying what it is that he believes – namely, that his father is home. Or a normal child of this age who, pointing at the doll in your hands, says, "Give me doll," or "Me, doll," is saying what it is that she wants – namely, that you give her the doll in your hands. And yet, it would be extremely difficult to understand how these children could say what they want and believe if their desires and beliefs were not conscious. We do not usually say what we want or believe if our desire or belief is unconscious. It may be true, of course, that unconscious desires and beliefs are sometime linguistically expressed through Freudian slips; but it is highly implausible to suppose that every time a normal 1- or 2-year-old child says what she wants or believes she is making such a slip.

The HOR theorist, it seems, is faced with a dilemma. He can either deny that the linguistically expressed beliefs and desires of 1- and 2-year-old children are conscious or attempt to explain their conscious status by claiming that these children are (or are disposed to be) conscious of having beliefs and desires on a regular basis. Neither option is very attractive, however. To hold that these children do not and cannot say what they consciously want or believe, but only say what they unconsciously want or believe, is rather counterintuitive.

The other option is equally unattractive. For there does not appear to be any *independent* reason (i.e., a reason that does not presuppose the truth of HOR) to suppose that normal children between 1 and 2 years of age are (or are disposed to be) conscious of having beliefs and desires. The linguistic be-

haviors of normal children at this age, for instance, do not appear to call out for an explanation in terms of the children being (or being disposed to be) so conscious. For intuitively, the only type of linguistic behavior of these children that *would* clearly require such an explanation would be their *reporting* that they have beliefs and desires. But typically, children do not begin to show signs of reporting that they have such states of mind as beliefs and desires until about their third year. This is not to say, of course, that children younger than 3 years do not use terms such as 'belief' or 'want' (or variations of) in their speech. In fact, they appear to start using such terms around 2 ½ years. But, what has been discovered by Shatz, Wellman, and Silber (1983) is that children at this age appear to use such words – if at all – not for the purpose of *reporting* that they have beliefs or desires but for the purpose of *expressing* their desires ("I want cookie") or their tentative beliefs ("It's a ball, I think"), or for certain conversational purposes, such as to gain attention ("It's a hat, you know"), or to introduce an activity ("I thoughted we'd eat some cake"). By recording and analyzing the spontaneous speech of children with their parents during mealtime and playtime, Shatz and colleagues discovered that it is not until about 3 years of age and older that children begin to show some signs of using terms like 'want' and 'believe' for the purpose of reporting that they (and others) have or are in the states of mind denoted by such terms (e.g., "I thought there wasn't any socks, but when I looked I saw them"). So, if their reporting that they have beliefs and desires is the only type of linguistic behavior in normal 1- and 2-year-olds that would clearly require an explanation in terms of their being (or being disposed to be) conscious of having beliefs and desires – an assumption that, on its face, is intuitively plausible – then, with respect to their linguistic behaviors, there does not appear to be any independent reason to suppose that normal children of this age are (or being disposed to be) conscious of having beliefs and desires.[9]

Now, in addition to the linguistic behaviors of normal 1- and 2- year-olds not appearing to require an explanation in terms of these children being (or being disposed to be) conscious of having beliefs and desires, the *non-linguistic* behaviors of these children do not appear to require such an explanation. The non-linguistic behaviors of children at this age are typically directed at manipulating objects in their environment so as to satisfy first-order desires for non-mental items, such as for a toy, to pet the dog, to get a sweet, and so on. There would appear to be little need to account for such behaviors in terms of the children being (or being disposed to be) conscious of having beliefs and desires. Their being conscious of concrete items in their external environment,

and their having first-order beliefs and first-order desires directed at such items would appear to be sufficient.

It should be made clear what *is* and what is *not* being argued for here. What is *not* being argued for is the claim that normal children of 1 or 2 years are incapable of being conscious of having beliefs and desires. For all we know, children of this age are capable of being conscious of having beliefs and desires. Rather, what *is* being argued for is the claim that we have no independent reason at the present time to suppose that normal children of this age are (or are disposed to be) so conscious, and that a representational theory that can account for conscious beliefs and desires in children of this age without attributing to them mental capacities that we have no independent reason to suppose that they have is preferable to a representational theory that cannot.

So, bringing this point together with those made earlier against FOR, the general moral seems to be that representationalists are advised to answer Q1 – Q3 in the negative, if they can. The question, of course, is whether representationalists can answer Q1 – Q3 in the negative? The answer, I believe, is that they can.

4. An alternative kind of transitive-conscious state

To see how a representationalist can answer Q1 – Q3 in the negative, an alternative type of transitive-conscious state must be introduced. Recall that FO transitive-conscious states take concrete items in a subject's external environment as their immediate intentional objects, and HO transitive-conscious states take the subject's own current intentional states as their immediate intentional objects. But, these are not the only kinds of entities that a subject can be said to be immediately conscious of. A subject can also be said to be immediately conscious of the *intentional contents* of his intentional states. Roughly speaking, the intentional content of an intentional state is the thing that determines the state's truth, satisfaction, or veridicality conditions (Searle 1983: 12). If a subject's belief, desire, or perceptual experience is true, satisfied, or veridical (respectively) if and only if *p*, then the intentional content of the subject's belief, desire, or perceptual experience is whatever it is that is expressed by the sentence '*p*' – which we can call, following conventional usage, the proposition that *p*.

With respect to one's own beliefs, one can make their intentional contents explicit by simply reporting *what* it is that one believes. If one reports that what one believes is *that the kettle is boiling*, then one has made explicit what

the intentional content of one's belief is – namely, the proposition expressed by the "that-clause," 'that the kettle is boiling.' The practice of making explicit the intentional contents of one's desires or perceptual experiences is not always as straightforward as the practice of making explicit the intentional contents of one's beliefs, though. For sometimes we use "what-phrases" of the form, 'what S is perceptually experiencing' and 'what S desires,' merely as a way to make explicit the intentional object of one's perceptual experience or desire, while leaving implicit the state's intentional content. At other times, though, we use these "what-phrases" to make explicit (or, at least, more explicit) the intentional content of one's perceptual experience or desire. If, for example, a subject says that what he is visually experiencing, or what he desires, is a ripe tomato, then the subject is using the "what-phrase" to make explicit the intentional object of his visual experience or desire, while leaving implicit the state's intentional content. However, if a subject says that what he is visually experiencing is that there is a ripe tomato on the table in front of him, or that what he desires is that he eat the ripe tomato on the table in front of him, then the subject is using the "what-phrase" to make explicit the intentional content of his visual experience or desire. For in these latter cases, the subject is making explicit the conditions under which his visual experience or desire would be veridical or satisfied, respectively. To mark the distinction between these two uses of "what-phrases," I shall use 'what_o' to identify the intentional object of a subject's intentional state and 'what_c' to identify the intentional content of the subject's intentional state.

The intentional content of an intentional state, therefore, should not be confused with the intentional object of the state. I may have a belief about the kettle on the stove, for instance; and though the kettle on the stove is the intentional object of my belief, it is not what_c I believe – it is not the thing that determines the truth conditions of my belief. For I may have many different beliefs about the kettle, all of which have different truth conditions. The intentional content of an intentional state should also not be identified with any concrete item in a subject's external environment. For the concrete items in a subject's external environment may change without necessarily affecting a change in the intentional content of a subject's intentional state. I might continue to believe that the kettle on the stove is boiling, for instance, even if the kettle, unknown to me, suddenly ceased to boil or to exist. Finally, it should also be mentioned that the intentional content of an intentional state is not the intentional state itself. The intentional content of my token belief is something to which my belief is *related*, it is not my token belief itself. My belief may be related to the same intentional content as your belief, since what_c it is that I believe may be

Table 1. Types of transitive conscious states

Type	Immediate Intentional Object
HO transitive-conscious states	Current intentional states of the subject
FO transitive-conscious states	Concrete items in the subject's external environment
SO transitive conscious states	Intentional contents of intentional states of the subject

what$_c$ it is that you believe. Bringing these points together, then, we can say that the intentional content of an intentional state is that which (1) determines the truth, satisfaction, or veridicality conditions of the state, (2) is distinct from the intentional object of the state, (3) is not a concrete item in a subject's external environment, and (4) is distinct from the intentional state itself.[10]

Having made these points of clarification, let us return to the alternative form of transitive-consciousness mentioned at the start of this section. It seems quite obvious that we are often immediately conscious of the intentional contents of (at least some of) our intentional states. This is evidenced by the fact that we can immediately (i.e., without the aid of inference or observation) report the intentional contents of (some) of our intentional states, if we so choose. If, for instance, someone were to ask me what$_c$ it is I believe about or desire for the big game tomorrow – that is, what would make my beliefs about or my desires for the big game true or satisfied – I could give an immediate answer, one that required no observation, reflection, investigation, or consultation on my part. And if this person were to go on to ask me how I determined that this is what$_c$ I believe about or desire for the big game, I would, if I thought the person was serious, say that I know that this is what$_c$ I believe or desire because I am me, and I am directly aware of what$_c$ I believe or desire. I assume that most people would answer questions of this (queer) sort in the same way. So, I assume that most people acknowledge that we are, at times, directly aware of the intentional contents of (at least some) of our intentional states.

What we seem to have, then, is a type of transitive-conscious state that is neither HO nor FO. To be immediately conscious of what$_c$ one believes, desires, or perceptually experiences is to be in a transitive-conscious state that has neither a concrete item in the external environment nor a current intentional state in one's mind as its intentional object. Given that the immediate intentional objects of such transitive-conscious states in a subject are the *same* as the intentional contents of other intentional states in the subject, we can, perhaps, call such transitive-conscious states, same-order (SO) transitive-conscious states. Table 1 illustrates the three different types of transitive-conscious states under consideration.

Table 2. Responses of representational theories to questions pertaining to GES

	FOR	HOR	SOR
Q1: Is M = M*?	Yes	No	No
Q2: Is $x = y$?	Yes	No	No
Q3: Is y = M?	No	Yes	No

Once SO transitive-conscious states are recognized, an alternative representational theory of state-consciousness presents itself, which we can call *same-order representationalism* (SOR). According to SOR, what makes a subject's belief, desire, or perceptual experience conscious is not that the subject's belief, desire, or perceptual experience itself makes the subject (or enables him to be) conscious of something (as FOR requires), or that the subject is (or is disposed to be) conscious of having the belief, desire, or perceptual experience in question (as HOR requires), but the fact that the subject is (or is disposed to be) conscious of the intentional content of his belief, desire, or perceptual experience in question – conscious *of what*$_c$ it is that he believes, desires, or perceptually experiences, not of his *believing*, *desiring*, or *perceptually experiencing* something. We can express SOR more formally as follows:

> *SOR*: Where M is a conscious intentional state that is of or about x and is possessed by a subject S, what makes M conscious is the fact that S is (or is disposed to be) in a contemporaneous transitive-conscious state M* that makes S conscious of the intentional content of M.[11,12]

As one can see from its description and from Table 2 SOR answers Q1 – Q3 above in the negative and is, therefore, formally distinct from FOR and HOR.

Of course, presenting a representational theory that is formally distinct from the two dominant representational theories is one thing, but showing that it is a plausible alternative is quite another. And so, I turn my attention to this latter project.

5. The case for SOR

If SOR is at all a plausible theory of state-consciousness, then it ought to be able to explain the distinction between those intentional states in ourselves – that is, in *normal, adult human beings* – that we intuitively take to be conscious from those that we intuitively take to be non-conscious. Does SOR carve our intuitions here at their joints? I believe so. For we appear to count as conscious those (current or recently past) intentional states in ourselves whose intentional con-

tents we can immediately (i.e., without the aid of inference or observation) report. To illustrate, consider the following three cases.

The weather case: Suppose that you currently believe that it's going to rain tomorrow. And suppose that someone asks you, "What do you believe the weather will be like tomorrow?" Whereupon, you immediately and sincerely reply, "It's going to rain tomorrow." In such a case, I believe, we would find it quite natural to take your current belief that it's going to rain tomorrow to be a conscious belief of yours.

The refrigerator case: Suppose that you believed that there is a six-pack of beer in the refrigerator, but when you open the refrigerator, you do not find a six-pack of beer in there. Suppose further that someone, upon seeing your disappointed expression, asks you, "What did you think was in the refrigerator?" Whereupon, you immediately and sincerely reply, "That there was a six-pack of beer in there." In such a case, I believe, we would find it quite natural to take your immediately past belief about the contents of the refrigerator to have been a conscious belief of yours.

The restaurant case: Finally, suppose that you are at a restaurant, and that you currently desire to have the grilled tuna for your entrée. Suppose further that your waiter asks you, "What do you want for you're entrée?" Whereupon, you immediately and sincerely reply, "I'll have the grilled tuna." Again, I believe we would find it quite natural to take your current desire in this case as a conscious desire of yours.

What these three cases, and others like them, appear to illustrate is our tendency to count as conscious those (current or recently past) intentional states in ourselves whose intentional contents we can immediately report. (It should be mentioned that the above cases are *not* intended to differentiate SO transitive-consciousness from HO transitive-consciousness. For admittedly, the subjects in the cases are not only conscious of what$_c$ they believe or desire, they are also conscious of having certain beliefs or desires. The cases are merely intended to show that there is an apparent correlation between intentional states whose intentional content we can immediately report and intentional states that are conscious.)

On the other hand, what we appear to find with respect to our judgments about *non-conscious* intentional state in ourselves is just the reverse: we appear to judge as non-conscious those (current or recently past) intentional states in ourselves whose intentional contents we *cannot* immediately report. Numerous cases can be given to illustrate this. I shall describe four that I take to be rather typical.

The Nisbett & Wilson shock case: Nisbett and Wilson (1977) describe a number of studies that researchers have taken as paradigm cases of unconscious thought and thought processes. In one such study, subjects were given a series of electric shocks of increasing intensity. Prior to the test trial, half of the subjects were given a placebo pill which, they were told, would produce symptoms such as heart palpitations, breathing irregularities, hand tremors, etc. – that is, symptoms typical of electric shocks. It was predicted that the pill subjects would form the belief that their shock-related symptoms were due to the pill (and not to the electric shock) and, as a result, would tolerate more amperage than the subjects who did not receive the pill. And this is precisely what happened. Pill subjects took four-times as much amperage as non-pill subjects. In the interview process that immediately followed the test trial, pill subjects were asked to explain why they took four-times as much amperage as other subjects; and they were asked whether the pill that they had taken earlier had anything to do with their ability to tolerate more amperage. The vast majority of pill subjects sincerely denied that the pill had any influence on their ability to tolerate more than the average amperage; and many gave confabulatory explanations for why they were able to tolerate more than the average amperage. Researchers have taken the results of these interviews as evidence that the beliefs the pill subjects had about the pill that explained their behaviors during the test trial were unconscious. What is important for us to note, however, is the fact that pill subjects were apparently unable to report the intentional contents of their beliefs about the pill that explained their behaviors during the test trial; they were, in other words, incapable of saying what$_c$ it was they (in fact) believed about the pill that explained their behaviors. Pill subjects were, of course, also apparently unable to report *that they had beliefs* about the pill that explained their behaviors; and this inability of pill subjects is certainly important. But, it is the pill subjects' inability to report what$_c$ it was they (in fact) believed about the pill, not their inability to report that they had beliefs about the pill, of which I wish the reader to take note.

Feeling-of-knowing cases: In feeling-of-knowing (FOK) cases, subjects know that they know something even though they are unable at the time to say what exactly it is that they know (Miner & Reder 1994). To illustrate, suppose that you know (and hence correctly believe) that Muhammad Ali's other name is Cassius Clay, but when someone asks you what Ali's other name is, you go blank and cannot say. Although at the time you are unable to say what Ali's other name is, you do *know that you know* (and hence correctly believe) what his other name is: you are quite sure, that is, that you have a true belief about what Ali's other name is. Should we say, though, that at this moment this belief

of yours is conscious? Intuitively, this seems wrong. It doesn't seem correct to call your belief conscious until you are able to say what$_c$ it is that you believe – that is, until you can access the intentional content of your belief. It seems intuitively inappropriate to say of you, for example, that, during the time when you cannot say what$_c$ it is you believe Ali's other name to be, your belief that Ali's other name is Cassius Clay is a conscious belief of yours.

Blindsight cases: Blindsight subjects have blind regions (scotomas) in their visual field as a result of damage to certain parts of their visual cortex. Items that fall within a blindsight subject's scotoma are treated by the subject as if they are not seen. If a flash of light is presented in a subject's scotoma, for instance, and the subject is asked whether there was anything flashed in the region of space before him, the subject will sincerely declare that there was not. However, on forced-choice test trials, subjects are able to guess correctly (from a list of options) what it was that was presented in their scotomas. These results have led researchers (Marcel 1998; Weiskrantz 1986) to hypothesize that blindsight subjects have non-conscious visual perceptions of items in their scotomas.

Subliminal perception cases: Similar sorts of results were discovered with normal subjects on visual masking studies. In one such study (Marcel 1983), a word (e.g., 'alarm') was flashed on a screen for a few milliseconds and immediately followed by a mask pattern composed of random parts of letters. When queried, subjects sincerely declared that no word was flashed on the screen. The subjects were then presented with a card with two choices of words (e.g., 'warning' and 'flesh') and were asked to guess which of the two words was similar in meaning (or graphically) to the word flashed on the screen below threshold. It was discovered that subjects typically perform well above chance on such forced-choice test trials, leading researchers to hypothesize that the subjects in the study had non-conscious perceptions of the target word's semantic (or graphic) properties.

What is relevant for our purpose here is that in both the blindsight cases and the subliminal perception cases, subjects are unable to report the intentional contents of the perceptual states that account for their behaviors in the forced-choice test trials. Although blindsight subjects appear to have perceptions of certain items in their scotomas, and subliminal-perception subjects appear to have perceptions of words flashed below threshold, and the perceptual states of both types of subjects have intentional contents, neither type of subject could immediately report the intentional content of his or her respective perceptual state. Neither the blindsight subjects nor the subliminal-perception subjects were able to say what$_c$ they were perceiving while undergoing the per-

Table 3. Conscious/non-conscious distinction in intentional states for normal, adult human subjects

Conscious	**Non-conscious**
Intentional states whose intentional contents subjects can immediately report	Intentional states whose intentional contents subjects cannot immediately report
Examples	**Examples**
The weather case, the refrigerator case, and the restaurant case	The Nisbett & Wilson shock case, the FOK cases, the blindsight case, and the subliminal perception case

ceptual states that explained their discriminatory abilities on forced-choice test trials.[13]

What the seven cases above suggest is that we intuitively count as conscious all and only those (current or recently past) intentional states in ourselves whose intentional contents we can immediately report. This intuitive distinction between conscious and non-conscious intentional states is displayed in Table 3.

Assuming that our intuitive judgments here are correct, SOR has a ready explanation for why it is that all and only those intentional states in ourselves whose intentional contents we can immediately report are conscious. The reason is simply that all and only such intentional states are *the ones whose intentional contents of which we are (or are disposed to be) immediately conscious*. Those cases in which a subject is able to immediately report the intentional content of his belief, desire, or perceptual experiences, according to SOR, are simply those cases in which the subject is (or is disposed to be) conscious of the intentional content of his belief, desire, or perceptual experience; and those cases in which a subject is not able to immediately report the intentional content of his belief, desire, or perceptual experience are simply those cases in which the subject is not (or is not disposed to be) conscious of the intentional content of his belief, desire, or perceptual experience. And where a subject is (or is disposed to be) conscious of the intentional content of his intentional state, according to SOR, his intentional state is conscious; otherwise, it is not.

The question is whether FOR and HOR can offer a more plausible explanation of our intuitive distinction between conscious and non-conscious intentional states in ourselves? I do not believe that they can. With respect to FOR, its reasonable to suppose that proponents of this theory would offer something like the following sort of explanation: the reason that all and only those intentional states whose intentional contents we can immediately report are

conscious is that all and only such intentional states are *the ones that make us (or enable us to be) immediately conscious of their intentional objects*. The problem with this explanation, however, is that not all of our conscious intentional states whose intentional contents we can immediately report make us (or enable us to be) conscious of their intentional objects. In many cases, we can immediately report the intentional contents of our conscious conative states and hallucinatory perceptions, even though, as noted in Section 3, such intentional states do not make subjects (or enable them to be) conscious of anything. The FOR explanation here may be plausible for true beliefs and veridical perceptions – intentional states that make subjects (or enable them to be) conscious of things – but it looks to be rather inadequate for conative states and hallucinatory perceptions.

The HOR theorist, on the other hand, would presumably offer something like the following sort of explanation: the reason that all and only those intentional states whose intentional contents we can immediately report are conscious is that all and only such intentional states are *the ones that we are (or are disposed to be) immediately conscious of having*. But, such an explanation, on the face of it, is deficient. For it does not appear to be the case that *all* intentional states of which we are (or are disposed to be) immediately conscious of having are intentional states whose intentional contents we can immediately report. As the FOK cases appear to show, we are sometimes immediately conscious of having a belief or desire even though we are quite incapable at the time of reporting the intentional content of the belief or desire. In the FOK case described above, for example, you are conscious of having a true belief about what Ali's other name is, even though you are not able to say what$_c$ it is that you truly believe. FOK subjects, one could say, are like individuals who are conscious of another person's preparedness to say something on some topic but are unable at the time to say what$_c$ it is the other person is prepared to say – except that in FOK cases, the 'individual' and the 'person' here are the same.[14]

In response, the HOR theorist could argue that, in order for an intentional state to be conscious, the subject not only must be (or be disposed to be) conscious of having the state, he must also be (or be disposed to be) conscious of the intentional content of the state. (Rosenthal 2000: 278, appears to argue for such a position). And that, therefore, the reason that all and only those intentional states whose intentional contents we can immediately report are conscious is that all and only such intentional states are the ones (1) *that we are* (*or are disposed to be*) *immediately conscious of having* and (2) *whose intentional contents of which we are (or are disposed to be) immediately conscious*.

Now, such a response would certainly save the HOR explanation from being deficient, but only, it seems to me, at the cost of making it more complex than is necessary. For while this HOR explanation purports to account for why all and only those intentional states whose intentional contents we can immediately report are conscious in terms of conditions (1) *and* (2), the SOR explanation outlined above purports to account for this correlation in terms of condition (2) *alone*. And so, the question naturally arises: Why endorse the more complex HOR explanation over the simpler SOR explanation? On the face of it, I can see no good reason.[15] It will not do for the HOR theorist to respond that without condition (2), a subject's intentional states will not be conscious. For such a response merely begs the question against SOR as a theory of state-consciousness. And it will not do for the HOR theorist to respond that without condition (2), a subject will be unable to be conscious of his intentional states and, thereby, be unable to think and reason about them. For such a response, although true, is irrelevant to the question at hand. (It should be noted again (see note 12) that SOR does not deny that adult human beings have HO transitive-conscious states, or that such states enable us to think and reason about our own intentional states.) The question is *not* why are all and only those intentional states whose intentional contents we can immediately report *the ones about which we are able to think and reason*. The question, rather, is why are all and only those intentional states whose intentional contents we can immediately report *conscious*. And with respect to *this* question, the SOR explanation is simpler than the HOR explanation. So, unless the SOR explanation can be shown to be deficient in some way with respect to answering *this* question, it has the advantage over the HOR explanation of being more parsimonious. This, then, is surely a place where Occam's razor should be used to adjudicate between competing explanations.

If what I have argued thus far is sound, then it would appear that of the three representational theories under consideration, SOR offers the more plausible account of our intuitive distinction between conscious and non-conscious intentional states in normal adult human beings. SOR also has the additional virtue of avoiding those problems that, as we saw in Section 3, FOR and HOR face as a result of their respective answers to Q1 – Q3. SOR avoids the problem that FOR faces with respect to explaining state-consciousness in conative states. On SOR, a subject's desire is conscious not because the desire itself makes the subject (or enables him to be) conscious of something, but because the subject is (or is disposed to be) conscious of the intentional content of his desire. And SOR avoids the problem that FOR faces with respect to explaining state-consciousness in intentional states whose intentional objects do not actually

exist. On SOR, a subject's visual hallucination of a pink elephant is conscious not because the visual state involved in the hallucination makes the subject (or enables him to be) conscious of its intentional object, but because the subject is (or is disposed to be) conscious of the intentional content of the visual state. And though there may be no pink elephant of which the subject can be conscious, there is the intentional content of his visual experience of which he can be conscious.

Finally, SOR avoids the problem that HOR faces with respect to explaining conscious beliefs and desires in normal 1- and 2-year-old children. On SOR, normal 1- and 2-year-olds need not be (or be disposed to be) conscious of having beliefs or desires in order for them to have conscious beliefs and desires; rather, these children need only be (or be disposed to be) conscious of what$_c$ they believe and desire. And the fact that normal children of this age can and often do *say* what$_c$ it is they believe and desire strongly indicates that these children are (and are disposed to be) conscious of the intentional contents of their beliefs and desires.

To illustrate, suppose that a normal 2-year-old pulls his mother in front of a high countertop on which a jar of cookies sits and, while looking intently into his mother's face, gestures vigorously toward the jar. Suppose further that, upon noticing the way that her child is behaving, the mother suddenly becomes conscious of what$_c$ it is that her child believes and wants with respect to the contents of the jar: he believes that the jar contains cookies and he wants her to give him a cookie from the jar. Now, although *the mother* in this case is conscious of what$_c$ her child believes and wants with respect to the contents of the jar, we might wonder whether *the child* too is conscious of what$_c$ he believes and wants with respect to the contents of the jar, or whether he is like a pill subject in a Nisbett and Wilson experiment who has a particular belief (or desire) about some item (the pill's effects) but is not conscious of what$_c$ it is that he believes (or desires) with respect to this item. To determine whether the child is conscious of what$_c$ he believes with respect to the contents of the jar, we might ask him, "Where are the cookies?" or "What's in the jar?" If he immediately and sincerely says, "Cookies are in jar," then this would appear to be rather good evidence that the child in this case *is* conscious of what$_c$ he believes with respect to the contents of the jar and is *not*, therefore, like the pill subjects in Nisbett and Wilson's experiments. And if the mother put her hand in the jar, pulled out a ping-pong ball, handed it to the child, and the child said, "No! Give me *cookie*," (or if she pulled out a cookie and began eating it in front of the child, and the child said, "No! Give *me* cookie"), then this would appear to be rather good evidence that the child in this case is *also* conscious

of what$_c$ he wants with respect to the contents of the jar and, again, is unlike pill subjects. So, the fact that normal 1- and 2-year-olds are able to tell us in an immediate way what$_c$ it is that they believe and want about certain things, I would argue, gives us rather good grounds for supposing that they are (or are disposed to be) conscious of what$_c$ it is that they believe and want about certain things. After all, this is precisely the sort of evidence we sometimes use to determine whether normal, adult human beings are conscious of what$_c$ it is they believe or want with respect to some item.

One might object here that a subject could not be (or be disposed to be) conscious of what$_c$ he believes or desires unless he is (or is disposed to be) conscious of having beliefs and desires; and that, therefore, a child could not be SO transitive-conscious without being HO transitive-conscious. Now, it may be the case, as this objection claims, that a subject cannot be SO transitive-conscious without being HO transitive-conscious, but such a claim is anything but obvious and is, therefore, in need of an argument. I, for one, do not find it at all inconceivable that there should be subjects who are conscious of what$_c$ they believe and desire but incapable of being conscious of having beliefs and desires. And again, given what we know about very young children – that they appear to show some signs of being conscious of what$_c$ they believe and desire but no signs of being conscious of having beliefs and desires – it would appear that this can and does occur quite frequently. To assert otherwise requires an argument.

In reply, one might argue, as I once did (Lurz 2002), that in order for a subject to be conscious of the intentional content of his token intentional state M, the subject's token intentional state M must be singled out from the many other token intentional states that he currently has, and that the only way for this to happen is for the subject to be conscious of his token intentional state M. Hence, a subject can be SO transitive-conscious only if he can be HO transitive-conscious. But, I no longer find this line of argument persuasive. I do not see why a subject's being conscious of a particular intentional state is the only way for that token intentional state to be singled out as the one whose intentional content is the intentional object of the subject's SO transitive-conscious state. For I see no reason why the question,

> Why does a subject's SO transitive-conscious state M* have as its intentional object the intentional content of the subject's token intentional state M, and not the intentional content of some other token intentional state that the subject is currently in?

cannot be reasonably answered as follows:

> Because the subject's token intentional state M, and not some other token intentional state the subject is currently in, caused (or is causally responsible for the (continued) existences of) the subject's SO transitive-conscious state M*.

It might also be argued that in order for a subject to be conscious of the intentional content of his belief or desire, he must be conscious of it *as the intentional content of his belief or desire*, which, in turn, would require that he be HO conscious of having a particular belief or desire. And if so, then SO transitive-consciousness would require HO transitive-consciousness.

Not so, though. I certainly agree that in order for a subject to be conscious of the intentional content of his belief or desire, he must be conscious of it as being something or other. But, I see no reason to hold that he must be conscious of it *as being the intentional content of some belief or desire that he has*. If the subject believes that *p*, for instance, and is conscious of the intentional content of his belief, I see no reason why the subject cannot be conscious of the intentional content of his belief simply *as being that p*. After all, the intentional content of his belief simply *is* that *p*.

If these are the only reasons for thinking that a subject could not be SO transitive-conscious of what$_c$ he believes and desires without thereby being HO transitive-conscious of having beliefs and desires, then I see no good reason to doubt that a subject could have SO transitive-conscious states without HO transitive-conscious states. And again, given what we know about normal 1- and 2-year-old children, this would appear to be what actually takes place in some cases.

6. Conclusion

I have argued that there is a type of transitive-conscious states – the SO variety – that is neither FO nor HO, and a representational theory of state-consciousness – SOR – that is neither FOR nor HOR. I have also argued that SOR avoids certain problems that FOR and HOR face as a result of their respective answers to Q1 – Q3 above; and that SOR offers a more plausible explanation of our intuitive distinction between conscious and non-conscious intentional states in normal adult human beings. If my arguments here are sound, then, contrary to what some have claimed (e.g., Carruthers this volume), SOR is the theory of state-consciousness that representationalists are advised to endorse.[16]

Notes

1. The philosophers that are typically classified as representationalist are Armstrong (1997), Carruthers (2000), Dretske (1995), Gennaro (1996), Lycan (1996), Rosenthal (1997), and Tye (1995). The term 'representationalism' is perhaps slightly misleading, since it is sometimes used to denote a particular view about the nature of propositional attitudes (conscious or otherwise) (see Fodor 1979; Field 1978). Perhaps, a less misleading name for the approach is 'the transitivity approach' or 'transitivism,' since the general idea behind it is that state-consciousness can be explained or reduced to transitive-consciousness. However, for ease of exposition, and for continuity with the (somewhat) standard nomenclature in the literature, I shall use the term 'representationalism.'

2. Phenomenal-consciousness is, of course, an important property of intentional states, and it is a property that some representational theories aim to explain (Carruthers 2000 and this volume). But, it is not this aim of representational theories with which I am concerned in this paper.

3. By 'concrete external item' I mean (roughly) any spatial or temporal object, property or condition of an object, event, or state of affairs in the subject's mind-independent environment – that is, the sorts mind-independent items that a subject can literally be said to see, feel, touch, smell, taste, or hear.

4. Since I shall be using these symbols throughout the paper, the following key may be convenient to refer to when in doubt:

> M = a conscious intentional state of a subject S (e.g., a conscious belief, desire, or perceptual experience of S)
> M* = a transitive-conscious state of S (i.e., a state of S that makes him conscious *of* something)
> x = the intentional object of M
> y = the immediate intentional object of M*.

5. Since it is not entirely clear from what Dretske writes here whether he requires that, in order for M to be conscious, it must *actually* make S conscious of x or simply *disposes* or *enable* S to be conscious of x, I have placed the dispositional alternative inside parentheses.

6. In reply, the FOR theorist could maintain that conscious conative states are conscious in a way that is different from the way that conscious cognitive states are conscious, and that, therefore, state-consciousness in conative states deserves a different explanation from the one given of state-consciousness in cognitive states. The problem with this move, however, is that there appears to be nothing to recommend it. We apply the term 'conscious' to both conative and cognitive states, and there is nothing in our use of the term that indicates that we are using it to pick out a different type of property when applying it to conative states from the type of property we pick out when applying it to cognitive states. It may be true, of course, that *what it's like* to undergo conative states is different from *what it's like* to undergo cognitive states, and that, therefore, conative states have a different phenomenal character from cognitive states. This, however, is not a reason to think that conative states have a different form of state-consciousness from cognitive states (see Section 2 above).

7. Armstrong (1963) was perhaps the first HOR theorist to argue for the distinct existence of a subject's HO transitive-conscious state M* from the subject's (lower-order) intentional state M, the intentional object of M*. And Rosenthal (1986) was perhaps the first HOR theorist to argue that a subject's conscious state M must be the immediate intentional object of the subject's HO transitive-conscious state M* and be contemporaneous with M*, if it is to be made conscious by M*.

8. There are two (fairly) well-known family squabbles among HOR theorists. One squabble is over the attitudinal status of M*. *Higher-order thought* (HOT) theorists (Rosenthal 1997; Carruthers 2000; Gennaro 1996) maintain that M* takes the form of a higher-order thought; whereas, *higher-order perception* (HOP) theorists (Armstrong 1997; Lycan 1996) maintain that M* takes the form of a higher-order perception. The other squabble is over whether the subject S must actually be in the transitive-conscious state M* or simply be disposed to be in M* in order to make S's lower-order intentional state M conscious. *Actualist* HOR theorists (most notably Rosenthal 1996) maintain the former; whereas, *dispositional* HOR theorists (most notably Carruthers 2000) maintain the latter. I mention these two types of divisions in HOR theories only to put them aside. For my interest is not with what divides HOR theorists, but with what unites them – in particular, their negative answers to Q1 and Q2 and, most important, their affirmative answer to Q3. My case against HOR is intended as a case against any of the four versions of HOR mentioned above.

9. The HOR theorist might object that what the study by Shatz and colleagues shows is *not* that 1- and 2-year-olds fail to manifest linguistic signs of being *reflexively* conscious of having beliefs and desires (i.e., what I have been calling HO transitive-consciousness), but that these children fail to manifest linguistic signs of being *introspectively* conscious of having beliefs and desires – where *introspective consciousness*, unlike mere reflexive consciousness, is a *deliberate* mental act that results in a subject becoming *conscious of his being conscious* of an intentional state that he is in (Rosenthal 1997: 745). For after all, it may be argued, these children are being asked to report on the kinds of *conscious* intentional states they are in, and their compliance with such a task would require, on the HOR model of state-consciousness, that they *deliberately* attend to the intentional states *of which they are reflexively conscious*. In reply, it should be noted first that there was no such task in the Shatz and colleagues' study. The children in the study were not asked by their parents or the psychologists to try to figure out what type of intentional states they were in and then to report what they introspectively discovered. If the children in the study were found to use the terms 'belief' or 'want' (or variations of) in their spontaneous speech with their parents to report that they believed or desired something, it was not in response to their being asked to identify the kind of intentional state they were currently in. Second, even if the act of reporting that one believes or desire something always involves introspective consciousness (which is highly doubtful), it doesn't follow that (1) the act does not *also* involve reflexive consciousness, or (2) that there are *other* types of linguistic behaviors in normal 1- and 2-year-olds *besides their reporting that they have beliefs and desires* that would clearly require an explanation in terms of these children being reflexively conscious of having beliefs and desires. Consequently, if (1) and (2) are false, then the results of Shatz and colleagues' study can be taken to show two things at once: that, based on their linguistic behaviors alone, there is no independent reason to suppose that these children are *either* reflexively *or* introspectively conscious of

having beliefs and desires. In order to maintain that the study shows only the latter disjunct, the HOT theorist must prove either (1) or (2) true. No such proof, though, is offered.

10. It may be, as some have argued, that the kinds of things that determine the conditions of truth and satisfaction for beliefs and desires (i.e., propositions) are not the kinds of things that determine the conditions of veridicality for perceptual experiences (e.g., scenarios (see Peacocke 1992)), that the contents of propositional attitudes are different from the contents of perceptual experiences. For the purposes of this paper, I need not take a stand on this issue. What is important is that the intentional content of an intentional state – whether it be a proposition, a scenario, or whatever – is neither the intentional object of the intentional state, nor a concrete item in a subject's environment, nor the intentional state itself. On such a view of intentional contents, there would appear to be little disagreement. For ease of exposition, though, I shall call the intentional contents of intentional states propositions, using 'proposition' as a placeholder for whatever it is that determines the truth, veridicality, or satisfaction conditions of intentional states.

11. There are three points that I would like to mention about my presentation of SOR here. First, SOR quantifies over intentional contents and, therefore, it is committed to their existence. There are philosophers who have argued against the existence of intentional contents (e.g., Stich 1983), but none of them to my knowledge is a representationalists. Moreover, I find the standard eliminativist's arguments against the existence of intentional contents unconvincing (see Horgan & Woodward 1990), and I find some of the standard realist's arguments in favor of the existence of intentional contents rather convincing (see Fodor 1985). Second, since intentional contents are abstracta – like numbers and uninstantiated universals – a subject cannot be said to be conscious of them by perceiving them. Therefore, SOR must hold that the attitudinal status of the transitive-conscious state M* that makes a subject conscious of the intentional content of his intentional state M is not perception-like but is thought-like. And third, I have presented SOR here in its fullest generality, eliding the difference between *actualist* versions of SOR and *dispositionalist* versions of SOR. Although I favor an actualist version – one that requires that the subject *be* conscious of the intentional content of his intentional state M in order for M to be conscious – my case for SOR over FOR and HOR does not depend on this.

12. It should be noted that SOR (and FOR) does not deny the existence of HO transitive-conscious states. In fact, it seems quite obvious that normal, adult human beings have them, and that our having them enables us to do many valuable things. What SOR (and FOR) denies is that our having HO transitive-conscious states is what *makes* our conscious intentional states *conscious*.

13. A blindsight or subliminal-perception subject could certainly guess or infer the intentional contents of the perceptual states that explained his discriminatory behaviors on forced-test trials. But, being able to guess or infer what$_c$ it is that one perceives is not the same as being able to immediately report what$_c$ it is that one perceives. And it is the absence of the latter ability in blindsight and subliminal-perception subjects that is relevant here.

14. One might object that FOK subjects are not, in fact, conscious of having a *particular* belief or desire, that if they were, they would be conscious of the intentional content of that particular belief or desire. But, I see no reason to believe this. In fact, FOK cases appear to provide rather clear counterexamples to this claim. Furthermore, if FOK subjects are not

conscious of having particular beliefs and desire (the intentional contents of which they cannot at the time access), then what are FOK subjects conscious of having when they correctly say that they know that they know something in particular? On the face of it, it certainly appears as if these subjects are reporting their knowledge of being in a particular state of knowledge (and, hence, true belief). Now, it is true that FOK subjects are not conscious of their particular (FOK) beliefs and desires *as* beliefs and desires with particular intentional contents. You are not conscious of your belief about what Ali's other name is, for example, *as* the belief that Ali's other name is Cassius Clay. But, this does not mean that you are not conscious of having that particular belief. It just means that you are not conscious of it under *that* description.

15. At one point, Rosenthal appears to suggest that there *is* some further work to be done by requiring that a subject be conscious of his intentional states as well as their intentional contents. He writes that "[i]f the way one is conscious of a state characterizes it only in terms of some content, that will not differ *subjectively* from one's being conscious only of a particular state type, rather than any individual token" (2000: 279) (emphasis added). Now, Rosenthal may be correct that what it's like subjectively to be conscious of one's intentional state *and* its intentional content is different from what it's like subjectively to be conscious *only* of the state's intentional content. But unless phenomenal-consciousness is being confused with state-consciousness, which I doubt, I fail to see how Rosenthal's point is relevant to our present concern over the nature of state-consciousness.

16. I would like to thank Rocco Gennaro, Fred Dretske, Abe Witonsky, Gene Witmer, Jonathan Adler, and Mary Jane Clarke for their extremely helpful comments on earlier versions of this paper. This chapter was supported by a grant from The City University of New York PSC-CUNY Research Award Program.

References

Armstrong, D. (1963). Is introspective knowledge incorrigible? *Philosophical Review*, 4, 417–432.

Armstrong, D. (1997). What is consciousness? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness* (pp. 721–728). Cambridge, MA: MIT Press.

Block, N. (1995). On a confusion about the function of consciousness. *Behavioral and Brain Sciences, 18*, 227–247.

Carruthers, P. (2000). *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. (this volume). HOP over FOR, HOT theory. In R. Gennaro's (Ed.), *Higher-order theories of consciousness*. Amsterdam & Philadelphia: John Benjamins.

Dretske, F. (1999). The mind's awareness of itself. *Philosophical Studies*, *95*, 1–22.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Dretske, F. (1993). Conscious experience. *Mind*, *102*, 263–283.

Field, H. (1978). Mental representation. In N. Block's (Ed.), *Readings in Philosophy of Psychology 2* (pp. 78–114). Cambridge, MA.: Harvard University Press.

Fodor, J. (1985). Banish discontent. In J. Butterfield's (Ed.), *Language, Mind, and Logic*. Cambridge: Cambridge University Press.

Fodor, J. (1978). Propositional attitudes. *The Monist*, *61*, 501–531.
Gennaro, R. (1996). *Consciousness and Self-Consciousness*. Amsterdam & Philadelphia: John Benjamins Publishing.
Horgan, T. & Woodward, J. (1990). Folk psychology is here to stay. In William Lycan (Ed.), *Mind and Cognition* (pp. 399–420). Cambridge: Basil Blackwell.
Lurz, R. (2002). Neither HOT nor COLD: An alternative account of consciousness. *Psyche*, 9, <http://psyche.cs.monash.edu.au/v9/psyche-9-01-lurz.html.>
Lurz, R. (2003). Advancing the debate between HOT and FO theories of consciousness. *Journal of Philosophical Research*, *27*, 25–46.
Lycan, W. (1996). *Consciousness and Experience*. Cambridge, MA: MIT Press.
Marcel, A. (1983). Conscious and unconscious perception: experiments on visual masking and world recognition. *Cognitive Psychology*, *15*, 197–237.
Marcel, A. (1998). Blindsight and shape perception: Deficits of visual consciousness or of visual function? *Brain*, *121*, 1565–1588.
Miner, A. & Reder, L. (1994). A new look at feeling of knowing. In J. Metcalfe & A. Shimamura's (Eds.), *Metacognition* (pp. 46–70). Cambridge, MA: MIT Press.
Nisbett, R. & Wilson, T. (1977). Telling more than we can know. *Psychological Review*, *84*, 231–259.
Peacocke, C. (1992). Scenarios, concepts, and perception. In T. Crane (Ed.), *The Contents of Experience* (pp. 105–135). Cambridge: Cambridge University Press.
Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies*, *49*, 329–359.
Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness* (pp. 729–753). Cambridge, MA: MIT Press.
Rosenthal, D. (2000). Consciousness and metacognition. In D. Sperber (Ed.), *Metarepresentations* (pp. 265–298). Oxford: Oxford University Press.
Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.
Stich, S. (1983). *From Folk Psychology to Cognitive Science*. Cambridge, MA: MIT Press.
Shatz, M. Wellman, H. M., & Silber, S. (1983). The acquisition of mental verbs: A systematic investigation of first reference to mental states. *Cognition*, *14*, 301–321.
Tye, Michael (1995). *Ten Problems of Consciousness*. Cambridge, MA.: MIT Press.
Weiskrantz, L. (1986). *Blindsight*. Oxford: Oxford University Press.

Chapter 11

A cold look at HOT theory

William Seager

1. HOT theory, introspection and concepts

To count as a higher order thought (HOT) theory of consciousness (of which there are now several versions: e.g. Carruthers (2000), Gennaro (1996), Rosenthal (1986)), a theory must of course assert that it is *thoughts* that engender consciousness. A basic problem then is the question of why possession of the ability to think about mental states should be a *necessary* condition for being conscious. Even if one conceded that only thinking beings are conscious (a controversial claim in itself) it would remain mysterious why the ability to think about *mental states* was critical. Why should being able to feel pain require being able to think about pain?

Now, being able to consciously *introspect* one's states of consciousness obviously does require the ability to think about mental states: that is simply what introspection is. It is a fundamental feature of HOT theory that basic, non-introspective consciousness is equivalent to unconscious introspection. *Conscious* introspection is then explained rather elegantly via the unconscious introspection which is basic consciousness becoming conscious via the normal HOT mechanism. But then, since introspection is normally accurate it follows that the consciousness engendering thought must itself characterize the target conscious state in some detail and categorize it as belonging to the domain of mental states. We might diagram the situation as follows:

$$\text{MS} \leftarrow \text{T[MS]} \leftarrow \text{T[T[MS]},$$

where MS is the original or first-order mental state, and T[...] is the higher-order thought forming operator responsible for generating consciousness of its target (i.e. whatever is enclosed within the square brackets).

This schema portrays someone who is consciously introspecting a mental state, say a pain. It is evident that the content of the third order state must mir-

ror the content of the second order state if the introspection is to be accurate. Since the conscious introspection in our example is of a *pain*, the notion of pain must be present in the second order state (along with much else of course, such as details of the phenomenology of the pain which are also generally accessible to introspection). Thus for the HOT theory, consciousness presupposes those 'thought-contents' necessary to think about mental states. Traditionally, such contents are, or are composed of, *concepts.* In this case, the concepts required are concepts of those mental states which can be conscious, including paradigmatically phenomenal states such as pain but also intentional mental states, such as belief, which can be consciously entertained as well as mixed phenomenal/intentional states such as emotions.

This consequence of HOT theory immediately raises a serious problem which has frequently been noted. HOT theory appears to impose far too great a burden of conceptual ability on consciousness, a burden which animals and even human infants seem unable to bear. There are two possible responses: deny that animals and infants are conscious (Peter Carruthers (e.g. Carruthers 2000) takes this route) or maintain that animals and infants have greater conceptual resources than they are usually given credit for (David Rosenthal (e.g. 1993) and Rocco Gennaro (1996) favor – in their own ways – this line).

Since neither of these options is intuitively attractive, HOT theory finds itself in a dilemma. My aim is to show that the fault lies with HOT theory itself via an examination of the probable reasons for possessing, or biological function of, both the ability to conceptualize mental states and consciousness itself. I do not aim to provide a demonstrative argument against HOT theory, but I think the weight of evidence strongly suggests that it is in serious trouble on this issue. Let's begin with thinking about minds.

2. Do animals attribute mental states?

Most animals seem completely oblivious to the mental states of others. But our own species is so far from oblivious that it can and incessantly does deploy explicit and complex mentalistic models serving overtly explanatory and predictive ends. And yet we evolved from creatures of the former category. With two extreme poles here, a natural question arises about the transition from one to the other. Was there a gradual change from creatures capable of 'basic' cognition to creatures whose cognitive abilities encompass the attribution of mental states to others? Or was there a sudden transition into a new cognitive regime that included thoughts about minds? It is important to note that I am not con-

cerned here with the *conscious* attribution of mental states. According to HOT theory, it is a precondition of consciousness that creatures possess mentalistic concepts, or, equivalently, can attribute mental states to at least themselves (we shall discuss below the link between self and other attribution). Our discussion is about this first-order ability. I am not making any claims about the ability to deploy such attributions in the formation of higher-order thoughts about mental states.

There has been a large amount of work on the question of the extent to which animals are aware of the mental states of other animals with no definitive answers yet provided. There is a fundamental philosophical difficulty here, which is that, at the level of typical animal behavior, it is in principle hard to distinguish between those responses based only upon expectation of behavior from responses based on the attribution of mental states. Such attributions represent a kind of intervening variable which in most cases seems to be a dispensable complication. Our uncertainty peaks at some unknown point as we move down the spectrum of complexity of animal behavior. Presumably there is absolutely no difficulty about denying paramecia the ability to attribute mental states and of course no difficulty in affirming it when we reach *homo sapiens*.

Let us begin nearer the unproblematic end of the spectrum. The famous dance of the honey bee is clearly a 'content carrying' activity which certainly appears to convey to other bees information about the location and quality of various items of interest (e.g. water and nectar sources, potential new hive sites, good places to find sticky tree sap, etc.). While there remains some small residual controversy about the effectiveness or even the function of the dance, there now seems to be little doubt that it does serve the function of recruiting other bees to specific ends with considerable success (see Sherman & Visscher 2002). In fact, it is hard to understand how the bee dance could have evolved unless it succeeded in transferring behavior modifying information to other bees; it would be a miracle if individual bees shared a behavioral disposition whose exercise could be strongly correlated with distant objects of intrinsic interest to bees but to which other bees were entirely indifferent, for what then could provide the selection pressure on the dance behavior?

Consider how the information transfer is effected. At the initial stage this involves tactile *perception* of the various dance parameters by spectator bees, but there is no 'fixed action' response to the dance. Honey bees are not stupid. For example, the food source version of the dance usually occurs within the pitch dark hive interior. Nonetheless, the location of the target is presented in a sun-relative way – the direction is given as an angle from the sun as viewed from the hive entrance. So the bees adopt a 'convention' and perform their dance so

that the short axis of the figure-eight dance (called the straight run) is oriented at an angle from vertical within the hive which corresponds to the angle to the target location relative to the sun at the hive entrance. But the dance is also used to indicate the direction of potential new hive sites after the bees have swarmed. This version of the dance is performed in the open, under the sun, and then the bees just 'aim' their dance directly at their target. The audience bees 'know' this and adjust their behavior appropriately. In short, it is fair to say that some cognition intervenes between the perception of the dance and the ensuing behavior.

Are the bees interacting in a Gricean fashion? Perhaps an audience bee recognizes a communicative intention in the dancing bee, knows how such intentions are conventionally encoded and recognizes that the dance is also intended to get it to recognize this intention and thus is a deliberately communicative act. Maybe, but somehow I don't think so. But this would be an interpretation which grants to the bee an ability to attribute mental states to other bees and of course it would explain how the information transfer is effected in a familiar way. It is, however, surely more plausible and more economical (in keeping with the venerable Morgan's canon) to ascribe to bees only the ability to form beliefs (or belief-like cognitive states[1]) about the location and quality of the dance's target. Bees are smart enough to figure out what the dance indicates under varying circumstances but this does not require any attribution of mental states.

Taking a big step up the continuum of complexity of behavior (though arguably bees are the more 'highly' evolved creatures), consider the behavioral clues that dogs use to signal, as we might naturally say, various states of mind. One example (see Allen & Bekoff 1997) is the so-called play-bow, in which a dog crouches on its front legs with its rear end high in the air and tail wagging vigorously. Anyone who has been around dogs at all knows that this means 'I want to play'. Although it is reasonable to regard such ritualized displays as indeed signaling a state of mind and perhaps also as an intentional offering of an invitation to social interaction, there doesn't seem to be any need to suppose that other dogs need to attribute states of mind to interpret such signals. They only need to be able to form expectations about what kind of behavior is likely to ensue after the giving of such a signal. Of course it does take a powerful mind to be able to form such expectations about future behavior which are sensitive to very subtle distinctions, such as that between 'true' and 'play' aggression, but it does not require a mind so powerful as to be able to attribute mental states which underlie and explain the expected behavior. Dogs need to be able to think that another *will* play, but not that another *wants* to play.

Nature is full of lying and deceit, and one might reasonably think that the realm of chicanery is most likely to require or at least favor the development of the cognitive abilities involved in the attribution of mental states to others. Yet deflationary accounts can be applied to almost all cases of natural deception. After all, few involve any agent cognition, as in cases of deceptive coloring such as the eye-spots on some moth's wings, or in complex structural deceptions such as the incredibly elaborate forms of certain orchids which mimic whole insects. And although many animals engage in 'intentional deception' there is little reason to think this demands any attribution of mental states to the victim. The well known injury feigning of many species of ground nesting birds is explicable in terms of behavioral expectations but does not require the bird to think about what a predator is thinking or feeling. As another example, many species of animals make alarm calls that alert others (conspecifics typically but not exclusively) to a variety of dangers (see Hauser 2000; Seyfarth & Cheney 1990). Many of these animals also use these calls deceptively, that is, produce them in the absence of the danger the call is supposed to indicate in order to gain some advantage. For example, a Peruvian jungle bird called the antshrike uses an eagle warning call to distract other birds so that it can steal insect food. As in the case of the dog's play-bow or the feigning bird, there seems to be no reason at all to suppose that the antshrike has any beliefs about the mental states of its dupes; it need only have a reliable expectation of what behavior its faked warning call will (tend to) elicit. As Hauser relates (2000: 141ff.), this is truly a pathetic case, since the dupes never learn to ignore these false calls, presumably, as Hauser conjectures, because the consequences of doing so when there actually is an eagle nearby are catastrophic. Here we see a kind of avian Pascal's wager at work but surely operating in the complete absence of any conscious awareness of the costs and benefits involved. These failures reinforce our skepticism that such creatures possess sophisticated abilities to attribute mental states.

It is not until we reach the higher primates that there is anything like a serious case supporting the ability to attribute mental states to others but even there the case is not all that strong. Dennett (1987) has pointed out that we may need to ascend to a kind of higher-order deception if we are to use deceit to make a case for the ability to attribute mental states. But it is very hard to find evidence of such super-sophisticated forms of deception in animals. Dennett discusses the MIO (Moving Into the Open) call of the vervet monkey which is made when a monkey intends to leave the forest cover and move into open ground. If the monkey's companions fail to echo the MIO call, it will not move but rather will wait a while and then utter another MIO call. Once the call is

returned the monkey will finally move. Perhaps the best interpretation of this call is that it expresses, when first uttered, something like 'I'm going to move now' with the reply call meaning 'that looks ok to me'.

Now, Dennett, Seyfarth and Cheney were intrigued by the possibility of observing a case where one monkey's MIO call is deceptively echoed by a rival monkey who knows there is danger lurking nearby and has reason to believe, or would have reason to believe if capable of such sophisticated beliefs, that the original caller does not know of the danger. This would be a case of quite sophisticated deception. No such event has ever been observed, and Dennett et al. despaired of engineering an appropriate situation. But even if the scenario envisaged was observed, would we then have a case that proved that these monkeys attribute beliefs, desires and other mental states to their fellows? It again seems possible instead to explain such deception in terms of expectations and desires about how others will behave. Perhaps the rival believes that he can get the other monkey into satisfying trouble by making the call, but only when the danger is not 'in plain view' of the other monkey.

And the notion of 'in plain view' here need not really depend upon the mentalistic concept of 'seeing', but simply upon associations between actions and objects which meet certain constraints of position relative to the other monkey. Nonetheless, the interpretation of the 'invisibility' or 'in plain view' condition remains a key point here, for why would this matter unless monkeys had some idea of what other monkeys would *believe* given the visual scene from *their* perspective? Maybe that would be the most reasonable interpretation, but it is not impossible to conceive of complex behavioral expectations based upon an appreciation of relative perspectives without an intervening attribution of belief. Some further support for such mental deflationism comes from the distressing performance of chimps in experiments explicitly designed to reveal whether or not they have a concept of 'seeing' (see Heyes 1998 for discussion). Chimps apparently fail to realize that someone with a paper bag over their head cannot acquire information about the local environment (more accurately, although chimps can quickly, and unsurprisingly, learn to discriminate a trainer's having a bag over his head from a trainer who does not with respect to reward directed behavior, there is little sign that this reflects any concept of 'sight' or 'visually induced belief').

I do not wish to denigrate the chimpanzees who are among the most intelligent animals on the planet; maybe they do deploy attributions of mental states to predict the behavior of others. Laboratory studies of captive chimps may not be the best place to discover whether these animals are able to attribute mental states to others. Christopher Boesch and Hedwige Boesch-Achermann

(2000) make this point in a striking way: "no captive study has so far attempted to study the *chimpanzee's* theory of mind, but all have confronted the chimpanzees with totally new situations to pass tests to show the *human's* theory of mind" (243, my emphasis). Perhaps, that is, chimpanzees have a distinctive conception of the mind which is, of course, normally deployed in real-world chimpanzee behavior and which would best be discovered through observation of such behavior in the field. The meticulous and long term field work of Boesch and Boesch-Achermann goes some way towards making this claim and deserves some examination here since it will in fact lead to another *difficulty* with the idea that chimpanzees can attribute mental states to others.

The behavior which Boesch and Boesch-Achermann find most compelling is that of cooperative hunting (see 238 ff.), described thus:

> . . . imagine the situation is which Darwin (a young and not very experienced hunter) and Brutus (the most experienced hunter) are in some trees with monkeys, driving them towards Kendo, who climbs a tree to block some of them. If Kendo now chases the monkeys away from Brutus and Darwin, he is hunting on his own. This situation happens at Taï with inexperienced hunters. If, however, Kendo remembers that Darwin and Brutus are in the trees from which the colobus came, he might push them back towards Darwin. This would be an error, for Darwin does not have the competence to anticipate that far ahead and would have already gone to the ground to follow the hunt. Kendo should remember where Brutus is and push the monkeys in his direction, for Brutus can anticipate such a move and wait for Kendo to chase the prey back in his direction. (Boesch & Boesch-Achermann 2000: 239)

The Taï chimpanzees hunt frequently and there is no doubt that it involves sophisticated discriminations, attributions, anticipations and advanced spatial reasoning. But it is not clear to me that it requires that a chimpanzee attribute any mental states to fellow hunters. I don't dispute that someone observing the chimpanzees hunt might have a compelling sense that they are coordinating actions via mutual awareness of each others' mental states. It might be very natural to interpret some of Kendo's actions as informed by the thought: 'Brutus *wants* me to go over to that tree now because he *intends* to drive the monkeys in that direction'. It is natural for us to make interpretations of this kind because this is of course how we ourselves often interact. But to carry over a human-style psychological interpretation of the chimpanzees' actions would be to commit the inverse of the mistake that Boesch and Boesch-Achermann warn us against.

It is not particularly difficult to re-interpret the hunting story given above so that it does not require mental state attribution. The chimpanzees are well

aware of individual differences in behavioral capability and this does not require the ability to attribute mental states. Chimpanzees are evidently very good at extrapolating the evolution of a dynamically changing situation which depends upon the actions of several creatures who are naturally acting from their own mental states (and this is true not only in hunting but even more in all the various fluid social situations crucial to chimpanzee life). If we are to put thoughts into their heads, thoughts such as 'Brutus is driving the monkeys in that direction so I should cut them off by going up *that* tree', or 'I should drive the monkeys towards Brutus rather than Darwin because Darwin is not a good hunter' suggest themselves as pretty effective hunting thoughts. These are doubtless remarkable thoughts to attribute to a chimpanzee, thoughts powerful enough to underlie cooperative hunting I would think, but which do not require any ability to attribute mental states to others.

One might well ask, how could chimpanzees have such thoughts as these if they could not attribute mental states to others? That is a good question to ask of (adult) humans but not of chimpanzees, for it is certain that even if chimpanzees do attribute mental states they do so only to predict behavior, and not to explain it, for chimpanzees simply do not engage in any kind of *explaining*. So nobody, including himself, is going to ask Brutus 'why do you think Darwin is a bad hunter'. In the absence of any explanatory role for mental state attribution, it will be very difficult to establish that they are necessary to explain chimpanzee behavior. In fact, despite their ready acceptance of the idea that chimpanzees do have a 'theory of mind' Boesch and Boesch-Achermann seem to recognize this when they say "studies of a theory of mind in chimpanzees should orientate their investigations towards attribution of competence and abilities in the physical domain of social partners. Whether some aspects of the human theory of mind such as the attribution of beliefs and emotions to others, or the intentional manipulation of such knowledge in a malicious way, are part of the chimpanzee's domain, is still open" (252). It is clear from this passage that the chimpanzee 'theory of mind' as conceived by Boesch and Boesch-Achermann does not demand any ability to attribute mental states to others, but rather only the ability to maintain a rich, diverse and dynamic database about what behavior to expect of others in an impressive variety of social and environmental situations, where much behavior reflects the mental states of other animals.

The possibility of forming such complex expectations is supported by another cognitive possibility. It seems entirely possible that animals could *track* the mental states *of* others without any ability to attribute mental states *to* others (this idea is explored in some depth in Sterelny 2000). Roughly speak-

ing, tracking involves the ability to detect something which remains present through a variety of quite distinct manifestations of and even in the temporary absence of the tracked object. For example, some tropical honey bees forage after sunset and will dance to indicate locations in the absence of the sun. They orientate the dance to the position of the sun *below* the horizon (and they do not use a visible full moon for this purpose). The audience bees correctly interpret these bizarrely oriented dances. Bees can thus *track* the position of the sun even when it is unobservable (see Gould & Gould 1995: 38). While this shows that bees have, in some sense, a 'cognitive representation' of the sun, the conclusion that bees have a *concept* of the sun or can think about the sun would hardly be justified.

In a world where appearances are constantly changing and frequently mask important features of the environment tracking is crucial. Perceptual systems can track a variety of features in the face of rapidly changing stimuli (leading to, for example, color constancy). It is obvious why this is a good idea since it is much more important to keep track of the ecologically significant properties and potentialities of things than their surface appearance. General categorization itself is a form of tracking which presumably depends upon the more basic tracking abilities of the perceptual systems. It seems plainly impossible – since viciously circular – that categorization and lower level tracking should depend upon conceptual abilities to attribute properties to objects; rather, basic categorization abilities are what underlie the more sophisticated capacities.

Now our overall question can be phrased as: is the ability to attribute mental states necessary to predict intentional behavior? The answer is clearly 'no' if it is possible to track the inner states which direct such behavior in the absence of the ability to attribute mental states to others. I have tried to present a case for just this claim. For a massively large range of animal behavioral repertoires including behaviors sensitive to the mental states of other creatures there is little or no reason to think that anything like attribution of mental states has any involvement whatsoever.

3. Consequences for HOT theory

This fact is important and substantiates our objection to HOT theory if we accept a highly plausible principle of evolution: animals don't possess unused abilities. In particular, animals do not possess *cognitive* abilities never deployed in behavior. This bald principle has to qualified in various ways. Bears, for example, can ride bicycles, a behavioral capacity never observed in the wild

(I trust). In a more purely cognitive domain, pigeons are able reliably to discriminate Monets from Picassos (Watanabe et al. 1995), a facility presumably of little use to the average wild pigeon. At least two things need to be said about this point. First, these bizarre – from mother nature's point of view – abilities are clear extensions or even instances of more general abilities which are clearly useful in these animals' normal environment. Bears frequently adopt an upright posture and thus need sophisticated balancing mechanisms which can then be pressed into service at the circus; pigeons face any number of situations calling for subtle visual discriminations and, after all, the works of Picasso and Monet are rather dissimilar. Second, and more important here, the fact that an animal can be trained into a certain behavioral or cognitive capability does not show that the animal has that ability prior to training. The worry here exploits a modal ambiguity in the term 'ability' which is, so to speak, doubly potential or suppresses a difference between first and second order versions of abilities. A first order ability is the ability to do something on demand; a second order ability is to have the ability to acquire the target ability. To say that a bear can ride a bicycle can mean (first order) that if put on a bicycle, it rides, or it may merely mean (second order) that the bear can learn to ride a bicycle.

The principle enunciated above gains still more plausibility if restricted to first order abilities (although it is clear that very few if any animals other than human beings can learn to attribute mental states to others). Bears in the wild do *not* know how to ride bicycles, wild pigeons are not able to tell a Monet from a Picasso and they would have these abilities, as first order abilities, only if these animals were in the habit of exercising them as they pursued their ordinary lives.

Since we see little or no indication that animals are using an ability to attribute mental states, it follows that the vast majority of animal species do not possess the ability to attribute mental states to others in the prediction (nor of course the explanation) of behavior.

We can also infer from this that almost all animals also lack the ability to attribute mental states to themselves because those who can self-attribute will be a subset of those who can other-attribute. It is, in the first place, of doubtful logical possibility that a being which lacked the ability to attribute mental states to others could attribute them to itself. For example, such an asymmetry would seem to run counter to Evans's (1982) 'generality condition' on concept possession (the claim that one cannot have a concept C unless for any object O, one can have the thought that 'O is C'). Note that this is not to say that any being that can attribute mental states to itself must or even will attribute them to others, but only that it must be able to. After all, the being

in question might be a solipsist or believe that it lives among zombies. What is incoherent is the notion that one could conceptualize one's own possession of mentalistic attributes while being completely unable to form a thought of another being having mental states.

But even if it might be logically possible for a creature to attribute mental states to itself while lacking the ability to attribute such states to others, it is extremely unlikely that such a being could actually exist or have evolved naturally. This is because other-attribution is much more useful than self-attribution. The primary, and considerable, benefit of the former ability is to predict what others will do in various circumstances. In contrast, creatures have little or no use for the ability to predict what they will do, since their actions are governed by their own plans and intentions.

To be a little more precise about this, we should distinguish two domains of prediction: short-term and long-term. There is great benefit in the short-term prediction of the behavior of others, and the ability to attribute mental states is constantly used by us for such predictive purposes. By contrast, there is little or no benefit from and correspondingly no use for an ability to make short-term predictions about our own behavior. In the short-term we *know* what we are going to do because we are in the process of intentionally generating that behavior. Consider haggling over the price of something. One's ability to attribute mental states to one's 'opponent' is vital; there is no place to use this ability on oneself. For short-term prediction at least it seems clear that the primary benefit of being able to attribute mental states stems from attribution to others and not to oneself (in fact, any serious attempt to self-attribute mental states is more likely to paralyze action than facilitate it). Thus it is reasonable, at least in the domain of short-term prediction, to assert that any self-attributing creature will also be an other-attributer. That is enough for the present argument.

All of the foregoing very strongly suggests, although it does not of course absolutely *prove*, that no animals other than (adult?) human beings have the ability to attribute mental states to others and based upon that little or no reason to think that any animals (other than – again – humans) have the ability to attribute mental states to themselves. So any version of HOT theory which took hold of the second horn of our original dilemma – that animals have the ability to think about mental states – becomes extremely implausible. It remains possible to grasp the first horn: perhaps hardly any animals, including even human infants, are actually conscious.

4. Tests for consciousness

So let us consider now the second mental characteristic germane to our criticism of HOT theory: consciousness. By 'consciousness' I mean phenomenal awareness, having states for which 'there is something it is like' to be in them. Here too we can look across the range of animal species and expect to find some pattern of distribution in the existence and form of their respective consciousnesses. What is important here at first is the seemingly obvious fact that consciousness is *much* more extensively distributed throughout the animal kingdom than is the ability to attribute mental states. Now, I would fooling no one if I claimed to *know, absolutely,* how to tell whether or not a certain animal, or species, enjoys consciousness. Some philosophers think that it is blindingly obvious and unquestionable that many animals are conscious. John Searle thinks that there simply is "no possibility of doubt" that dogs are conscious (2002: 62). Indeed, it is the existence of supposedly genuine philosophical worries about dog consciousness which is, for Searle, the "interesting question". It is pathological to deny consciousness to dogs and while philosophical pathologies may have some Wittgensteinian interest, they should not affect our beliefs about animal consciousness.

Other philosophers take the problem more seriously and look for behavioral evidence for consciousness in animals. Taking the problem 'seriously' evidently means *not* taking seriously the ordinary evidence that leads everyone when in non-philosophical moods to grant consciousness to animals. That is, the evidence has to somehow go beyond standard strategies of attribution, such as those based upon noting such facts as that stepping on a dog's paw causes yelping, rapid and insistent efforts to get the paw out from under one's step, paw-favoring, paw-licking, future suspiciousness by the dog about where one's foot is going, etc. All these behaviors have clear analogies with behavior caused by consciously painful events within us. Similarly, on the side of more positive emotions, when we watch a child at play with a dog, are we really supposed to think that seriousness demands that we ignore one-half of the absolutely equivalent signs of joy and pleasure emanating from both participants and grant consciousness only to the child (or should we go so far as to rethink our 'concession' that young children are conscious beings)?

Although I think the idea is almost reprehensible, the rather uncritical granting of consciousness to animals urged above is sometimes said to be 'naive' or 'unscientific'. But how there could be a genuinely scientific attitude towards the attribution of consciousness given our present state of ignorance strikes me as quite mysterious, unless all that is meant is that we ought to be

open minded about how the issue might eventually be settled. I – along with everyone else – have no objections to that innocuous percept. In any case, once the commonplace bases of attribution of consciousness are rejected, the remaining evidence that is supposed to be taken seriously tends to raise the bar on attributions of consciousness. It has to be evidence that comes as close as possible to *guaranteeing* the existence of consciousness. For example, Colin Allen and Mark Bekoff (1997: 148ff.) have suggested that one test for consciousness is the ability of an animal to correct for perceptual errors while remaining able to exploit the content of the illusory perceptual state (in just the way that a human being can know the Müller-Lyer lines are the same length even while being fully aware that a particular line *looks* longer). This would reveal an ability to 'access' appearances while at the same time in some measure discounting those appearances. If we assume that phenomenal consciousness involves a kind of access to the 'way things appear' such sensitivity to the difference between appearance and reality would provide evidence for phenomenal consciousness.

But Allen and Bekoff concede that this test will not answer the purely philosophical, or metaphysical, problem of consciousness. It will rather turn out that the existence of consciousness will provide a good (or perhaps the best) explanation of how animals are able to discount mis-leading perception while still exploiting the content of that perception. This means that at best this test is a sufficient condition for attributing consciousness to a creature and not much of a guide to what is absolutely required for consciousness.

Furthermore, the proposed test seems to involve a very sophisticated form of consciousness. As Allen and Bekoff note, what is at issue is the ability to cognize about the way things appear to one. That is, to pass the test the creature at issue must know the difference between appearance and reality, which is to say that the creature must have a notion of appearance as a kind of representational state which may or may not be accurate. Thus stated, the test seems even more stringent and not at all likely to test effectively the lower end of the spectrum of conscious beings.

What is worse, the test may in fact presuppose consciousness. This would be the case if the notion of appearance that is at issue essentially involves consciousness. To me, this is the best interpretation of the test as presented, which is then better thought of a test for the sophisticated ability to be aware of one's own awareness. However, we may suppose that there is a notion of appearance or 'potentially mis-leading perceptual information' that does not presuppose consciousness. Some such notion would be necessary if we want to allow that there can be non-conscious forms of perception – something I think most

would wish to endorse. In that case, there is room for doubt that the test is a test for consciousness at all.

Consider the following thought-experimental extension of a classic psychological experiment. Murphy and Zajonc (1993) showed that non-conscious perception of 'affectively loaded' stimuli had effects on judgment. Chinese ideographs were presented to non-Chinese reading subjects who were to decide whether the ideograph represented a 'good' or 'bad' concept. Before presentation of the ideograph either an angry or happy human face was presented for merely 4 ms. The affect of the face influenced the subject's decision about the ideograph, but without conscious awareness of the faces. Suppose now that we replicate the experiment with subjects who have been fully informed of the nature of the experiment and its outcome. I would predict that such subjects could come to have some knowledge of the affective significance of the subliminally presented faces via reflection on their feelings about the presented ideographs. This could then be used to modify their judgements about the ideographs. This, I think, would not make the faces, or even just their affective dimension, 'phenomenally conscious' to the subjects.

Many other examples could be given where knowledge of the non-consciously 'given' does not generate phenomenal consciousness. For example, blindsighters can know that they are right to guess that the target is, say, a square rather than a circle but this does not make the target phenomenally conscious. Victims of cerebral achromatopsia can know that color information underlies their ability to see shapes (which have been designed so there is no non-color basis for discrimination) but this does not make the colour phenomenally apparent to them.

So, unless we covertly assume that 'appearances' are conscious states, it will be hard to show that access to the content of mis-leading appearances will necessarily involve consciousness.

Perhaps a behavioural test of consciousness is not the best way to go in any case. We might hope instead to descend into the brain and search for the elusive neural correlates of consciousness. Unfortunately, we have at the present time very little information about what these are. Significantly, for the argument presented here, even a severely damaged brain appears to be able to sustain consciousness.

This is revealed in the most dramatic way possible in a study that assessed a set of patients diagnosed as being in a 'persistent vegetative state' and who had remained in that state (putatively) for more than six months (see Andrews et al. 1996). PVS is occasioned by severe and generally extensive brain damage and patients are left with a living body but no mind. Sleep-wake cycles are pre-

served and there is 'reflex' response to stimuli, but no consciousness. There are no neuro-diagnostic tests for PVS which can reliably distinguish it from sometimes superficially similar but radically different conditions, such as 'locked-in' syndrome. But the diagnosis of PVS is often of importance because it can form the basis for crucial decisions, such as withdrawal of feeding tubes and other life supports.

In an amazing and distressing clinical experiment, Andrews et al. found that fully 43% of the studied patients were not actually in PVS. Through exceptionally patient and creative attempts to communicate with these patients Andrews et al. discovered that subtle physical movements could be used by the patients to communicate with care-givers. One particular example is telling:

> we did not identify [B's] responses until 25 weeks after his admission [into Andrews's study], though it was obvious from subsequent conversations with him that he had not been vegetative for some time. This patient was admitted with very severe joint contractures which required surgical release and a prolonged physical management programme before he could be seated appropriately in a special seating system. Only when he was satisfactorily seated was it identified that he had a slight shoulder shrug which could be used for communication purposes. (1996: 14)

It is an obvious inference to the conclusion that other patients who unfortunately happen to be physically incapable of responding might nonetheless be fully, or partially, conscious. That consciousness can persist in the face of general and severe brain damage means that neurological tests for consciousness must await much more sophisticated knowledge about the details of how the brain implements consciousness despite or in the face of such damage.

Even more worrying from the point of view of using some form of neurological measure to assess the presence of consciousness stems from studying the behavior of victims of hydranencephaly, which is a condition, the causes of which are not well understood, where only the brainstem fully develops and the afflicted child is born more or less without a cortex. Such children can survive for some time (cases up to 19 years are known). Hydranencephaly must be distinguished from the even more severe anencephaly, a disorder of the neural tube beginning very early in development, which leads to virtually no brain development. Anencephalics do not survive long (though survival over two years has occurred). It should also be distinguished from hydrocephalus (nonetheless frequently associated with hydranencephaly) which commonly afflicts developmentally normal brains.[2]

The almost total absence of cortical structures in hydranencephalics would apparently mandate a complete absence of consciousness. What, for example, the JAMA (1995) says of anencephalics: "because anencephalics lack functioning cerebral hemispheres, they never experience any degree of consciousness. They never have thoughts, feelings, sensations, desires or emotions. There is no purposeful action, social interaction, memory, pain, or suffering" (1615), should also hold true of hydranencephalics. And yet some rather heartrending behavioral evidence suggests otherwise.

In Shewmon (1999) there is a series of reports of the behavioral capacities of some relatively long lived hydranencephalics. Although almost entirely lacking any cortical structure, these children exhibit some interesting kinds of behavior which includes discrimination amongst people and objects with affective response, affective responses to music and goal-directed activities. Taken together, these behavioral capacities strongly suggest that these children are conscious beings (though of course not self-conscious in any conceptual way). Shewmon[3] at least has no doubts: "any one of these cases suffices to disprove that all content of consciousness, including pain and suffering, is necessarily mediated by the cortex" (1999:370). Shewmon also postulates that perhaps the remaining brain tissues have, over time, reorganized themselves to support some minimal forms of consciousness. He calls this 'vertical plasticity'. It is of course what philosophers have long called 'multiple realizability', and perhaps the most striking real-world example of it in our possession.

5. The function of consciousness

In the face of our extreme ignorance about the kinds of brain architectures and neural processes which are necessary for the implementation of consciousness, there is little prospect of any sensitive brain based test for the presence of consciousness. We should turn, I think, to the function of consciousness in biological systems. Here we face two fundamental questions: 'does consciousness have a function' and, given an affirmative answer, 'is the (or a) role of consciousness to carry information about the system's own *mental* states, or is the role to carry information about the state of the environment (including the state of the system itself)'?

Leaving aside the issue of classical epiphenomenalism, consciousness could still be functionless if it was a mere evolutionary accident or side-effect of some other development (of course, once in place consciousness could take up a role in selection, perhaps being exapted for some function or other). This is

precisely the view of Carruthers (2000); it follows from the idea that the conceptual resources of the folk psychological theory of mind were selected for, and selected for the social advantages they confer, and phenomenal consciousness then arose from the application of those concepts onto certain of our own mental states according to the mechanisms of HOT theory.

We must ask here, without illicitly presupposing HOT theory, whether it seems in the slightest degree plausible that consciousness has no biological function. And, far from being plausible, this seems on the face of it ridiculous. As Dretske says: "let an animal – a gazelle, say – who is aware of prowling lions – where they are and what they are doing – compete with one who is not and the outcome is predictable" (1997:5).

It will doubtless be replied that Dretske here conflates consciousness with a possible non-conscious sense of 'awareness'. There is some superficial justice to this charge. After all, Dretske's original account of consciousness (1995) entails that swampman – Davidson's (1987) thought experimental atom-for-atom duplicate of a human being generated *de novo* – is entirely unconscious. Thus a 'swamp-gazelle' would similarly be unconscious but would evidently compete equally with conscious gazelles. However, this is not a fair complaint in the present context. Even if a non-conscious swamp-gazelle was a 'metaphysical possibility', this would not impugn the idea that consciousness serves to provide salient information about the environment to gazelles, just as the fact – if it is a fact – that swamp-gazelle's 'heart' would *not* have the function of pumping blood would not impugn the idea that this is precisely the function of the normal gazelle's heart.

If anything is evident from our own conscious experience it is that consciousness is vital for intentional action. What consciousness appears to provide is a rich field of real-time information about what is likely to be important for current and planned action. One of the most striking aspects of phenomenal consciousness is its 'evaluative valence'. There is a kind of primitive or basic desirability and undesirability of things that is a feature of the way we consciously experience them. It would seem that pain *hurts* so that we cannot ignore it (or not very easily at any rate), and similarly, *mutatis mutandis* for pleasure. One might then be forgiven for thinking that perhaps consciousness evolved as a means of presenting highly salient, often critical, information to behaviorally flexible systems possessed of high-bandwidth sensory systems. If so, we would expect to find consciousness appearing quite early in the evolution of animals in co-development with increasing behavioral and sensory capacities. For as the flexibility of behavioral responses and sensory capacity grows there is a concomitant need for information management if the

new flood of information is not to overwhelm the system. This is a version of the infamous problem of relevance (a sub-problem of which is the so-called frame problem) which still bedevils research in artificial intelligence (see Ford & Pylyshyn 1996; Dennett 1984).

Consciousness appears, from our own experience, to help mitigate the problem of relevance. Observe how your own consciousness presents objects and events as potentially worthy of attention and how important or potentially important information springs unbidden to your mind. Consciousness and attention should not be conflated however; I regard the former as something like a field of salience, onto which attention is selectively focused. Function can sometimes be highlighted by the breakdown of a system. So consider obsessive compulsive disorder (OCD), which can perhaps be regarded as a malfunction in the relevance registering function of consciousness. Victims of OCD are beset by feelings which compel them to perform (otherwise) irrational actions, such as repeated hand-washing or 'checking'. For example, one OCD suffers reports:

> When my hands were dirty, or I thought they were, they had a special feel about them. They felt huge and as though they were vibrating. I tended to hold them away from myself. Along with the sensation in my hands came a gripping feeling in my stomach. My thoughts were, 'I can't feel right until I wash my hands. I must feel right immediately or something bad will happen.' The issue shifted from the dirt on the hands to the feeling in the stomach and the vibrating feelings I felt in my hands... My head was saying, "Not dirty – no danger", but my stomach was still saying, "Danger, danger, do something quick!" (Dumont 1996)

It is significant that the experience is compelling yet at odds with active rational thought processes which are fully conscious, introspectible and even actually introspected. It is the conscious feelings, in some way independent of 'higher' levels of mind, that are more powerfully efficacious in directing behavior here. It seems very odd to suggest that such feelings cannot exist without the ability to introspect them (even if unconsciously) as states of phenomenal consciousness. Many animals engage in analogous pathological behavior, which can be induced by certain drugs (Szechtman et al. 1998). The compulsively checking rats of the Szechtman study give no indication of thinking about their plight or of recognizing that there is something peculiar about their behavior. In this, they are strikingly different from human beings, whose affliction gains an additional and much more serious degree of painfulness from introspective knowledge of the behavior and the feelings which prompt it. But on the face of it,

there seems no reason to doubt that the rats' behavior is brought about by powerful feelings of anxiety and 'restlessness' – a feeling of 'not-rightness'.

The task of presenting integrated information to a cognitive system which is evaluationally graded seems to provide a highly important, and phenomenologically plausible, job for consciousness. With a plausible functional role for consciousness in place there is no reason to deny it to creatures that give evidence of systems which fulfill that function. Here we can look at behavioral evidence from the animal kingdom. We see a huge range of animals, including even many species of insects, which have quite sophisticated behavioral repertoires coupled to highly capable sensory systems. All these animals need some way to organize sensory input, directing salient information to the cognitive systems governing behavior. So I suggest it is reasonable to regard them as all conscious though in varying degrees. Personally, I would expect that any creature whose behavior is mediated by sensory systems in a way that is not purely reflexive would be to some extent conscious. But these creatures would not have to possess even the slightest hint of self-consciousness, and very limited conceptual abilities. Certainly, hardly any of them would have any conceptual grasp of minds or mental states.

Of course, my proposal is speculative, but the point here is only to provide support for our compelling intuition that animals are conscious beings. I also want to stress again that this proposal about the function of consciousness is not intended to go, and does not go, any way towards solving the metaphysical problem of consciousness. Metaphysically, the existence of phenomenality in a purely physical world is bizarre. But since it is in the world and seems to be active in the world, its biological function may well be that of providing a relevance metric.

Coupled with the argument in the first section of this paper that animals are not, by and large, capable of thoughts about mental states, it seems fair to state that there is a grave mismatch between the possession of consciousness and the possession of thoughts about mental states. This is exactly what our ordinary intuitions about thought and consciousness would suggest. I've tried to provide some reasons to think that these intuitions ought to be respected. If so, no version of HOT theory can be correct. On the other hand, first order representational theories (such as Dretske 1995 or Tye 1995) are quite in line with our (now bolstered) intuitions and thus are to be favored at least on that score.

Notes

1. What is necessary for a cognitive system to possess 'genuine' beliefs is controversial and is a part of the same spectrum of problems here under consideration. There are philosophers who maintain that to have any beliefs at all one must have the concept of belief and hence be in a position to attribute mental states to others. See for example Davidson (1975), who goes so far as to assert that 'only a creature that can interpret speech can have the concept of a thought'. It would seem to follow fairly quickly that if Davidson is right then HOT theory implies that only language users could be conscious – an astonishing conclusion. In any event, I assume here that no matter the details of the resolution of this issue, there is a reasonable notion of cognitive states which carry information and whose effect on behavior is a function of the information carried and which operate in a way analogous to belief states, but which do not require an explicit conception of them in the cognizer.

2. Hydrocephalus can be used to bolster my point however. As noted by Gennaro (1996), there are some remarkable cases of victims of hydrocephalus living cognitively and consciously normal lives despite apparently lacking almost all – or at least a very great deal of – cortical structure (see Lorber 1980). As discussed below, these cases are perhaps explicable in terms of the multiple realizability of consciousness in a plastic brain struggling to preserve function in the face of dramatic challenges. A CAT scan and 'normal' conceptions of the brain would lead to the conclusion that these people are unconscious and incapable of thought – which is patently false. Another interesting paper which highlights the astonishingly normal adult behavior of neonatally decorticated cats is Bjursten et al. (1976). These cats, although not fully normal, give every indication of being conscious.

3. I am aware that Dr. Shewmon has a somewhat controversial advocacy role in debates over the status of those in various forms of so-called vegetative states, but I don't think this in any way impugns the data presented here.

References

Allen, C. & M. Bekoff (1997). *Species of Mind*. Cambridge, MA: MIT Press.

Andrews, K., L. Murphy, R. Munday, & C. Littlewood (1996). Misdiagnosis of the vegetative state: Retrospective study in a rehabilitation unit. *British Medical Journal, 313* (7048), 13–16.

Bjursten, L., K. Norrsell, & U. Norrsell (1976). Behavioral repertory of cats without cerebral cortex from infancy. *Experimental Brain Research, 25*, 115–130.

Boesch, C. & H. Boesch-Achermann (2000). *The Chimpanzees of the Taï Forest*. Oxford: Oxford University Press.

Carruthers, P. (2000). *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Cheney, D. & R. Seyfarth (1990). *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Davidson, D. (1975). Thought and talk. In S. Guttenplan (Ed.), *Mind and Language* (pp. 7–24). Oxford: Oxford University Press.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*, 441–458.
Dennett, D. (1984). Cognitive wheels: The frame problem of AI. In C. Hookaway (Ed.), *Minds, Machines and Evolution* (pp. 129–151). Cambridge: Cambridge University Press.
Dennett, D. (1987). Intentional systems in cognitive ethology: The 'Panglossian paradigm' defended. In Dennett's (Ed.), *The Intentional Stance*. Cambridge, MA: MIT Press.
Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.
Dretske, F. (1997). What good is consciousness? *Canadian Journal of Philosophy, 27*, 1–16.
Dumont, R. (1996). *The Sky is Falling: Understanding and Coping With Phobias, Panic, and Obsessive-Compulsive Disorders*. New York: Norton.
Evans, G. (1982). *Varieties of Reference*. Oxford: Clarendon Press.
Ford, K. & Z. Pylyshyn (Eds.) (1996). *The Robot's Dilemma Revisited*. Norwood, NT: Ablex.
Gennaro, R. (1996). *Consciousness and Self-consciousness*. Amsterdam: John Benjamins.
Gould, J. & C. Gould (1995). *The Honeybee*. New York: Scientific American Library.
Hauser, M. (2000). *Wild Minds*. New York: Henry Holt.
Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*, 101–134.
JAMA (1995). The use of anencephalic neonates as organ donors. *Journal of the American Medical Association, 273*, 1614–1618.
Lorber, J. (1980). Is your brain really necessary? *World Medicine, 3*, 21–24.
Murphy, S. & R. Zajonc (1993). Affect, cognition, and awareness – affective priming with optimal and suboptimal stimulus exposures. *Journal of Personality and Social Psychology, 64*, 723–739.
Rosenthal, David (1986). Two Concepts of Consciousness. *Philosophical Studies, 49*, 329–359.
Rosenthal, D. (1993). Thinking that one thinks. In M. Davies & G. Humphreys (Eds.), *Consciousness*. Oxford: Blackwell.
Seager, W. (1999). *Theories of Consciousness*. London: Routledge.
Searle, J. (2002). *Consciousness and Language*. Cambridge: Cambridge University Press.
Sherman, G. & P. Visscher (2002). Honeybee colonies achieve fitness through dancing. *Nature, 419*, 920–922.
Shewmon, D. (1999). Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. *Developmental Medicine and Child Neurology, 41*, 364–374.
Sterelny, K. (2000). Primate worlds. In C. Heyes & L. Huber (Eds.), *The Evolution of Cognition*. Cambridge, MA: MIT Press.
Szechtman, H., W. Sulis, & D. Eilam (1998). Quinpirole induces compulsive checking behavior in rats: a potential animal model of obsessive-compulsive disorder (OCD). *Behavioral Neuroscience, 12*, 1475–1485.
Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge: MA: MIT Press.
Watanabe, S., J. Sakamoto, & M. Wakita (1995). Pigeon's discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior, 63*, 165–174.

Chapter 12

HOT theories of consciousness

More sad tales of philosophical intuitions gone astray

Valerie Gray Hardcastle

Let me lay my cards right on the table: I think that higher ordered thought (HOT) theories of consciousness are wrong. Not only are all the individual theories proposed under this rubric wrong, but the entire approach is also fundamentally a bad idea. In this chapter, I shall explain why I believe that all this is true in some detail, though the short answer is that philosophers who rely on HOT theories have difficulty picking out genuinely empirical issues, taking them instead to be questions of intuition. And philosophical intuitions aren't worth the ink spilt over them when it comes to consciousness studies.

Champions of HOT theories are also inconsistent in their standards of good reasoning, selectively applying such criteria to hypotheses they reject, while ignoring the demands of good reason when contemplating hypotheses they like. I hope that in outlining where and how HOT theorists go astray, we shall learn how to theorize about consciousness that much better.

1. HOT theories

HOT theories assert that what makes any given mental state conscious is that attached to it is a higher order thought that one is having that very mental state (cf., Armstrong 1968, 1980; Carruthers 1989, 2000; Rosenthal 1986, 1997, 2002, forthcoming; see also Lycan 1987, 1990, 1996).[1] Let me back up a bit to explain how HOT theorists get to this *prima facie* rather odd view.

Let us suppose that we are materialists and we want to explain consciousness in as scientific a manner as possible. Furthermore, we believe that conscious mental states are not metaphysically special in any way; whatever it is

that makes a conscious state conscious is going to be some part of the universe's furniture we already know and love. So far, so good. Even though each of these assumptions has been the subject of numerous objections in the philosophical literature, they are relatively common and mild starting assumptions.

Where do we go from here? HOT theorists reason as follows. (At least, David Rosenthal reasons as follows (cf., 1986, 2002, forthcoming, see esp. 1997: 735–747).) There are two possibilities for what it could be that makes a mental state conscious: either there is something about the brain that makes these states conscious – a special neuron, a special location, a special activation pattern, or some such – or there is something about the mind. But if it is something about the brain, then we wouldn't be able to understand how it is that that thing causes consciousness (this is just Joe Levine's good old explanatory divide (1983)). In other words, it doesn't really help us to know that coincident oscillations in the brain co-vary with consciousness, for example. It doesn't help because brain oscillations are just too different from conscious mental states. We have no intellectual or theoretical bridge from one to the other. Without such a bridge, knowing about these sorts of co-variations isn't going to be explanatory, and if it isn't explanatory, then it doesn't provide a good foundation for a theory.

So: if it isn't something about the brain that makes our mental states conscious, then it must be something about the mind, since that is the only other option. But, from the HOT theorists' point of view, there are only two relevant aspects to our mental states: their phenomenal feel, which is what we want to explain, and their content, which is not what we want to explain (right now). Obviously, we are not going to be able to explain the consciousness of mental states in terms of their phenomenal feel, for to do so would be to stand on our own tail. Therefore, all that is left is the content of mental states.

Something about mental content must make some mental states conscious. It could either be something about the content of the very state that is conscious, or it could be something about the content of some other state. Since the content of conscious mental states is no different than the content of unconscious mental states, it can't be anything about the content of the conscious states that makes them conscious. It has to be something about the content of some other mental state, which is somehow then attached to the mental state it is making conscious. The only way we know that mental states "attach" themselves to anything is via their intentionality. Therefore, if the content of some other mental state is going to make a mental state conscious, then the content of that other mental state has to be directed to, be about, the mental state it is affecting. It is then just a simple skip in reasoning to conclude the content that

one is having a thought with the content of the conscious mental state is what makes that mental state conscious. Thus, we can see that a conscious mental state is "a compound state, consisting of the state one is conscious of together with a HOT" (Rosenthal 2002:416).

2. Examining the reasoning

Let us go through this chain of reasoning again, examining each point in turn. We shall find that none of them is actually justified or well-supported.

Reasoning step one: *From a materialist's point of view, there are only two possibilities for what makes a mental state conscious. Either something about the brain does it, or something about the mind does it.* Already we can see that we are off on the wrong foot. If one is going to be a genuine materialist, then of course whatever the mind is is most likely something of the brain. Buying that assertion is just what it is to be a materialist. But if the mind reduces to the brain (in admittedly some fashion that we don't understand yet, but are working very hard to), then it is problematic to set up a metaphysical dichotomy between brain and mind. I'm not saying that one can't do it. It can be done, but one has to be very careful about how one goes about making such claims.

Let us take an analogous situation. Let us suppose that we want to explain ripples in a pond. Let us further suppose that we are materialists with respect to pond ripples. Whatever makes up and causes ripples is going to be some aspect of the universe that we already know about. From our materialist point of view, there are only two possibilities for what makes the pond's surface ripple: either something about the pond does it, or something about the ripples themselves do it.

We can see the error in reasoning here. Pond ripples cannot be separated from the pond. Whatever explains pond ripples will have to do so in terms of the pond and its interactions with the world. There aren't ripples in any interesting sense apart from the pond (or other ripple substrates).

This is not to say that we can't talk about ripples *as though* they have an independent existence. We can. We can also measure them, track them, theorize over them, make predictions about their behavior, and so forth. But no one really believes that there are any ripples above and beyond the ripple's substrate. That is just what it is to be a materialist about ripples.

Minds and brains work in the same way. If we are materialists, then we have to see minds as brain ripples in a very real sense. They don't exist above and beyond their substrates, even though we might talk and act as though they do.

We might want to abstract over the brain and talk about its higher-level properties, like having content and referring to the world. We might also want to talk about these properties interacting with each other. Finally, it might also be the case that to talk in these higher-level abstractions is the most fruitful way to understand our brain's behavior or its causal patterns. If we do these things, then we might be using mind language to explain mind activities.

Ah, you say, isn't all this just much ado about nothing? Reasoning step one is shorthand for the explanatory expansion just provided. There will be two ways to explain the consciousness of certain mental states: either we will do so in lower-level brain language, or we will do so in higher-level mental language. Furthermore, we can think of the two possibilities for what causes the consciousness: either it is something at a lower level of brain organization or something at a higher level.

But this complaint isn't much ado about sloppy speech, for in recognizing that minds are really (aspects of) brain, we lose our simple dichotomy between the two. It is no longer obviously the case that either some lower-level brain property or some higher-level brain property causes some mental states to be conscious. It could be that a combination of both sorts is required. Or maybe there is some intermediate level of brain organization we haven't yet figured out how to describe that is the most causally relevant. While the possibilities for explaining consciousness don't suddenly become endless, they do become more than a simplistic either-brain-or-mind proposition.

Reasoning step two: *No brain property is going to explain consciousness because we can't bridge the conceptual divide between brain and mind. Therefore, it must be some mental property that will explain consciousness.* (NB: This is Rosenthal's argument for why he is a HOT theorist. Other – maybe all other – HOT theories don't believe this to be true.) Once again, we are confronted with a false dilemma. Philosophical intuitions tell us that brain properties are somehow fundamentally different than our mind properties (even though they are properties of the very same metaphysical object). From my perspective, this tells me to beware philosophical intuitions. Be that as it may, others apparently find this intuition appealing.

But: if there is an unbridgeable conceptual divide between neural and mental properties, which are properties of the same object, then there might be unbridgeable conceptual divides between mental and other mental proper-

ties or between neural and other neural properties. If this is just the way the brain is put together for us, then there is no reason to assume that the higher-level/lower-level divide is special. Maybe the whole darn thing is a mystery and nothing can be conceptually connected to anything else. Maybe.

In any event, consciousness is a really weird property of some mental states, and, for the life of me, I can't see why it would be more likely that something like intentionality or directedness is going to explain it than something like activation patterns in the brain. Both seem equally unlikely. (Actually, I would think that intentionality is less likely to be a player in explaining consciousness since we really don't know what that is either. At least we know what some activation patterns in the brain are.)

We know that something is making these mental states conscious. There is some causal factor out there. But we don't know what it is, not even approximately. Everything seems wrong for an explanation. It is therefore just a flat-footed error in judgment to presume that some mental property is going to explain consciousness because we can't see how any other property is going to do it. We haven't the foggiest how any property is going to do it.

Reasoning step three: *Mental states have only two relevant (higher level) properties: their feel and their content. And their feel is what we are trying to explain.* Strangely enough, it is this step that most other philosophers seem to complain about the most. They argue that their feel just is their content (e.g., Tye 1995), or that we can separate their feel from what we are trying to explain (e.g., Block 1995; Shoemaker 2001). Others have hashed this step over much better than I ever could, so I propose just to move on.

But before I do, let me point out that generally when philosophers make such claims as what we find in reasoning step three, they do so without any good supporting data. What properties mental states have is an empirical claim, no different that asking after the properties of sodium chloride. It might be the case that philosophers right now can't think of any other "relevant" properties. It might even be the case that they haven't been able to think of any others for many years. They might not ever be able to think of any. But all this points out is the limited imaginations of philosophers; it doesn't say anything deeply metaphysical.

I note in passing that many who do empirical work on the mind disagree with philosopher's claims. Some think, for example, that valence is an important dimension to all mental states, for how we assess a mental event determines what we do with it cognitively (cf., Hardcastle & Dietrich 2001, for more

discussion of this point). Maybe, just maybe, a mental state's affective intensity has something to do with whether it becomes conscious.

Reasoning step four: *Mental content is all that is left to explain consciousness with*. If we can't explain consciousness by reference to consciousness – which we can't – and if the only other properties mental states have is content, then we have to explain consciousness in virtue of that.

But, even in their discussion of conscious mental states, Rosenthal and other HOT theorists mention other potentially relevant properties. For example, one example that seems to pervade all philosophical discussion of HOT theories is zoning out while driving. It is a good example, since most of us have experienced it ourselves at some point in our locomotive lives. (It is a bad example for reasons I discuss below.) The point that Rosenthal makes is that, because we weren't paying attention when we were driving, we weren't conscious of the world around us. Similarly, because we weren't conscious of the world around us, we can't introspect the contents of our conscious thoughts. Ergo, introspectibility and ability to pay attention to are two properties of conscious mental thoughts. Rosenthal himself notes this fact: "Mental states are conscious just in case they are introspectible" (1997:745). (He then goes on to explain introspectibility in terms of 3rd order thoughts about the consciousness-making HOT, which in turn makes the HOT conscious. Notice first how he limits his theoretical scope to the contents of various mental states and second the ease and confidence he exudes in explaining what introspection is. And he does all this without so much as a single experiment to verify his approach.)

What should we make of these two putative properties of conscious mental thoughts? At the moment, we can't make much, since, despite what Rosenthal claims, we don't really know what introspection or attention amount to. Some think introspection is a relative of our sensory capacities; others think it something else entirely. Saying, as Rosenthal does, that it is a thought about a thought about a thought doesn't clear up any mysteries though, since what we want to know are how it is this 3rd order thought gets "directed" at another thought, how the thought "learns" the content of another thought, what it is about the other thoughts that allow them to be mentally inspected, and so forth. Until we understand at least some of the causal mechanisms involved, we haven't make much progress. All of these things, of course, are empirical issues, and we shall either have to run the relevant experiments ourselves (once we figure out what they are) or wait until some psychologists do before we can say with any sort of surety what the property of being introspectible means with regard to being conscious.

Attention, on the other hand, has received lots of attention, especially by cognitive psychologists. However, we still don't know very much about how it operates or why, really (cf., Hardcastle 1998). And we certainly don't understand the apparent tight link between attention and consciousness, what it is about being able to be attended to also means the state is conscious. (I should note for the record that Rosenthal too notices the connection between attention and consciousness ["Some mental states seem to be conscious only some of the time largely through shifts in attention" (1997:746).], but he doesn't appreciate the theoretical significance of such facts.)

Nevertheless, it seems to me that by investigating introspection and attention we will likely learn more about what makes some mental states conscious. Being able to introspect them and pay attention to them are two hallmarks of conscious mental states. If we could understand how we do these things, and what it is about some mental states that lets us do these things, then we would learn some important things about what consciousness really is (or at least what co-varies with it psychologically).

Reasoning step five: *The only difference between conscious and unconscious mental states is that conscious mental states are conscious. Therefore, whatever makes conscious mental states conscious can't be anything about the conscious state itself.* If you believe that just about all there is to mental states are their content and their feel, then this step more-or-less falls out of this belief. In any event, Rosenthal notes that "content will be invariant whether or not the state is conscious" (2002:416). Indeed, "it is not uncommon that a particular mental state is sometimes conscious and sometimes not" (Rosenthal 1997:744).

HOT philosophers claim to know an awful lot about our unconscious mental states. I find this strange, since even those who do a lot of empirical work on our unconscious mental states don't know very much about them. We certainly don't know how much they resemble our conscious ones, including how and whether the contents are similar. Indeed, I would hazard to guess that they don't resemble them very much at all. Below is a chart that lists some of the more obvious differences between the sorts of cognitive processes we find associated with conscious and unconscious mental states (drawn from Hardcastle 1995). This list is not exhaustive, by any means, but it does give a flavor of the sorts of differences we find in the psychological literature.

Let me highlight that the hypothesis underlying this list is diametrically opposed to what HOT theorists claim, for they believe that "whether or not a state is conscious will not affect the state's role in planning and reasoning" (Rosenthal 2002:416); that is, it will not affect the causal properties of the state. Once

Figure 1. Differences in processing between conscious and unconscious mental states (after Hardcastle 1995).

Conscious mental states	Unconscious mental states
Attentional effects	No attentional effects
Directed forgetting	No directed forgetting
Effects decay over time	Little decay over time
Levels of processing effects	No levels of processing effects
Only one semantic interpretation active at a time	Multiple interpretations active, or no interpretations active
Shows mnemonic interference	No mnemonic interference
Requires elaborative processing for storage	Requires minimal processing for storage
Little affective priming	No affective priming
Slower controlled identification	Automatic rapid identification
Few exposure duration effects	Exposure duration effects

again, we find the philosophers conveniently ignoring a wealth of empirical data to follow their own muse instead.

Of course, differences in processing do not necessarily translate into differences in structure. It is entirely possible that conscious and unconscious mental states are structurally identical to one another and what separates the two are the sorts of processing they undergo. Conscious and unconscious mental states would then just be poised differently in our cognitive economies. This option appears to be what HOT theorists presume.

This turn of events is possible, but not likely. With such a long list of orthogonal processes, it is unlikely that such huge functional differences could obtain without some differences in structure. If one were to bet, one would be wisest to bet that large and antagonistic differences in function translate into some differences in structure.

(If, on the other hand, you think that mental states are defined functionally and that their structure is largely irrelevant to psychology or cognitive science, then obviously the HOT theorists' assumption that the only difference between conscious and unconscious mental states is consciousness has to be wrong, because of all the functional differences I just listed.)

Conscious mental states aren't just unconscious mental states with an extra ingredient added; they are fundamentally different beasts. I believe that disregarding or overlooking this fact is what has tripped HOT theorists up the most. They start with the implicit assumption of a billiard ball mental economy, with all our mental states having roughly the same characteristics and all being roughly interchangeable with one another. I don't have any deep insight

into why some philosophers start reasoning from here, especially given all the psychological data to the contrary, but they do, and they are wrong in doing so.

Reasoning step six: *Mental states only interact with one another via their content. If a mental state is going to "attach" itself to another mental state, then that mental state's content must be about the state it is attaching itself to.* Let us suppose that the first sentence in this step is true. (I truly doubt that it is, but discussing that would require defining mental content and its scope, a quagmire I prefer to avoid.) Rosenthal tells us that we should think of conscious mental states as compound states: a mental state about some other event or object conjoined with the (unconscious) belief state that one is having the previous mental state. Furthermore, he means compounding in only the loosest sense of co-occurring. He explicitly denies any particular causal connection between the two states (1997).

At this point, the burden falls onto the HOT theorists to explain exactly what they think is going on here and why that results in consciousness. They now wading deep into empirical waters. If they don't follow through at this point with ways to operationalize their account, then they are open to the charge that they are explaining the mysterious by the more mysterious. The ultimate aim of HOT theories is to explain conscious mental states in such a way that the mysteriousness of consciousness is removed (or at least diminished). But if they are telling us that consciousness results from two structurally similar mental states being co-active, then this obviously does not lessen any mysteries. It simply shifts them over from the qualia to the co-activation. Furthermore, it still leaves wide open the question of why it is *this* co-activation that causes or results in consciousness.

This burden escapes Rosenthal entirely. He reasons thus: "HOTs are supposed to make the mental states they are about conscious. How can nonconscious HOTs do this? . . . HOTs confer intransitive consciousness on the mental states they are about because it is in virtue of those thoughts that we are transitively conscious of those mental states" (1997:743). And later: "HOTs result in conscious qualities because they make us *conscious of ourselves as being in certain qualitative states*" (2002:415). "The only plausible explanation is that a sensory quality's being conscious does actually consist in our having a HOT about that quality" (2002:414).

But these suggestions don't really answer the query. Why should a HOT confer consciousness, of any sort and at all? Without further development, we have no explanation. In virtue of what do HOTs make us conscious of anything? Without answering this question, then Rosenthal is either begging the

question by trying to explain consciousness in terms of consciousness or is waving his hands at the problem and calling it solved.

As discussed above, the reason Rosenthal opts for a psychological account of consciousness is that he buys into a version of the explanatory gap: he can't for the life of him see how anything neural could cause something as weird and as grand as a conscious experience. However, nothing that he has done so far reduces that very same claim against him. Why should two co-occurring mental states result in qualia? That is just as wild and weird as neuronal interactions giving rise to phenomenology.

If the point behind these explanations of consciousness is to increase its intuitive plausibility, then the HOT theorists have failed as grandly as anyone else. Every proposed solution is implausible, but we are conscious nonetheless. In my humble opinion, it makes more sense to focus on the second point – we are conscious – and worry about explaining it without feeling the need to kowtow to anyone's intuitions, educated or not. So far as I am concerned, nothing is intuitive when talking about the science of consciousness, except the fact that we are conscious, which, when you get right down to it, doesn't say very much at all.

Grand conclusion: *The content that one is having a thought with the content of the conscious mental conscious state is what makes that mental state conscious.* "When a mental state is conscious, we are conscious of being in that state; so the content of our HOT must be, roughly, that one is in that very state" (Rosenthal 2002: 409). From a purely logical point of view, this quotation borders on nonsense. How on earth could we possibly know that when a mental state is conscious, we are conscious of being in that state? It could be the case that when my mental state "seeing an orange ball" is conscious, I myself am conscious of smelling smoke in the living room, or I am conscious of nothing. Logically, there is nothing to rule out such scenarios as long as one can distinguish the self from some subset of presently occurring mental states. I wouldn't know whether my seeing-an-orange-ball mental state is conscious unless I experience it as such, but there is nothing in being a conscious mental state that guarantees I experience it as anything.

(If this discussion resembles too much science fiction and intro logic courses on a bad day, think about it in terms of Dissociative Identity Disorder. The disparate selves housed in a single body are each individually conscious of a subset of experiences of the body's existence, yet none of them is conscious of all the conscious mental states housed in the body.)

Moreover, insofar as we have no idea how a HOT might actually result in consciousness and insofar as these alleged HOTs are unconscious, so we have no first-person access to them, it is premature to conclude anything about their content. How do we know that if the seeing-an-orange-ball state is conscious that the corresponding HOT is about seeing an orange ball and not about smelling smoke? Logically speaking, it could be otherwise. It could be radically otherwise. Without embedding suggestions of what the content of our unconscious mental states are in a solid theoretical framework supported by lots of well-established empirical data, we simply can't proceed any further.

My whole point here is that much work needs to be done before we get to draw even the apparently simple conclusions that HOT theorists want. Without knowing an awful lot about what selves are and how they function in cognitive creatures, we can't make any statements about they are related to individual conscious mental states. We certainly can't conclude that a self having some HOT is going to make some other mental state conscious in virtue of the HOT co-occurring with the mental state.

Notice, however, that it is possible to refute the argument HOT theorists want to use in defense of their theory and yet the theory still be true. It still could be the case that the content that one is having a thought with the content of the conscious mental conscious state is what makes that mental state conscious, even though HOT theorists do a lousy job of arguing for its empirical adequacy.

In the next section, I examine the empirical evidence HOT theorists bring to bear in defense of their theory and suggest the sorts of experiments that would have to be done in order to give reason for us to buy some version of a HOT theory of consciousness. Not surprisingly, I argue that the evidence HOT theorists use falls far short of what they need and relies on unstable and questionable intuitions about when one is conscious, and that the evidence they do need, while probably not too hard to get, hasn't, so far as I know, been produced yet.

3. Evidence for HOT theories

It is an empirical claim that by recognizing (unconsciously) that one is conscious of something makes one conscious of exactly that something, and we need to treat it as such, looking for experimental data relevant to the claim. Philosophical intuitions should be treated very warily here, if at all.

As mentioned above, one common example that pervades the HOT discussion is the experience (or non-experience) of driving on automatic pilot. HOT theorists want to use this example in support of their approach because they believe that being on automatic pilot means that we are experiencing the world around us without being conscious of it. This is a first step in arguing that not all mental states are conscious (though who seriously believes that all mental states are conscious in this day and age is beyond me) and therefore something must be going on to make some states conscious.

HOT theorists might be right, if we knew that being on automatic pilot means that we are not conscious of our surroundings. But we don't know this. We only know that later, after the fact, we cannot report being conscious of our surroundings. There are at least two plausible explanations of this phenomenon. One is the HOT theorists', that we weren't conscious of our surroundings. The second is that we were conscious of our surroundings as we were experiencing them, but these experiences never made it into any of our memory systems, so we have no way to report later that we had such experiences.

If HOT theorists want to use this example, then they are going to have to give us reason to believe that people on automatic pilot are not just not-remembering their experiences, but that they are genuinely unconscious of their surroundings as they are experiencing them. Needless to say, HOT theorists don't give us any reasons. Nor, for that matter, has anyone else. No one knows at this point what we are conscious of and what we make memories of when we are piloting automatically. And how it seems to philosophers, after the fact, is useless in helping resolve this question.

The second empirical fact that HOT theorists rely on to support their theory is that as our conceptual structures become richer through experience and learning, we become conscious of more things in our experiences. The two favored examples are of wine-tasting and music appreciation. As one learns more about wine, then one is able to distinguish more flavors in each sip. And as one learns more about music composition, one is better able to hear the individual notes in complex performances.

Using these sorts of examples is not new in consciousness studies. I remember in graduate school class discussions about jazz performances and wondering whether the yams I didn't like as a child taste the same to me as the yams I like as an adult as my palate has matured. Just about everyone who has written much at all on consciousness has discussed these facts. The question is what to make of them.

HOT theorists want to use them to support a HOT theory. Here is Rosenthal's attempt at connecting the data with his theory:

> Learning new concepts for sensory qualities is enough for us to come to be conscious of our sensory states as having those qualities. And on that basis, we can infer than nonconscious HOTs are responsible for there being something it's like for one to be conscious of our sensory state in that way. (2002: 415)

It is quite a leap to assert that new concepts entail new conscious states and then to conclude that a HOT is responsible for the new conscious states. That is in fact too much of a leap, for there is nothing that precludes having the new concepts be part and parcel of the conscious mental state itself. (Dretske 1993; Hardcastle 1995, argue for this conclusion.) Only if one antecedently believes the HOT theory could one draw such a conclusion. The data themselves do not support it.

Indeed, the data really aren't data, for we don't really know that more or different concepts result in more or different sensory experiences. It could be that more or different concepts only give us new ways of expressing sensations that were already there and new ways of labeling them in memory. Consider, for example, one of Rosenthal's claims: "If one's HOT couldn't classify one's sensations in terms of the sound of an oboe but only that of some undifferentiated woodwind, having that sensation could not be for one like hearing an oboe" (2002: 413–414).

I have to say that I find this statement to be just false. Risking confusing an anecdote for data, let me relay my own recent personal experience. This past spring, I was driving down a country road when I heard a strange whining sound through the open window. Without recognizing the sound as anything in particular (except that it was probably animal-made), I wondered what it was. Coincidentally, at that very moment, my front-seat companion said to me, "Do you hear all those peepers?" Ah, peepers. Now I knew what was making the sound; I knew better how to label my experiences and I could tie that experience to my admittedly rather slim knowledge of frogs. Did any of this change, for me, the sound I heard? I submit it did not. I merely gained a new label "peepers' sound" for the same auditory experience (previously labeled "strange whining sound").

How would we test these claims to determine who is correct here, and, more importantly, the general claim that adopting new concepts alters conscious sensory experiences? Because linguistic and other related behavioral reports are the only paths by which we can access the conscious states of others, it will be difficult. For all we can ever do is report what we remember, and if we

need a concept in order to tag an experience in memory, then rich experiences for which we have no labels at all cannot be retrieved, reported, and therefore cannot be accessed.

At the same time, we can already see that whatever story is correct is going to be a complicated one. One difference, for example, that we might point to between the wine-tasting and the peeper examples is that in learning how to taste wine, one is learning to focus one's attention on different aspects of a complex gustatory experience, while in learning what the peepers' call is called, one is not trying to distinguish between aspects of the auditory experience. In naming the peeper experience, one is simply (metaphorically speaking) putting the largely unexamined experience in a mental box. One is not trying to learn how to distinguish the peepers' song from some other very similar amphibian noise. In doing the latter, one might very well learn to distinguish tones or sounds that one hadn't noticed before.[2]

What these examples tell us is that not all new-concept learning are created equal and what sort of learning is going on is directly relevant to how and whether one's conscious experience/memories change. They tell us that it is overly simplistic to draw a strict correlation between concepts and consciousness. Without more and better empirical investigation, we can't say much about the nature of consciousness relative to our conceptual structure of the world.

It is now time for me to put my money where my mouth is. Thus far, I have been complaining both about HOT theories and how HOT theorists try to support their claims. Let me now suggest at least some ways that one might actually test HOT hypotheses, ways that follow accepted and well-known psychological testing methods. Despite everything I have said, it is possible that some version of a HOT theory might be true after all. (That is, I might be wrong. I don't think I am, of course, but it *is* a possibility.) How might cognitive psychologists test this claim?

Experimental psychology has used sublimal priming to study how unconscious processing influences our conscious mental states since the late 1800s (e.g., Peirce & Jastrow 1884). Noting that subjects who could not report the contents of letters or numbers printed on cards held out of their visual range could nonetheless guess the characters at above chance levels, an early proponent of this view concluded that there must be "the presence within us of a secondary subwaking self that perceives things which the primary waking self is unable to get at" (Sidis 1898:171, as quoted in Merikle 1992). Getting a handle on our unconsciousness has been a cottage industry in some corner of psychology ever since.

Many cognitive psychologists advocate that we should think of our unconscious states as being merely structural in nature, and not semantic or intentional. The classic experiment associated with this perspective is Anthony Marcel's (1983a, 1983b) subliminal priming task. If two semantically related words are flashed to a subject one at a time, then the subject is faster to recognize the second word. Psychologists say that the subjects were "primed" by the meaning of the first word. For example, if we see CAT followed by DOG, then we are faster to recognize DOG as a word than if we see PIN followed by DOG. Marcel showed that if we are flashed FOOT and then PALM and then TREE, there is no priming effect for TREE, presumably because FOOT primes us to read PALM as a body part and not as a plant. However, if FOOT and PALM are flashed so quickly that we do not consciously register them, then TREE is primed. Most take this sort of evidence to support the conclusion that only the structure P-A-L-M is activated in unconscious memory and so it would then be free to prime all its associates, including TREE (as well as HAND). Interestingly enough, Marcel himself offers a different interpretation. He thinks that when words are accessed subliminally, all their semantic interpretations are activated instead of none.

Regardless, though, of exactly how sophisticated unconscious mental states are, we can exploit their tight relationship to conscious mental states to measure differential processing depending upon which unconscious mental state is active. For example, Jacoby et al. (1992) briefly presented completed words before presenting a target word stem. (They might present RESPOND first followed by ___OND.) In an opposition condition in which unconscious processing is supposed to be working against conscious processing, subjects were told not to use the completed word in suggesting a word that would complete the stem. (They could not give RESPOND as an answer to the query ___OND; instead, they would have to give something like SECOND.) However, subjects would presumably be primed unconsciously to give the flashed answer (since unconscious priming occurs apart from the specific context), even though they were instructed to disregard that information. As a result, subjects take longer to answer the queries for which they had just been primed with an answer that they could not use.

In an in-concert condition in which unconscious and conscious processing are supposed to be working together, subjects were told to use the completed word. (They were supposed to give the answer RESPOND to the query ___OND.) In this case, the unconscious priming would work to the advantage of the subject, for they would be more ready to give the flashed word as they correct answer. Their reaction times should be shorter. By comparing how

long it takes subjects to respond under the opposition condition with how long it takes them to respond under the in-concert condition and with how long it takes them to respond in purely neutral conditions (where the completed word does not match the stem), we can get an idea of the influence unconscious perception has on the conscious production of the answers.

We can run similar experiments using error rates as a variable. Error rates generally track reaction times, so subjects would make fewer errors under the in-concert condition than they would under the opposition condition. If we compare error rates from these two conditions with what happens under neutral conditions, then we can again get an idea of what sort of influence unconscious priming has on conscious perception and behavior.

We can and should use a similar methodology to determine whether we have unconscious HOTs of the sort HOT theorists propose co-active with any conscious states. If what we know about subliminal priming is true, unconsciously being aware that you are aware of something should prime other thoughts. More specifically, we need a priming task that would test whether we can recognize that we were aware of a series of target conscious events faster or with fewer errors than other aspects of the same events. If we can, then that would be some evidence that we are unconsciously aware that we are aware. If we cannot, then that would be some evidence that no such HOTs co-occur with any regularity with our conscious mental states.

Here is one way such an experiment might go. We flash a series of simple scenes (a cat on a mat, say, or a dog with a bone) for half a second or so, long enough for subjects to see them consciously, and then replace each scene with the same masking stimulus (a set of hash marks or some such). Stimulus masks allow subjects to perceive a target stimulus, but prevent them from studying it. After each scene, when the mask is in place, we then query the subject with one of two sorts of questions. Either we ask about the fact of their conscious experience, e.g., did you see a cat? did you see a bone?; or we ask about the structure of the actual scene: was the cat on the mat? was the dog next to the bone? Half of the questions we ask in each category should actually reflect the previously viewed scene and half should not reflect any of the scenes viewed in the experiment. The types of questions asked after each scene should be randomly distributed across the stimuli set. If we indeed have unconscious HOTs accompanying all conscious experiences, then that HOT should prime our behavior in regard to reacting to the fact that we are conscious, and we should answer queries in category (1) with fewer errors, on average, than we do those in category (2). (We should also answer category (1) questions faster on aver-

age, but reaction times are hard to measure under these conditions, since the time used in asking the question can be highly variable.)

It is these sorts of experiments – experiments that build upon a wealth of accepted empirical research – that will be most telling regarding any proposals about consciousness. It simply unacceptable to use anecdotal accounts of one's own or of someone else's experiences as genuine support for any hypothesis. We are not going to get a serious theory of consciousness through considering first principles or our intuitions. Science just doesn't work that way. And thank goodness for that!

Notes

1. We can make finer distinctions among the players – some advocate higher order perception theories and others more straightforwardly higher order thought theories. The question arises whether there are any crucial differences between the views (see, e.g., Güzeldere 2000). What I have to say doesn't address this issue and I am not concerned with any potential divide in sorts of higher order mental state theories of consciousness.

2. We can't run the equation in the other direction either and claim that "being able to form intentional states about certain sensory qualities must somehow result in our being able to experience those qualities consciously" (Rosenthal 2002b: 414). Dianne Raffman (1993) has clearly demonstrated that while we can have concepts referring to various closely related tones in music, we cannot physically distinguish among them. Indeed, people with sensory deficits who nonetheless have concepts referring to aspects of the missing sensory experiences have intentional states about certain sensory qualities, but they do not consciously experience them.

References

Armstrong, D. (1968). *A materialist theory of mind*. New York: Humanities Press.

Armstrong, D. (1980). *The nature of mind and other essays*. Ithaca, NY: Cornell University Press.

Block, N. (1995). On a confusion about a function of consciousness, *Behavioral and Brain Sciences, 18*, 227–247.

Carruthers, P. (1989). Brute experience. *Journal of Philosophy, 86*, 258–269.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. New York: Cambridge University Press.

Dretske, F. (1993). Conscious experience. *Mind, 102*, 263–283.

Hardcastle, V. G. (1995). *Locating consciousness*. Amsterdam: John Benjamins Press.

Hardcastle, V. G. (1998). The puzzle of attention, the importance of metaphors. *Philosophical Psychology, 11*, 331–352.

Hardcastle, V. G. & E. Dietrich (2001). Toward a book of counter-examples in cognitive science: Dynamic systems, emotions, and aardvarks. *Danish Yearbook of Philosophy*, *36*, 35–48.

Jacoby, L. L., D. S. Lindsay, D. S., & J. P. Toth (1992). Unconscious influences revealed: Attention, awareness, and control. *American Psychologist, 47*, 802–809.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly, 64*, 354–361.

Lycan, W. G. (1987). *Consciousness*. Cambridge, MA: The MIT Press.

Lycan, W. G. (1990). Consciousness as internal monitoring, I. In. J. Tomberlin (Ed.), *Philosophical perspectives, Vol. 9* (pp. 1–14). Atascadero, CA: Ridgeview Publishing Company.

Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: The MIT Press.

Marcel, A. J. (1983a). Consciousness, masking, and word recognition. *Cognitive Psychology, 15*, 198–237.

Marcel, A. J. (1983b). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15*, 238–300.

Merikle, P. M. (1992). Perception without awareness: Critical issues. *American Psychologist, 47*, 792–795.

Peirce, C. S. & J. Jastrow (1884). On small differences in sensation. *Memoirs of the National Academy of Sciences, 3*, 73–83.

Raffman, D. (1993). *Language, Music, and Mind*. Cambridge, MA: The MIT Press.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 94*, 329–359.

Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–753). Cambridge, MA: The MIT Press.

Rosenthal, D. (2002). Explaining consciousness. In D. J. Chalmers (Ed.), *Philosophy of mind: Classical and contemporary readings* (pp. 406–421). New York: Oxford University Press.

Rosenthal, D. (forthcoming). *Consciousness and the mind*. New York: Oxford University Press.

Shoemaker, S. (2001). Introspection and phenomenal character, *Philosophical Topics*.

Sidis, B. (1898). *The psychology of suggestion*. New York: Appleton.

Tye, M. (1995). *Ten problems of consciousness: A representational theory of the phenomenal mind*. Cambridge, MA: The MIT Press.

Chapter 13

A few thoughts too many?

William S. Robinson

1. Background

There are many occasions on which we express in subvocal speech something we have realized for the first time. For example, perhaps I have been observing a conversation between Jones and Smith, and have intermittently attended to Smith's movements, smiles, frowns, and so on. Suddenly, I find myself subvocally saying "Smith is really not enjoying this conversation with Jones!" Or, perhaps I am silently counting prime numbers, and when I come to 17, I say to myself "17 is the 7th prime – now, there is food for thought for someone taken with numerology".

These cases are interesting because they appear to offer two instances of consciousness: the thought expressed in subvocal speech is a conscious thought, and the auditory imagery in which it is expressed is conscious imagery. One response that I will discuss here is a "complex view", according to which each instance is conscious because of a (different) higher order representation (HOR). I shall argue that a simpler view that refuses a HOR for the thought seems adequate and, if so, should be preferred on grounds of parsimony. I shall consider two yet more parsimonious views, and argue that they are not clearly rejectable.

In one way, my aims here are modest – little is held to be firmly established. Nonetheless, the open possibility of views that are simpler than HOR theories, and not less explanatory of consciousness than they are, should raise a significant doubt about the claim that HORs are elements required for consciousness.

Not everyone reports thinking subvocally, and readers who do not will have to take the existence of cases of the kind described in the first paragraph on trust. But the phenomenon of subvocal speech is well recognized, and there is no serious question that we may take it as a background assumption that

(1) Subvocal speech exists.

Two assumptions may be regarded as somewhat more controversial. These are:

(2) The cases described are cases in which the subject has had a conscious thought.

(3) The auditory imagery of which subvocal speech is composed is, generally, conscious, quasi-sensory experience.

I am not going to attempt to mount a serious argument for these assumptions, because I think they come about as close to being data as any claims we are likely to find in discussions of consciousness. It is not easy to think of something more evident than these, from which they might be derived. A few words of explanation are, however, in order.

(2) might be questioned on the ground that a subvocal statement could be made and subsequently completely forgotten. Moreover, it might be forgotten before it or its causes had any other effect on the cognitive system. In such a case, one might be moved to say that Smith's irritation with Jones *almost* emerged into consciousness. One might regard the subvocalization as a symptom that typically indicates a conscious thought but that could fail to be such an indicator if further connection with the cognitive system were not to be forthcoming.

There are two responses to this suggestion. First, I think it would be preferable to say that Smith's irritation with Jones did emerge into consciousness for a moment. There is some experimental support for the view that items of which we were once conscious can be quickly forgotten (e.g., Wolfe 1999; Simons & Levin 1997; Muter 1980). It seems tendentious to rule out momentary conscious thought, i.e., to require conscious thoughts to be remembered.

But, second, for present purposes we need not try to resolve this matter decisively. We can simply stipulate that we are to consider *ordinary* cases of subvocal speech. In ordinary cases, if I subvocalize a statement, its content is available to my cognitive system; if contexts that arise soon afterward make the statement relevant, I will be able to adjust my behavior or draw inferences in an appropriate way. In such ordinary cases, there will be little temptation to say that I was not conscious of thinking that, e.g., Smith was irritated with Jones, until I made later use of that content. Suppose, for example, that later in the same week, someone relates some gossip about a break between Smith and Jones. I might say "I'm not surprised. I saw them a few days ago, and I remember thinking at the time that Smith was not enjoying his conversation

with Jones". It would seem ill motivated to hold that the thought could not have been conscious at the time, and becomes so only when it is remembered.

(3) claims that auditory imagery in subvocal speech is quasi-sensory experience. It is not sensory, because it is not heard. No eardrums are vibrated, no inner-ear cilia are stimulated, and the vividness of actual hearing is lacking. But auditory imagery is in certain respects like auditory experience. For example, it makes sense to say of it that it had the accent that matches that of my normal overt speech. Or, perhaps, that it did not: after all, on a just slightly different kind of occasion, I might have been in an odd mood, and subvocally expressed myself in a somewhat exaggerated tone – "Soooo, that Smith, he reeaally isn't enjoying Jones, is he?"[1]

(3) also claims that the auditory imagery in subvocal speech is generally conscious. This claim is compatible with allowing that there can be language-like representations of thought contents that are unconscious. However, when people say they talk to themselves when they think, they do not normally regard themselves as inferring language-like representations from some theory. Instead, they regard themselves as giving reports about occurrences they consciously experienced, that are not observable to others, and that others would not normally know about unless they were told about them.

It is possible to introspect one's subvocal speech, attending to its accent, for example, while one is saying it to oneself. But the cases I am considering here are ordinary cases in which one's subvocal speech is conscious, but is not the focus of special attention. Just as we usually hear ourselves speaking out loud without attending to our speech, we usually have our auditory imagery consciously, but without attending to it as such (although, of course, we *can* attend to both overt and subvocal speech).

It may be that some (or even most) people engage in silent soliloquy when asleep. I incline toward the views that such events, if they occur, are conscious, and that they are like dreams in this respect. These views, however, are not presupposed here. Assumption (3) is, by stipulation, concerned with ordinary cases that occur when we are awake.

2. The issue

With this background, I can now state the central question to be addressed here. The foregoing assumptions apparently imply the existence of two different cases of something's being conscious. There is a thought that is conscious, and there is a quasi-sensory experience that is conscious. The question is how

these two instances of consciousness are related. The interest of this question is that, as we shall see, it introduces difficulties for the plausibility of higher order accounts of consciousness.

Before turning to these difficulties, let me explain why I say there are distinct instances of consciousness. In having conscious auditory imagery, we have, as noted, something of which it makes sense to say that it had one accent or another. But such predicates do not apply to the thought that Smith is not enjoying his conversation with Jones. On the other hand, the thought (conscious or not) is true or false; but it makes no sense to say of my auditory imagery that it is true or false. The imagery's being conscious is thus a different fact from the conscious status of the thought that the words of the imagery may express.[2]

3. The complex account

HO theories hold that a mental state is conscious if and only if that mental state is the intentional object of an appropriate higher order representation.[3] So, if a thought is conscious, it exists and is also represented by a HOR, and if auditory imagery is conscious, it exists and is represented by a HOR. According to what I shall call "the complex account", the two conscious mental states (the thought, and the imagery) are represented by different HORs. Let us say that HOR1 is the higher order representation of the thought, *T* (e.g., the thought that Smith is not enjoying his conversation with Jones) and that HOR2 is the higher order representation of the auditory imagery, *I*. *T* is a conscious thought, because it is represented by HOR1, and *I* is conscious imagery because it is represented by HOR2.

The pictorial representation in Figure 1 will aid in understanding the complex account. In order to be able to interpret it correctly, let us first note that there is, presumably, an interesting relation between *T* and *I*. Assuming normal cognitive organization, HO theories will hold that *I* will match *T* in such a way

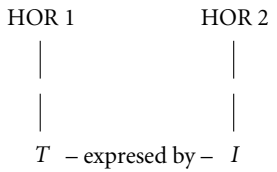

Figure 1.

that the words in *I* will be the normal expression, in the subject's language, of the content of *T*. Let us abbreviate this point by saying: *I expresses T*.

The complex account is a noncontradictory view. I shall argue, however, that it is an unattractive theory, because it is unparsimonious. In postulating HOR1, it commits itself to at least one too many higher order representations – one more, that is, than is needed in an alternative account that is at least as plausible as the complex account. Naturally, this point cannot be clear until we have seen a plausible alternative. So, I shall now proceed to describe an alternative view, and then return to explain the allegation of lack of parsimony in the complex view.

4. The simpler account

What I shall call "the simpler account" says that *T* is a conscious thought if and only if it gets expressed. I mean to allow overt speech to be one way of expressing a thought, but expression in subvocal speech is sufficient. I take it that one is conscious of one's overt speech, so whether a thought is expressed overtly or subvocally, its expression will be conscious. However, since I am focusing here upon the case of subvocal speech, I will be primarily concerned with that case in what follows.

In the pictorial representation of the simpler account in Figure 2, there is no HOR1, and the higher order thought about *I* is labeled with a "2" only because it is not supposed to be any different from the HOR labeled "2" in the previous diagram. The simpler account holds that what is in Figure 2 is sufficient for *T* to be conscious. This view is not to be taken to imply that *T* cannot be thought of, i.e., that there cannot be a HOR that is of *T*. But such a representation is not required in order for *T* to be conscious, and if there should happen to be one, it is not in virtue of that fact that *T* is conscious.

As before, I take it that if I actually express *T* (whether overtly or subvocally), it would normally be available to my cognitive system. Failures of such

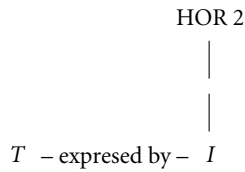

Figure 2.

availability could be treated as failures of consciousness, or (more plausibly, in my view) failures of conscious thoughts to be able to play their normal role. Preference for one or the other of these views should depend on what it is best to say about the normal case, and so I shall proceed to that discussion.

I take the comparative simplicity of the simpler view to be a ground for preferring it, providing it is adequate to the phenomena. I hold that the simpler view is adequate to the phenomena because I think that remembering having said something to oneself subvocally is sufficient for having had the thought consciously (in the normal cases on which I am focusing). If I remember saying "p" to myself, that seems to settle the question of whether I consciously thought that *p*. Perhaps I think I should not be blamed for actions that flow from unconscious intentions. But if I remember having said "I'm going to make life difficult for that so-and-so", I cannot excuse my act of revenge on the ground that I had no conscious intention. In considering a case of this kind, I have no sense that the question of consciousness of my intention rests on whether there was a further representation of my intention. Or, suppose I am inclined to claim credit for having thought of a solution to a problem. Perhaps others suggest that I only implicitly possessed the solution. Perhaps they even allow that I arrived at some results that seem to depend on the solution, but they think that those results were due only to the unconscious effect of background beliefs that *could* have led to a conscious recognition of the solution, but did not. Now, I may not be able to convince them otherwise, but if I remember having said the solution to myself, I will be justified in believing that I consciously thought of the solution. Again, in a case of this kind, I have no sense that whether I consciously thought of the solution depends on whether there was a further thought about the thought of the solution.

Does the possibility of introspection show that there are always HORs of our conscious thoughts – HORs that *can* become conscious, even though they are not normally conscious? This view is evidently consistent, but it is not needed. On the simpler view, introspection (when it does occur) consists in our saying to ourselves sentences that are tantamount to "I think that *p*".

If these remarks are accepted, then there is pressure against the complex view, on the ground of lack of parsimony. HOR1, it seems, can be lopped off without losing an adequate account of conscious thought. Unconscious thought can be understood on the simpler view: it is thought that does not receive any expression, but that may have other effects on the cognitive system.

The complex view might be rescued if one could show that there would not have been an expression of *T* unless there had been a higher order representation of it (i.e., if HOR1 had not been present). I do not think our present

knowledge of cognitive systems permits a decisive refutation of this idea, but I do not think it would be good scientific practice to accept the view. For it is also true that there is nothing in our present knowledge of cognitive systems that suggests that (or explains why) there should have to be a representation of a thought in order for that thought to be expressed. According to HO theories, HOR1 would not, in general, be a conscious thought. Therefore, if HOR1 were held to be required for *T* to be expressed, it would be implied that unconscious thoughts can be efficacious. (A HOR would be a required part of the only available sufficient condition of a subvocal or overt expressing of *T*.) But if that is allowed, there would seem to be no ground for denying that *T* can be efficacious (whether it is regarded as conscious or not). If that much is allowed, there seems to be no ground for denying that *T* itself could be a state of cognitive organization that can affect the language production centers in such a way as to produce a (subvocal or overt) expression of its content. Thus, once again, the complex view lacks parsimony; it seems to postulate a higher order representation that is not required for any job that we have good reason to suppose is done (unless, of course, we are already assuming a higher order representation theory).

These considerations seem to me to establish a reasonable suspicion that the simpler view is preferable to the complex view. In the remainder of my remarks, I will assume this preferability, and attempt to follow out where we may be led if we accept it.

A consequence of this assumption is that there now seems to be some mild pressure against the right side of our diagram. The right side, let us recall, says that the consciousness of my auditory imagery consists in its being represented by a higher order representation – the HOR2 of the diagram. Now, the view that sensory consciousness requires higher order representations is sometimes thought to be supported by analogy with conscious thoughts. That is, it is sometimes thought that HO theories are plausible *at least* as accounts of what makes beliefs and desires conscious, and then it is reasoned that if HO representation accounts for the consciousness of beliefs and desires, it would be natural that HO representation would likewise account for the consciousness of sensory states. If this is accepted, and if it is further accepted that auditory imagery is like sensory consciousness, it would seem natural to extend the HO theory to auditory imagery, and say that it is conscious only if it occurs and is represented by a higher order representation.

But on the simpler view, the consciousness of a thought does not consist in its being represented by a higher order representation. It consists in its being expressed by a conscious event, i.e., either an overt speaking or a subvocalization.

On this view, therefore, there is no model in the case of conscious thoughts that suggests that consciousness consists in being represented by a higher order representation.

Evidently, none of this is a refutation of the view that the consciousness of auditory imagery consists in that imagery being represented – that is why I said that the preferability of the simpler view puts only mild pressure on the right side of Figure 2. The consequence I have been alleging is only that to adopt the simpler view is to give up a parallel model that may be thought to support a HO theory of the consciousness of auditory imagery (and, more generally, of sensory consciousness).

5. The very simple view

What I shall call "the very simple view" gives up the HO account for consciousness of auditory imagery. Such imagery, represented by *I* in Figure 3, is taken to be conscious in its own right. It is not something that is conscious only in virtue of something else that represents it. As in the simpler view, *T* is conscious if it is expressed by a conscious occurrence. This conscious occurrence will be auditory imagery in the focal cases for this paper, but, as already noted, *T* will also be conscious if it is expressed in overt speech.

The very simple view evidently allows for unconscious thoughts – these are thoughts that may have effects on our cognitive systems, but fail to have their content expressed by a conscious occurrence. But it does not allow for a distinction between conscious and unconscious auditory imagery, for it takes imagery to be conscious as a matter of its intrinsic nature. In my view, this result is just as it should be, for I do not think we have reason to introduce unconscious auditory imagery.

We cannot rule out that unconscious thinking may involve our language centers.[4] It is possible that there is a great deal of unconscious accessing of words. It is plausible that grammatical structure is accessed and processed in unconscious thought, and operations on grammatical structure would be as unconscious in unconscious thought as they routinely are when we are consciously thinking. Perhaps there are large stretches of unconscious thinking

T – expresed by – *I*

Figure 3.

that have the same time course as overt speech. But allowing all this does not imply anything imagistic; it does not imply that there is anything auditory-like about unconscious accessing of words. Such thinking can be regarded as thought *without* imagery. This is not to say that I can derive a contradiction from the postulation of unconscious imagery. But there is no visible need for such a postulation, and making it seems unparsimonious.

6. Burdens of proof

I expect HO theorists to oppose the remarks of the last two paragraphs. On their view, the consciousness of conscious imagery consists in the possession of qualities by an image (which, just by itself, is not an instance of consciousness), plus the occurrence of a higher order representation of that image. The latter component of this pair can be described a little more fully and accurately as a higher order representation of the auditorily imagistic properties of the image. When we have conscious auditory imagery, we are not conscious of neural properties of events with which auditory imagery may be supposed to be identical, and we are conscious of properties that are not properties of the thought (or, meaning) that corresponds to the words that occur in the auditory imagery.

There are two problems for HO theories that are familiar and that I am only going to mention here, in order to help set off by contrast the issue that will be the focus of the present section. They are:

(P1) The problem of the stone. Representing a stone does not make a conscious stone, so why should representing a by-itself-unconscious image make a conscious image?

(P2) The problem of false representations. Representation, in general, goes with the possibility of misrepresentation. Suppose a higher order representation misrepresents the properties of an image, or represents the occurrence of an image when there is none. It seems that, from the subject's point of view, everything would be just the same as if there were an image. But then, consciousness cannot derive from the pairing of a first order representation and a HO representation, for it would be present when there is just the HO representation. If that is allowed, however, then HO theory must say what it is about a HO representation *tout court* that accounts for an occurrence of it to be a case of consciousness.

HO theorists have given answers to the problem of the stone (e.g., Rosenthal 1990; Lycan 1996; Gennaro, forthcoming). I am not persuaded that these answers are adequate, but I am not going to review this familiar territory yet again. Regarding the problem of false representations, I have had my say elsewhere (Robinson 1996; forthcoming b), as have others (e.g., Seager 1999), and will not say anything further about it here. What I will do in this section is to explore an issue about explanation and relative burdens of proof.

HO theorists can be expected to object to the very simple view of the preceding section on the ground that (*inter alia*) it does not contain an *explanation* of consciousness. It posits imagery as intrinsically conscious, but it does not define consciousness, nor does it explain how it comes about that imagery is conscious. I accept this point, but it is not one that can offer any comfort to HO theorists. The reason is that HO theories likewise do not *explain* how consciousness arises. They do not provide an explanatory answer to this question: How can two things, neither of which is conscious, constitute an instance of consciousness merely by having one of them represent (certain properties of) the other? And if we make one of the pair (the HO representation) itself already an instance of consciousness, then it will be evident that we can ask the same question of that theory that we can ask of the very simple view, namely, in virtue of what is the HO representation conscious?

It appears that we are at the following impasse. Those who think there are satisfying answers to (P1) and (P2) will think that HO theories have an advantage, in that they explain something that the very simple view does not explain. But some well-regarded philosophers do not think that HO theorists have satisfactory, explanatory answers to (P1) and (P2). If they are right, then the very simple view is not explanatorily worse off than HO theories; and in that case, it might well be preferred on grounds of parsimony.

At this juncture, a small advance for either side may tip the dialectical balance in its favor. The (putative) small advance on which I now want to focus concerns an attempted shift of a burden of proof onto proponents of the very simple view. According to that view, auditory imagery is intrinsically conscious. It is not made to be conscious by its standing in a relation to anything else. But this view implies that auditory imagery is *necessarily* conscious, i.e., that it has properties that could not occur except in a conscious event. Against this view, HO theorists may object that the "neutral" stance lies in favor of open possibilities, i.e., the burden of proof is upon those who claim necessary connections. So, the burden of proof about consciousness has to be borne by proponents of the very simple view. The alleged impasse is not really a symmetrical case; the advantage in parsimony of the very simple view is offset (or more than offset)

by the additional strength of its claim – its claim, that is, that auditory imagery is *necessarily* conscious.[5]

I am unmoved by this argument. The reason is that the necessity of the conscious character of auditory imagery is a consequence of its being a claim about the nature of such imagery, and HO theories equally make a claim about this nature. That is, HO theories claim that all of the properties of auditory imagery are such as to be able to occur in unconscious events. If that is their nature, then it is impossible that auditory imagery *by itself* should ever be a conscious occurrence. This claim is symmetrical with, and just as strong as, the claim of the very simple view, that auditory imagery is (by itself, or in its own right) conscious.[6]

It may aid clarity to put these modal reflections into the idiom of possible worlds. According to the very simple view, the nature of auditory imagery is to be conscious, and there is no possible world in which all the properties of auditory imagery occur together without constituting a conscious occurrence. A simple denial of this claim would amount to saying that there are some worlds in which an occurrence of auditory imagery is (just by itself) a conscious event, and other possible worlds in which an occurrence of auditory imagery (by itself) is not a conscious event. But HO theories do not leave the question of consciousness of auditory imagery (by itself) open in this way. They claim that the nature of auditory imagery (by itself) is not conscious. This does not have the sense of: In some worlds it is (by itself) conscious, in others not. Rather, the nonconsciousness of auditory imagery (by itself, or in its own right) in this world is taken as an indication of its intrinsic nature; and that implies that there is no possible world in which auditory imagery (by itself) is conscious. An equivalent way of putting this commitment of HO theories is to say that it is necessary that auditory imagery (by itself) is nonconscious.

The symmetry in the status of the claims of the very simple view and those of HO theories can be obscured by formulating the issue as the question whether sensory qualities must occur only in conscious events.[7] It will then seem natural to think that a neutral approach leaves this question open, and therefore the burden of proof is upon those who would close the question by giving a negative answer. The point of the last few paragraphs is that it is also a closing of the question to give a definitely positive answer. But a positive answer is implicitly advanced, if we start out by assuming that all relevant qualities of auditory imagery can occur in conscious or unconscious events. So, a terminology that may appear "neutral" on the question of the essentiality of consciousness for the qualities of auditory imagery may in fact be a terminology that is subtly prejudiced against the very simple view.

Perhaps an asymmetry can be reinstated by noting that, according to both the very simple view and HO theories, our concept of auditory imagery is formed in connection with cases, all of which are conscious. (This is true for the very simple view because all cases of auditory imagery are conscious occurrences; and true for HO theories because all the cases of auditory imagery that we use in acquiring the concept are ones that are represented by a HOR.) This fact might be thought to explain why it would be natural to come to think of auditory imagery as having to be conscious. But the importation of a condition of our concept formation into the nature of the object of which the concept is formed would be illicit. After all, we acquire the concept of a table from occasions on which we see or feel a table, but the objects that fall under that concept are not objects that are essentially seen or felt. (They may be essentially visible or touchable, but that is analogous merely to auditory imagery's being essentially representable, and that is a consequence that HO theorists can readily accept.)

This proposal is intriguing, but it cannot be used to shift a burden of proof onto proponents of the very simple view. The reason is that it already presupposes the falsity of that view; in offering an account of a mistake, it presupposes that a mistake has been made.

Perhaps, however, there is this asymmetry: HO views can explain how we came by the concept of intrinsically conscious auditory imagery, but the very simple view cannot. It is, after all, no explanation of the acquisition of a concept of X merely to say that there are Xs, or even to say that there are Xs that really do have the properties that we think they have.

The best answer to this suggestion would consist in giving a positive account of how the very simple view can explain the (proper) acquisition of the concepts it invokes. This task would, however, carry us far beyond the scope of the present paper.[8] For the present it will have to suffice to point out that a burden of proof cannot be imposed merely by showing that a mistake *could* have been made. A case would have to be made that there is no way of acquiring the concept of intrinsically conscious auditory imagery that does not rest on some error. This is a very strong claim, and the burden of proof for it is surely on those who would advance it.

The conclusion of this section is that there is no non-tendentious way of shifting a burden of proof onto proponents of the very simple view. If we are careful not to skew the terminology in which we formulate the issue, we cannot get beyond the dismal symmetry in the rival theories that amounts to what I called an impasse. So long as the issues in (P1) and (P2) are not decisively resolved, both HO theories and the very simple view have much work to do.

But the burden of proof, or the burden to explain how or why consciousness comes into the world, is no more upon one side than the other.

7. The resolutely simple view

What I shall call "the resolutely simple view" takes over the idea, from the very simple view, that auditory imagery is by nature conscious. It denies, however, that a thought, *T*, is required in order for there to be an expression of a thought. Occurrence of a subvocal saying produced in the normal manner is enough. The resolutely simple view can be diagramed as in Figure 4.

Even those who may have been inclined to share my respect for parsimony may suspect that I have now lost my grip on reality. Several points in what follows will help to allay this suspicion.[9]

The first thing to notice is that although there is much that is left out of the most recent diagram, what is left out is the same as what has been left out of all the others. Namely, I have always assumed, without representing it in the diagrams, that there is an organized brain that *produces* the items that are represented in the diagrams. I have already mentioned our grammatical capabilities. These presumably depend on brain operations. But all we are conscious of is the grammatical speech, overt or subvocal, that results from these operations. We are never directly conscious of the operations that produce the grammaticality of our verbal productions.

I suggest that it is the same with cognitive mechanisms that produce the content of our verbal productions. To see the point of a joke, for example, words must be understood, and, in general, many background facts about people, recent events, societal norms, and so on must be brought to bear. The accessing of relevant background typically takes place in a few seconds or less, and what we are conscious of is not the processing that leads us to see the joke, but only the result of that processing.[10] The resolutely simple view does not at all deny that this unconscious processing takes place. To the contrary, it insists that it does take place, and that it is unconscious.

What then does the resolutely simple view deny? It denies that behind a verbal production there must occur a unitary representation, the content of which matches the content of the words that occur in the verbal production.

I

Figure 4.

It denies that a case of a conscious thought has the following structure: First, there occurs a thought *that p*, which, just by itself, is not conscious, and then, if the thought is a conscious thought, there is an expression of it, either in overt speech or in subvocal speech. That is the implication of the "*T*" in the diagram of the very simple view, and that is what is lopped off in the resolutely simple view. The latter view says that there is no need to assume that everything comes together once in thought, and then it comes together again in the overt speech or subvocal saying of a sentence that expresses the thought.[11] Naturally, if this view can be sustained, i.e., if "*T*" can be dispensed with, then there will be no plausibility to the complex view – that would require a higher order representation of something that is not there to be represented.

Does the resolutely simple view deny that our words express our thoughts? No. Crucially, however, it does offer an analysis of "expressing a thought" that excludes the necessity of *matching* the content of a prior event if there is to be a case of expressing. The previous views that I have diagramed lend themselves to the idea that an overt or subvocal saying expresses a thought just in case the meaning of the words in the saying *match* the content of a distinct event, namely, the occurrence of the thought that is said to be expressed. That is what the resolutely simple view denies. It says that there is a cognitively organized brain that governs the production of our sayings, but it allows that the very first time that our cognitive organization produces a representation of a state of affairs can be its production of a (subvocal or overt) saying.[12] The resolutely simple view does not deny that there can be some cases in which an adequate unconscious representation of a whole state of affairs is produced by our cognitive organization, and that it is only later on, and partly as a causal consequence of the unconscious representation, that a saying with the same content is initiated. It does deny that this kind of prior structure is necessary for there to be an expression of a thought.

But what sense can be made of "my words express my thoughts" if not that my words have a content that matches a (prior or simultaneous) mental event? The answer can, perhaps, be most easily seen through use of the intermediate phrase, "what I think". My words express my thoughts if, when I say them (overtly or subvocally), I say what I think. And I say what I think if what I say fits in with – forms a coherent pattern with – what I have recently said and done, and what I go on to say and do.[13]

It may be suspected that the resolutely simple view cannot account for *mis*expression. To the contrary, the beginnings of an account can be roughly indicated, as follows. S's saying "p" misexpresses what S thinks if S makes a correction, or if S continues in a way that is evidently incompatible with accep-

tance of the view expressed by "p". Of course, if S's expression of views incompatible with *p* comes at a much later time, S may have had a change of mind (rather than an original misexpression). Generally, we take S's word for deciding whether an overt utterance was a misexpression or an adequate expression of a view that S has now abandoned. We also allow for cases in which some time has elapsed, and S is genuinely puzzled as to whether a previous statement was a misexpression or an adequate expression of a formerly held view.

I make no attempt here to refine the rough account of misexpression, as it would evidently be a formidable task to state necessary and sufficient conditions for an utterance to be a misexpression of a person's thought. I observe, however, that the theory of matching a previous, unconscious representation offers no help toward such an account. Such thoughts are in any case not directly accessible to others, and not directly accessible to subjects themselves. However it may be that we distinguish between misexpressions and changes of mind, it is not by comparing expressions with prior, unconscious thoughts, even when we are considering our own case. Proponents of the resolutely simple view thus have at their disposal the same resources for making that distinction as do proponents of less simple views.

It may be now be further objected that the resolutely simple view denies obvious truths of mental causation, e.g., that our actions are brought about by our intentions. Again, there is much that might be said here, and I shall only sketch the outline of the response of the resolutely simple view. We act intentionally when it seems to us that things are going well, and with no surprises. It pays here to think of very ordinary cases. For example, I hear the noise of a vehicle in the street. I look out the window to see what it is. Did I turn my head intentionally? Yes. Did I plan to turn my head? There is no reason to say so. Must my movement have been preceded by a representation of a desire to turn my head? That seems to be an unnecessarily complex account. I have a cognitively organized brain, and its state plus the input of noise from the street led to my turning my head. For this to happen, many events must take place in my brain, but none of them needs to be an adequate representation of a desire to turn my head. I wanted to do that, if I did it with no hesitation, and no feeling of regret or of being forced.

There are, of course, cases in which we plan ahead. "When I see Jones," I say to myself, "I won't even speak to him. I'll acknowledge his presence only with the minimum that is required by the barest civility." Then Jones appears, and I act accordingly. If we generalize from cases of this sort, we will form a picture, according to which intentional actions are actions that match the content of a plan. And if we then add that speaking is (in general) intentional action, our

picture will be that speaking is acting out a prior plan. If we apply this picture to subvocal saying, we will regard it too as acting out a prior plan; and then we will need thoughts behind the saying, whose content is matched in the saying.

But we should be wary of this picture. If subvocal saying is necessarily preceded by a plan, why must not the thoughts be so preceded? One can, of course, make a verbal distinction, and say that conscious thoughts must be preceded by an unconscious thought, but unconscious thoughts need no such predecessor. Or, if one objects to the precedence, one may say that conscious thoughts must have as a component, an element that would by itself be an unconscious thought with the same content. But this would not explain how an unconscious thought could come to have *its* content to be fully in order without a predecessor (or a component) that has that content. Requiring that kind of predecessor or component, however, would evidently lead to a regress. Thus, it seems that we must allow our cognitively organized brain to produce representations of states of affairs that have not been represented either in a previous thought or a component of a thought. If we allow that, there is no reason why there may not be many cases in which the *first* time that our cognitively organized brain produces an adequate representation of a state of affairs is when it produces a saying (overt or subvocal) of a sentence that affirms that state of affairs.

Do these views and arguments commit us to holding that ordinary people who say they acted *because* they wanted some result and believed their action would produce it are saying something false? No. The resolutely simple view only rejects an analysis of such remarks that requires prior, discrete desire thoughts and belief thoughts. In many ordinary cases, our cognitively organized brain not only produces complex and well sequenced actions that accomplish meaningful goals, but also produces speech that correctly describes the relation between immediate actions and long term goals. And it is often also true that the actions would not have been done had not the circumstances provided a relation between them and our goals, which our speech identifies. There are, of course, other cases in which there are divergences in content of speech and action (Nisbett & Wilson 1977a, 1977b), just as there are cases in which our nonverbal actions fail to be properly sequenced or consistently directed.

8. Conclusion

HO theories will remain attractive not only because they have able defenders, but because they present one of two fundamental ideas in coming to terms with consciousness. (The opposing fundamental idea is that consciousness cannot

be constructed out of elements that are, in themselves, not conscious.) There can be no question of establishing or refuting such a fundamental idea in a single paper, and I have not attempted anything so ambitious here.[14] Instead, I have addressed the important question, "If not HO theories, then what?" I have described two alternatives to HO theories, and explained something of their merits. I believe I have exhibited some genuinely attractive features of these alternatives, and I hope to have stimulated interest in their further consideration and development.[15]

Notes

1. Neural studies support the likeness of imagery and perception through findings of partially common activations in the two conditions. See Jeannerod (1994) and O'Craven & Kanwisher (2000).

Readers may wonder about schizophrenics' "hearing voices". This phenomenon may be more vivid than ordinary subvocal imagery, and less like ordinary speaking than is subvocal saying. It is not my topic here, however: (3) concerns only ordinary subvocal imagery.

2. The words in subvocal speech may be said to express a content that is true or false. If one adopts this style, it would be natural to express the point of this paragraph by saying that the auditory images that compose subvocal speech are conscious events that are constituted by non-semantic properties, and that their expressing a true or false content depends on facts additional to their having these non-semantic properties.

3. "Appropriate" abbreviates "non-inferential and non-observational". I suppress this aspect, because it does not figure in the argument to come.

4. I do not mean to suggest that this is all there is to thinking. Accessing of language centers will be driven by processes in other parts of the brain.

5. This argument is an explicit formulation of the sense of a number of remarks that have been put to me in conversation by several philosophers. The context of these remarks has been somewhat different on different occasions, and, naturally, never exactly the same as the context of the present paper. Nonetheless, I am confident that I am responding to a line of criticism that will occur to a number of readers.

6. Here, and throughout this discussion, I have used formulations that apply to HO theories that follow Rosenthal's (1990, 1991) understanding of consciousness of mental states that are not themselves thoughts. If auditory imagery is analyzed in such a way that a HOR is *part* of auditory imagery, then my remarks in this section have to be taken to apply to the *other part* of auditory imagery, and not to auditory imagery *tout court*. See Gennaro (1996) for an analysis of this kind.

7. Apart from using "states" rather than "events", this is the approach in Rosenthal (1991).

8. I develop an account of phenomenal concepts in Robinson (forthcoming b).

9. Am I being hasty in proceeding directly to the view in Figure 4 without considering the other way of reducing Figure 3 to two items? That reduction would eliminate *T* while leaving both *I* and HOR2. However, the issues involved in deciding between this case and the resolutely simple view I am about to discuss are the same as the issues involved in deciding between the simpler view of Figure 2 and the very simple view of Figure 3. Thus, the present and preceding sections include, in effect, discussion of the view that has just *I* and HOR2.

10. See Velmans (2000) for discussion and elaboration of this point.

11. The idea of "coming together" is derived from Dennett (1991a). Many of his cases are perceptual, and I have criticized some aspects of his view in Robinson (1994). But Dennett's fundamental idea seems right to me in the case of thoughts.

12. For further explanation of some of the ideas in this and the following paragraphs, see Robinson (1995, 1986, 1988).

13. See Robinson (1995), and Dennett (1991b). The description in the text would be objectionably behavioristic if it were assumed that what I have said and done and what I go on to say and do could be given an adequate description wholly in terms of physical motions, but this is not assumed (or implied) here. The fact that we cannot describe in purely physical terms what we expect of people who say "p" (for virtually any "p") does not imply that nothing they do will surprise us, and surprise us in a way we would not have been surprised if we had not heard them say "p". Nor does it imply that we must be surprised only because we have first postulated a thought (conscious or not) behind the saying of "p" that we have heard. See Robinson (1988, 1986) for further explanation of these points.

14. There is further argument toward various parts of the resolutely simple view in Robinson (1990, 1995, forthcoming a).

15. I would like to thank Rocco Gennaro and Peter B. M. Vranas for their comments on an earlier draft of this paper. Errors that may remain are, of course, my own.

References

Dennett, D. C. (1991a). *Consciousness explained*. Boston: Little, Brown & Co.

Dennett, D. C. (1991b). Real patterns. *The Journal of Philosophy, 88*, 27–51.

Gennaro, R. J. (1996). *Consciousness and self-consciousness: A defense of the higher order thought theory of consciousness*. Amsterdam & Philadelphia: John Benjamins.

Gennaro, R. J. (forthcoming). The HOT theory of consciousness: Between a rock and a hard place?

Jeannerod, M. (1994). The representing brain: Neural correlates of motor intention and imagery. *Behavioral and brain sciences, 17*, 187–245.

Lycan, W. G. (1996). *Consciousness and experience*. Cambridge, MA: MIT Press.

Muter, P. (1980). Very rapid forgetting. *Memory & Cognition, 8*, 174–179.

Nisbett, R. E. & Wilson, T. D. (1977a). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Nisbett, R. E. & Wilson, T. D. (1977b). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*, 250–256.

O'Craven, K. M. & Kanwisher, N. (2000). Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *Journal of Cognitive Neuroscience, 12*, 1013–1023.

Robinson, W. S. (1986). Ascription, intentionality and understanding. *The Monist, 69*, 584–597.

Robinson, W. S. (1988). *Brains and people*. Philadelphia: Temple University Press.

Robinson, W. S. (1990). States and beliefs. *Mind, 99*, 33–51.

Robinson, W. S. (1994). Orwell, Stalin and determinate qualia. *Pacific Philosophical Quarterly, 75*, 151–164.

Robinson, W. S. (1995). Mild realism, causation, and folk psychology. *Philosophical Psychology, 8*, 167–187.

Robinson, W. S. (1996). The hardness of the hard problem. *Journal of Consciousness Studies, 3*, 14–25.

Robinson, W. S. (forthcoming a). Thoughts without distinctive, non-imagistic phenomenology. *Philosphy and Phenomenological Research*.

Robinson, W. S. (forthcoming b). *Understanding phenomenal consciousness*. Cambridge: Cambridge University Press, 2004.

Rosenthal, D. (1990). *A theory of consciousness*, Report No. 40/1990, Research Group on Mind and Brain, Center for Interdisciplinary Research, University of Bielefeld, Germany.

Rosenthal, D. (1991). The independence of consciousness and sensory quality. In E. Villanueva (Ed.), *Consciousness: Philosophical issues, 1* (pp. 15–36). Atascadero: Ridgeview Publishing Co.

Seager, W. (1999). *Theories of consciousness*. London & New York: Routledge.

Simons, D. J. & Levin, D. T. (1997). Failure to detect changes to attended objects. *Investigative Ophthalmology and Visual Science, 38*, S707.

Velmans, M. (2000). *Understanding consciousness*. London & Philadelphia: Routledge.

Wolfe, J. M. (1999). Inattentional amnesia. In V. Coltheart (Ed.), *Fleeting memories* (pp. 71–94). Cambridge, MA: MIT Press/Bradford.

Chapter 14

Higher order representation in a mentalistic metatheory

Donelson E. Dulany

Must there be some kind of higher order representation for a mental state to be a conscious state? This is a wondrously old and tangled idea. Coming to this literature as a cognitive psychologist with more than the usual interest in consciousness, I see that higher order theories of consciousness range historically from Aristotle (Caston 2002), through Locke and Kant (Güzeldere 1997), and perhaps to Sartre (Gennaro 2002). Jaynes (1976) even argued, rather famously, that consciousness itself only emerged historically with the arrival of higher order reflection (too late for the Iliad but just in time for Socrates).

The mentalistic metatheory I have described (e.g. Dulany 1991, 1997, 2002) contrasts in the most fundamental ways with higher order representation theories of consciousness that I find in the philosophical literature. On the mentalistic metatheory, all mental states are conscious states, and higher order conscious states are yielded only by memory and/or inference from lower order conscious states. My aim is to contrast these views and to examine their relative credibility in the light of evidence and standards for empirical examination.

General nature of the philosophical literature

What I see as common to influential variants today is this: A first order mental state, one that represents the world beyond, can be a conscious mental state only if graced by some higher order mental state – of some kind, in some relation, in some way. Why? Drawing on experimental literatures purporting to reveal unconscious mental states (usually perceptual), theorists must explain how only some of those "naturally unconscious" mental states possess the "property" of consciousness. With a fear of mysterians at the gate, a fear

that besets (and perhaps besots) some in more than one discipline, these mysterious conscious states must somehow be "naturalized" – and how better than by some kind of relation to, and contrast with, all those naturally *non*conscious mental states? Different level state-mates must be anything but soul-mates.

The real conceptual tangle today lies in the way variants of the view contrast on several dimensions:

Nature of the higher order representation

What serves this role can be abstractly conceptual higher order thought (HOT, e.g. Rosenthal 1997), or with some kind of "monitoring", either a higher order perception (HOP, e.g. Armstrong 1968) or a quasi-perceptual higher order experience (HOE, e.g. Lycan 1996). Shift among these and we still have one claim that a mental state is enlivened with consciousness only if it can get some purchase on some kind of higher order representation (or HOR, if we must).

Things are further complicated with claims that the necessary second order representation is nonconscious (e.g. Rosenthal 1997) and either conscious or "possibly nonconscious" (e.g. Carruthers 2000: 227). And if the higher order representation is conscious, should we distinguish between being conscious *of* the lower order state and conscious *that* one has that state (Lurz 2003)?

Relation of first order to higher order representation

This relation has been said to be "roughly contemporaneous" (Rosenthal 1986: 335) and oddly called "actualist". Or the conscious state can only be "dispositional" to its HOR that temporally lags behind (Carruthers 2000). Or the necessary higher order representation is said to be in some way "intrinsic" to the conscious mental state (e.g. Gennaro 1996; Natsoulas 1999) – with the other two views characterized as "appendage" theories.

To twist the tangle further, the causal direction between first and higher order states could go one way or the other, but often seems ambiguous or unspecified. A first order representation can be dispositionally causal to its higher order state, the disposition rendering the first conscious (e.g. Carruthers 2000). But for Rosenthal (1997) the first and its second order representation seem to move in parallel. For that matter, anyone endorsing Locke's famous "Consciousness is the perception of what passes in a man's own mind" would dispense with causal direction by placing consciousness in the higher order representation exclusively.

What then, in essence, is the mentalistic metatheory, and where does it lie on these dimensions of contrast?

Mentalistic metatheory

This mentalistic view was first presented to a rather large audience at the Psychonomic Society (Dulany 1984). It was then elaborated and contrasted with the dominant cognitive metatheory, as well as used, as it had been earlier, in theories and experimental analyses of explicit and implicit learning in Dulany (1991, 1997). It was recently outlined in Dulany (2002), and aspects of the metatheory have been used by others (e.g. Carlson 1997, 2002; Perruchet & Vinter 2002; Tzelgov 1997).

Nothing as complex as consciousness submits to anything as simple as a definition. We may begin with common intuitive acquaintance and then lay out metatheoretical assertions that may imply empirically supportable theoretical assertions. Furthermore, "conscious of (or that) ___" and "aware of (or that) ___" in this usage mean the same thing and differ only in how gracefully they fit particular contexts.

Central principles

1. A *mental state* is a *conscious state*, with a conscious *mode* that is the sole carrier of *symbolic content*. These contents represent the present in perception, the past in remembrance, and the future in intentions, plans, hopes, fears – all conscious modes. Consciousness is no more a "property" of mental states than "my desk" is a property of "my desk".

   "Symbol" is not satisfactorily definable either. It is a theoretical construct with its scientific meaning, a network specification, given by the growing set of functional relations it enters into, the most central being these: (a) Symbols evoke other symbols, as in some idle moment when a thought of "Paris" evokes a thought of "Cafe Flore". (b) They serve as subjects and predicates of propositional thoughts, granting them a role in deliberative inferences and decisions. In fact, they can appear in the key proposition, "What I'm thinking of refers to ____ ". Within intentions, they may also (c) enter into the control of actions that are consistent with that key proposition. A simple action consistent with the symbolic value of thinking "coffee" is picking up the cup.

2. The person entertaining a conscious state ordinarily does so with some *sense of possession*, Brentano's (1774, 1973) "agency," a sense that varies in intensity over persons and conditions. It is sometimes intensely felt, sometimes weakly; and in psychotic episodes, the sense of possession may become so weak that thought seems alien and "inserted" – a psychiatric syndrome brought to the recent philosophical literature in Stephens and Graham (2000).
3. These conscious states come in two *classes of modes* with two *classes of contents*: *Propositional* modes carry contents that are propositional, such that we would speak of being aware that ___; and they come in sub-classes in which we believe that ___, perceive that ___, etc. *Sub-propositional* modes carry contents that are sub-propositional, such that we would speak of being aware of ___; and these, too, come in sub-classes of feeling of ___, fear of ___ , etc. Not only do modes vary in type, each type varies in degree, such as strength of belief or fear. We could speak conventionally of "propositional attitudes," and extend the usage to "sub-propositional attitudes," but "attitude" has long been claimed by social psychologists for modes such as feeling of ___and belief that ___ directed toward some specific target.
   In notation, then, we can say that a conscious mental state has this structure: I(mode[content]).
4. A *mental episode*, such as inferring or deciding or remembering, consists exclusively of conscious states interrelated by a nonconscious mental operation, an operation that is no more than the brain processes that interrelate and transform conscious states.

   In notation, a single mental episode could be written this way:

   $$\text{Cs State}_{in+1} \leftarrow \text{Ncs Op. (Cs State}_{\text{jn}},\ldots,\text{Cs State}_{\text{kn-m}}).$$

   And where conscious modes and/or contents can be quantitatively scaled, as in theories of intentional control and causal learning (e.g. Dulany 1968; Carlson & Dulany 1988), a set of mental episodes would be represented this way, with the set of operations represented by the function:

   $$\text{Cs States}_{in+1} = \mathbf{f}(\text{Cs States}_{\text{jn}},\ldots,\text{Cs States}_{\text{kn-m}}).$$

5. Mental episodes, then, are also of two general kinds: In an *evocative mental episode*, sub-propositional content evokes, by activational operation, another sub-propositional content. And in a *deliberative mental episode*, a deliberative operation of inference or decision yields one proposition from others – a conclusion in the case of inference, an intention in the case of decision. The kinds of episodes are causally interrelated: By repe-

tition, what was deliberative may become evocative. And by remembering what evokes what, the remembrance may be expressed propositionally and used deliberatively in inferences and decisions.

6. Symbols come in *codes.* In an *identity* code, which can be conceptual or analog, with the conscious content an identification as such (circle, person, trumpet sound). In a *literal* code, however, there is a fleetingly conscious representation of color or form, pitch or volume, sweet or sour, etc., depending upon the modality. Literal codes briefly precede identity codes in perception pre-attentively and surround these attentional identifications as a "fringe" (Dulany 2001).
7. The *domain of mental episodes* runs from the output of sensory transduction to the input to motor transduction. "Transduction" is a term from engineering sometimes used to label nonconscious neural processes that convert physical wave forms into a conscious sensory state, and convert final conscious intentions and sensations into muscular activity.
8. Inactive, *nonconscious memory* is constituted of non-symbolic neural networks that are established through associative or deliberative learning (e.g. Dulany 1997), and can when activated yield actions or symbols in conscious remembrance. Contents of consciousness are not "stored" and "retrieved," maintaining their symbolic functions and identities when outside consciousness. Something a scholar is said to "know" when not thinking about it does not even remotely function in the way what is consciously (and really) known does when its neural network is activated. With automatization, mental states drop out, not down to an unconscious, consistent with diminishing fMRI for the relevant network (e.g. Bolger & Schneider 2002).
9. *Orders of representation*? This makes use of all that precedes. Put simply, innumerable contents of first order conscious contents represent the world beyond mental activity in various modes; higher order conscious contents (sometimes called meta-awareness, metacognitive awareness, or reflective awareness) may represent other conscious states or mental episodes. They do so by memory and/or inference, resulting in either propositional or nonpropositional contents. In either recollective or inferred remembrance, the higher order awareness of a first order awareness state could be denoted this way:

   I[Aware_2(Cs State_1), where the first order state becomes the content for the higher order awareness mode.

With the use of memory and inference, awareness of a single mental episode might be represented this way:

$$\text{Aware}_{n+2}[\text{Cs State}_{in+1} \leftarrow \text{Ncs Op. (Cs State}_{\text{jn}},\ldots,\text{Cs State}_{\text{kn-m}})],$$

And awareness of a class of mental episodes, could be represented this way:

$$\text{Aware}_{n+2}[\text{Cs States}_{in+1} = \text{f(Cs States}_{\text{jn}},\ldots,\text{Cs States}_{\text{kn-m}})].$$

With abstraction over many states and episodes, various remembrances and inferences constitute a "theory of self".

This also amounts to a formalization of something essential to introspection, with the expectation of considerable but imperfect validity, depending upon conditions for memory of prior states, but much less validity with more fallible memory of, and inference to, more complex mental episodes. Of course, too, conditions of verbalization and disposition to report must also be met to increase the validity of introspective reports. Think for the moment of introspectively describing the difference between a Monet and a Manet – or a merlot and a cabernet.

Overall Nature

The mentalistic metatheory provides a modern analysis of the *intentionality* of consciousness, of what Brentano (1774, 1973) meant by its ability to entertain "aboutness." (An early resolve: to stay out of the "Meinongian jungle" and not go down "the old Chisholm trail".) It elaborates the view that conscious contents can represent a world beyond that particular conscious state – whether an outside world or an inner world of mental activity, and whether in reality or not (Brentano's "intentional inexistents".) The adaptive value of this kind of capability, for first order as well as higher order contents, is too obvious to argue. And if consciousness is the carrier of those symbolic contents, especially the sole carrier, we have a principled reason for the emergence and enrichment of consciousness throughout evolution. On this view, "cognition" as a nonconscious carrier of symbolic contents has been no more than an ether to call our own.

The mentalistic metatheory entails no ontological commitment, materialistic or non-materialistic – which is just as well since the "hard problem" (Chalmers 1996) remains hard and science as we know it offers no convincing way to solve it. The conduct of science, however, is unimpeded by ontological agnosticism. Nevertheless, mentalism as a metatheory may be endorsed while

holding either of these ontological positions, usually the first, as a working assumption.

Where then does the mentalistic metatheory lie on those dimensions separating current philosophical views of higher order representation – and why?

Nature of the higher order representation

Higher order representations can be either analogue (as in HOE) or conceptual (as in HOT) – as a substantial experimental literature shows (Gardiner & Richardson-Klavehn 2000). Reporting recognition of an item is by its nature report of a higher order representation: "I (agent) recognize (higher order mode) this item as experienced (first order mode) earlier in the experiment (content)". Some recognized items are reported as "recollected" (or "remembered") in the sense of imagery of its presentation, and others are reported as something only "believed" (or "known") to have been experienced in the absence of that recollective imagery. Interestingly, too, the literature shows that recognition in the "believed" modes is less subject to interference by a secondary task. It isn't just one, but either – in interestingly different ways. I would, however, regard both kinds of conscious content as "phenomenal" – in fact possessing nicely un-Quined qualia – but also generally susceptible to "access" for direct report or use in deliberative thinking (cf. Block 1995; Carruthers 2000).

Nevertheless, any reference to "higher order perception" (HOP) would be either a pleasant metaphor (as is "scanning") or an anatomical fantasy. We have exteroceptors, proprioceptors, and enteroceptors, but no "menteroceptors."

Furthermore, whether the recognition is recollective or believing in these two senses, higher order representation should consist of a sense of ___ that precedes the recognition that ___. Since all remembering, at consummation, comes by activation of a neural network, activation would yield an image of ___ or familiarity of ___ which may by recoding or inference yield a recollection that ___ or belief that ___ was experienced before. Much the same could be said of that other important measure of explicit memory: recall. "I (agent) recall (higher order mode) experiencing (first order mode) _____ (content)."

The preceding is episodic explicit memory. But does implicit memory, too loosely called "unconscious memory" (e.g. Anderson 2000: 236), provide evidence of unconscious higher order representation when it is dissociated from explicit memory? The dissociation comes with medial temporal, hippocampal damage, and in normals with certain experimental procedures. In a common priming paradigm for implicit memory, words are presented to awareness, such as "market" and later tested by presenting a fragment to awareness, such as

"mar__", or the whole word very briefly or in low illumination. Prior presentation primes in facilitating word completion (e.g. Graf, Squire, & Mandler 1984) and identification (e.g. Jacoby & Dallas 1981) and may do so independently of explicit memory of the word, assessed by recall or recognition.

It is widely accepted, however, that the priming effects are highly specific (e.g. Roediger & Scrivener 1993). Presentation of the word should strengthen the activational association sequentially within the word as measured in word completion, and from literal to identity awareness of the word as measured in identification. What is presented to awareness becomes more available to awareness – and there is no evidence that a higher order representation of what is not remembered explicitly, "I remember seeing the prime word before," is unconsciously remembered (Dulany 2000, 2002).

Relation of first order to higher order representation

In over a century of studies, assessed higher order states sometimes follow and sometimes precede some lower order state of interest, but I don't believe these studies provide any basis for saying they are even "roughly" contemporaneous. This is obviously true of those studies of explicit (episodic) memory. It is true of the many studies that have used introspective reports of remembrance that some prior state(s) occurred during learning (e.g. Dulany 1968). And Ericsson and Simon's (1993) "verbal protocols" in problem solving are reports of remembrance that some prior mental episode(s) occurred.

In the many studies of feeling of knowing (FOK), on the other hand, subjects report their conscious belief that a non-remembered paired associate will be remembered at a later test; and in studies of judgment of learning (JOL), they report a conscious belief that something being studied now will be remembered later. This is clearly elaborated in (Koriat 2000; Metcalf 2000; Nelson 2000) and their references. In both cases, this metacognitive belief is a tentative higher order representation of what will be consciously remembered later – although just what activates that metacognitive awareness varies with experimental circumstances. We also have significant analyses of the role of conscious intentions in controlling the future mental activities they metacognitively represent (Carlson 2002).

What then would I see as limitations of representative HOR theories of consciousness?

Influential higher order representation theories

"Actualist" theory

As Rosenthal (1997: 741–742) puts it, "...a mental state will be a conscious state if it is accompanied by a thought about that state...The core of the theory, then, is that a mental state is a conscious state when, and only when, it is accompanied by a suitable HOT...[which] makes us conscious of the mental state". But "we are seldom aware of such HOTs...a HOT will not itself be a conscious thought unless one also has a third-order thought about the second-order thought" – which is also nonconscious to avoid infinite regress. Only then would we speak of introspection, a third order thought and something that Rosenthal and others recognize as only occasional.

Psychologists will wonder why we should worry about infinite regress as a logical possibility when it is a human impossibility. We can naturally wonder, too, why we are conscious of something in the world beyond, a desk or a computer, only when the HOT makes us "conscious of the mental state" – which sounds like an intermediate state between the lower order state and its accompanying HOT. (A first-and-half order representation?) This feature of the theory is emphasized in Rosenthal (2000a: 203): "The leading idea behind the HOT hypothesis is that a mental state is conscious only if one is, in some way, conscious of that state." For that matter, we can wonder whether we could really have a nonconscious HOT, be aware of something beyond, and also aware of that awareness, too, all concurrently – and why evolution would have produced such a layered and labored system.

I pass over this aspect of the theory, however, because it seems to be contradicted by the concession that "I can be conscious of things without being conscious that I am" (Rosenthal 2000a: 206) – which I would interpret as a sensible denial of obligation to introspect. Nevertheless, the central role for those accompanying but nonconscious HOTs remains: "...the HOTS I have about these states are seldom, themselves, conscious", and "...the requirement I impose that HOTs be occurrent, rather than merely dispositional." (Rosenthal 2000b: 240).

Now, the most fundamental problem I see for the theory follows from these two circumstances: (a) If a lower order state is conscious "when, and only when" accompanied by its nonconscious HOT, these two states would have the same position in any network of theoretical relationships. Their causal antecedents and consequences would be the same. (b) Furthermore, although we have assessments of conscious mental states, we have no independent co-

ordination of nonconscious HOTs , and none could be expected to function independently of coordinations of their associated conscious states.

The problem I see? On the theoretical network, or functional, specification of the meaning of a theoretical construct – an analysis dating to Hempel (1952) in the philosophy of science and used in areas as diverse as test validation (Cronbach & Meehl 1956), intentional control (Dulany 1968), neurophilosophy (Churchland 1986), and even HOR theory of consciousness (Carruthers 2000: Ch. 1) – the meaning of a theoretical construct is given by the set of (usually causal) relations it enters into. Therefore, a "nonconscious HOT" would have no theoretical and empirical meaning apart from its "lower order (and now) conscious state."

Furthermore, on network logic applied to measure validation (termed "construct validation"), we gain confidence in the validity of a measure of an unobservable state to the degree the measure behaves in the way specified by theory of the state (Cronbach & Meehl 1956) and not in ways more credibly explained by theory of another state (Dulany 1968, 1991). The trouble for the "actualist" theory is this: Theoretical expectations for conscious states and their HOT mates would be the same, and so any proposed measures of HOTs, behavioral or neural, would more credibly reflect the conscious states themselves – where we have supportable and supported theory of their action, can assess them, and know they exist.

In short, on network specification of meaning, "nonconscious HOT" reduces to an odd alternative term for "lower order conscious state." And on a network logic of empirical support, saying that a HOT is necessary for a mental state to be a conscious state loses the possibility in principle of empirical confirmation or – conveniently – disconfirmation. Put more bluntly, the central claim of the theory becomes, in a fine old philosophical expression, "empirically vacuous."

Nevertheless, I have found the theory ingeniously and elegantly expressed, with surrounding aspects that I would agree with – for example, the significant ability of perception, memory, and introspection to fail or falsify under certain circumstances.

"Dispositional" theory

According to Carruthers (2000: 227–229), the essence of his dispositionalist HOT theory is that "the conscious status of an experience consists in its *availability* to HOT" – disposition to its later higher order representation. "In contrast with the actualist form of HOT theory, the HOTs which render M [a

mental state] conscious are not necessarily actual but potential". How? Perceptual contents enter "short term memory stores, C (conscious) and N (nonconscious)...But C itself is now defined, *inter alia*, by its relation to HOTs – any of the contents of C being apt to give rise to a HOT about itself, should circumstances (and what is going on elsewhere in the system) demand". What then? The "disposition" may be fulfilled through a process referred to as a non-inferential "mind-reading, HOT-generating system".

We can pass over whatever questions we might have about a habitat for a "mind reading" agency, or about a separately "boxed" STM system – or for that matter, what the theory might have to say about our first order *non*-analogue thoughts and feelings. What I see is a mixed set of claims, with the clearly defensible claims undercutting the central thesis:

Certainly it is important to recognize "...the richness of conscious experience in relation to the relative paucity of HOTs..." (Carruthers 2000: 229) – a recognition I take to refer to *conscious* HOTs, since the author would be in no position to know this about "possibly nonconscious" HOTs (Carruthers 2000: 227). Those "circumstances" and "what is going on elsewhere in the system" are highly variable and uncertain. Introspection upon earlier conscious states can be quite fallible. In decades of memory studies, too, items have been presented clearly to awareness with incidental learning instructions or interference, conditions that substantially block higher order representation in "elaborative processing" at the time. At a later memory test, there are numerous failures of higher order representation of that earlier conscious experience.

Certainly it is important, too, to recognize, as in Chapter 9, that with only an early conscious "feel," prior to attentional identification as such, there may at a later time be no higher order thought of the content it carries. We can give attention to the fleeting conscious *mode* carrying that preattentive content, and remember the mode – as does the author, a capability described in Dulany (2001). But the *content* must be in attention, an identification as such, in order to establish memory of those contents, as shown in Carlson and Dulany (1985) and studies reviewed by Perruchet and Vinter (2002). This is another clear case of conscious first order contents without a higher order representation. In fact, there are now examinations of "temporal" and "translation" dissociations between first order and higher order awareness (e.g. Schooler 2002).

This, I think, is the problem: With what we know about the highly variable relation of first order awareness to its higher order representation, and the variable memory and inference processes determining that relation, it is indefensible to say that it is a mental state's "disposition" to, or "potentiality" for, a HOT that "defines" its conscious status. Empirical assertions are not defini-

tional – not even stronger ones. Neither could "disposition" be identified in any way other than by identification of an independently conscious first order state. For that matter, I believe the acknowledged relationship of first order to higher states is too weak even to label this a higher order thought theory of consciousness.

Nevertheless, I find much in this work that is illuminating and valuable. But ironically, it is what I think inconsistent with a higher order thought theory of consciousness that seems most so.

"Intrinsic" theory

Gennaro (1996) clearly presents a version of "intrinsic theory," and over the last decade or so, Natsoulas has presented a series of eight scholarly analyses of variants of the theory (e.g. Natsoulas 2001). Gennaro (1996: 13) opens his account with the general claim that "When one is in a conscious state, it is natural to view its being conscious as an intrinsic property" (1995: 15). Natsoulas (2001: 219) directs the search for just what is intrinsic: "In our attempts to determine the intrinsic properties of states of consciousness we are well advised to attend to our inner awareness of them."

In Gennaro's theory (1996: 24), the heart of the matter becomes this: "The WIV [wide intrinsicality view] holds that the MET [metacognitive thought] is part of the conscious state which is rendered conscious. . .Consciousness is the intrinsic property of 'having a MET that one is in that state."' What distinguishes this theory from Rosenthal's, as he emphasizes, is that the metacognitive thought that confers consciousness on the state is not a separate state but an intrinsic part of a complex mental state. Like Rosenthal's HOT, however, Gennaro's MET would be a nonconscious representation, and introspection would be an occasional conscious representation, with that conscious status granted by its own internal and nonconscious MET.

What seems important about this view, and the range of theories reviewed by Natsoulas, is that conscious states are set apart by something intrinsic, perhaps unique, to conscious states. In replacing "appended HOT" with "HOT spot", however, the intrinsic theory carries over the "actualist" theory's insensitivity to empirical examination, a problem entailed by a HOT that is both concurrent and nonconscious. I also find it difficult to understand how any state that is "intrinsic to" in the sense of "part of" the lower order state can at the same time be a "higher order" state – or for that matter, how anything intrinsic to a conscious state could be *conveyed* by a higher order representation, rather than just *revealed* by it to the theorist.

Mentalism, the intrinsic, and sources of first order conscious representation

On a *mentalistic view of the intrinsic* , however, it is not some particular intrinsic property that confers consciousness on otherwise nonconscious states. Mental states *are* conscious states, and we need to recognize what is intrinsic to those conscious states. What is intrinsic, I suggest, is what is described in the mentalistic metatheory: Conscious modes that may be of various kinds – perception, belief, intention, and wish; modes that vary quantitatively in strength. Modes that can be propositional or non-propositional for the kinds of contents they carry. Contents with infinite variety that can symbolically represent the world beyond, or states and episodes within, in reality or non-reality, and in the past or present or future. Contents that can participate in deliberative and evocative episodes. Codes that can be identification or literal, conceptual or analog. A sense of possession that can vary in strength with persons and conditions. In fact, all that is intrinsic to conscious states is what can combine lawfully in the scientific investigation of consciousness.

In this sense, then, I would endorse an "intrinsic" theory of consciousness, but it would not be a higher order theory of consciousness.

The *source of a first order conscious state*? In summary terms, the first order perceptual consciousness is explained by covariation with stimulus values that activate conscious sensory content (through transduction), content that may activate conscious perceptual content, with attentional orientation and expectation variably entering. On the mentalistic metatheory, the route from sensory to perceptual is an evocative episode from literal awareness to identity awareness, an attentional content. Analogously, first order beliefs and intentions can follow with nonconscious operation upon prior conscious states with propositional contents; and later remembrances can follow from consciously represented cues that activate neural networks. These are processes with well studied underlying neural routes.

All of these, too, are processes that occur without the help of the complex and variable conditions for their occasional and sometimes fallible higher order representations. Put simply, with no HOR there or even potentially there, the first order conscious state would still be there. That conclusion, at least, would also agree with what I find in the Dretske (1995) and Tye (1995) strains of the philosophical literature.

Misguided rationales for higher order thought theories

"What it's like"?

Someone entering this literature without knowing "what it's like to be" a philosopher may be surprised by the almost obsessive concern with Nagel's famous expression as pointing to a characterization of conscious states. That characterization by its nature does require a higher order representation when theorizing, but this higher order theorizing is then oddly attributed to the subject of the theory – though made conveniently unconscious or only dispositional in non-theorizing main-streeters. Isn't characterizing "what it's like to be" simply a theorist's use of that representation to identify something intrinsic and perhaps unique to a conscious state?

A close cousin is the common intuition that "If we are aware of something, we can be aware of that awareness" – the flaw in the intuition, and in the theories it motivates, becoming clear with better answers to "How?" and "How often?"

Naturalizing?

I also find it odd that HOR theories are generally introduced by expressing a need to "naturalize" consciousness (as though it weren't "natural" enough as we find it). This apparently means either placement of conscious states within a fully deterministic, causal order (e.g. Carruthers 2000: 1–5) or affirming a physicalistic, material ontology (e.g. Rosenthal 2000: 735–737). First of all, the "hard problem of causation" is as hard as the hard ontological problem. In theories and in experimentation, causal analyses are necessarily selective in the moment and incomplete over time. Determinism, like materialism, can be a useful working assumption and at this point no more. I don't believe, however, that asserting that conscious mental states are set apart by concurrent or potential higher order representations can support either underlying position, as metaphysical commitment or working assumption.

Nonconscious mental states?

The most strongly reiterated aim seems to be a felt "need" to distinguish conscious mental states from what are said to be nonconscious mental states, arguing from various literatures in psychology. The HOR theorists are correct in saying that there has been some degree of consensus in psychology that non-

conscious mental states exist. As I have suggested (Dulany 1991, 1997, 2001, 2003), however, I believe this consensus would be better explained by the social psychology of the science than by a defensible appraisal of the evidence.

Why? From Watson (1924) to Wegner (2002) we have seen a common confusion of metatheoretical assertions of conscious control with two kinds of metaphysical assertions considered by many to be antithetical to science – those asserting non-materialism and indeterminism. Science could have no use for a soul of theology or free will in the sense of choices outside the physical causal order. Indeed, the same confusion, or at least blurring of distinctions, seems to animate eliminative materialism and homage to Rylean ancestry (e.g. Churchland 1993; Dennett 1991). The consequences? De-emphasize or reject consciousness in favor of a number of strategies that "put consciousness in its place" (Dulany 2003). At a recent conference, for example, two cognitivists at the table referred dismissively to the "C-word."

Among the strategies I have discussed is emphasis upon a "cognitive unconscious," meaning a working memory with fully formed mental episodes outside conscious attention (e.g. Baddeley 1986) and a level of cognition with consciousness only an occasional and noncausal emergent (e.g. Jackendoff 1987). With this kind of metatheoretical commitment, there has also been a strong, and socially supported, experimental strategy of searching for evidence of unconscious mental states; and as we need to recognize, methodological constraints can relax to accommodate deeply entrenched metatheory – especially when ideologically motivated.

To be brief, as I must, this will only illustrate and point. Interestingly, the literature drawn on most, and without noticeable regard for critiques, is that purporting to show "unconscious perception," what is probably the most controversial literature in our discipline.

Unconscious perception in normals? The investigation begins most prominently with McKay's (1973) evidence for what some now call "unconscious hearing" – using "subjective reports" of non-awareness. Presenting 27 ambiguous sentences to one ear with a disambiguating word to the other, and with instruction to track the one and ignore the other, subjects later reliably pick as presented before the one of two sentences consistent with the disambiguation – by interpretation, due to the "unconsciously perceived" disambiguation. Newstead and Dennis (1979) suspected, as many would, that normally curious undergraduates, instructed never ever to listen with one ear, might sometimes shift attention to that ear, then with professed obedience deny doing so when later asked for their "subjective report" of awareness – on only the last of 27 trials. They first replicated the McKay effect, showing mastery of his pro-

cedures, then in two subsequent experiments, removed what might artifactually draw attention to the forbidden ear – a time lag between sentences and the disambiguating word popping out of silence. With that change, evidence for "unconscious hearing" was no longer replicated. The widely read Velmans (1991:653) reported that the McKay finding "was replicated by Newstead and Dennis (1979)."

With the classical critique of the "subjective report" method by Holender (1986) (sometimes disregarded in prominent work, e.g. Mack & Rock 1998), the focus shifted to what has been known since Reingold and Merikle (1988) as the "direct objective" assessment of awareness of the stimulus. On this method, awareness of stimulus values is revealed when reports of stimulus values (sometimes as presence and absence) reliably covary with those stimulus values, as measured by the Signal Detection Theory (SDT) index of d' > 0. In fact, for decades workers with SDT had rigorously established that subjective reports of "I saw (or heard) ___" varies with the criterion subjects set for saying so (e.g. Macmillan & Creelman 1990). On this logic, then, claims of unconscious perception would require that an impoverished stimulus prime occurrence of something like a semantic associate, measured by a reliable "indirect objective" d', when d' for the "direct objective" measure of awareness is essentially zero. The d' is independent of criterion bias.

Nevertheless, Reingold and Merikle (1988) correctly argued that "null awareness" may be impossible to determine. In fact, all that is needed for a spurious finding of "unconscious perception" is conditions in which the direct measure is obtained with less sensitivity than the indirect. With use of a simultaneously masked stimulus, this kind of evidence can be ephemeral, as in Greenwald, Klinger, and Schuh (1995), where superiority of the indirect measure was found in two experiments (those usually cited), but not with two closely similar and four somewhat similar tasks – presumably because of variations in the relative sensitivity of the measures.

This lesser sensitivity of the direct measure can occur in various ways already critiqued (e.g. Dulany 1997, 2001; Holender 1986; Perruchet & Vinter 2002) – for example, a direct assessment at another time with fatigue and boredom, or prior to light adaptation, or on a smaller number of subjects, or even with different subjects or stimulus material. Another common way is the use of backward instead of simultaneous masking of the stimulus. Under comparable conditions, Greenwald and Klinger (1990) found semantic priming only with backward masking. This can permit literal awareness of the stimulus to prime the indirect measure's response – prior to the mask prohibiting the attentional identification required for remembering the stimulus at the direct measure-

ment (as elaborated in Dulany 2001). Haase and Fiske (2001, in press) significantly reveal still more conditions in which dissociation of direct and indirect measures can misleadingly suggest unconscious perception.

Still another way, surprisingly, is to give the role of "indirect" measure of "unconscious perception" to what is actually a "direct" measure of first order awareness, for example, reports of identity of one of four colors presented, then give the role of "direct" measure to the relation between correctness of response and reports of "high confidence" vs. "low confidence" in correctness. That confidence, of course, is actually a higher order awareness reflecting with still more error (and less sensitivity) the delay and necessary evaluation and inference required. For a reliable d' with the confidence measure, Kunimoto, Miller, and Pashler (2001:294) report that the necessary stimulus "durations were slightly but significantly longer".

They do, however, usefully show that still longer stimulus durations were required for subjects to abandon reports of "only guessing" or "at chance" – another higher order belief sometimes providing the basis for claims of unconscious perception. We can see why this is an even less sensitive higher order representation of first order awareness of a stimulus if we ask how one could believe otherwise of a weak but reliable relationship: Remember every R1 or R2 and to which S1 or S2, imagine the necessary 2X2 table, compute a X^2 in the head, and remember its position in its sampling distribution?

Recognizing reasons why a direct objective measure may not be sensitive enough to be exhaustive, some have turned to Jacoby's famous "process dissociation procedure" (e.g. Debner & Jacoby 1994) on which failure to follow instruction and exclude response to an impoverished stimulus is said to reveal its unconscious perception.. What Snodgrass' (2002:546) sophisticated SDT analysis shows, however, is that "exclusion methods are actually *subjective threshold* paradigms in disguise and as such are vulnerable to SDT-based criterion artifact. . ." – in essence, a return to the first procedure discredited. That being the case, it is unsurprising that Visser and Merikle (1999) showed that exclusion failures could be reduced even beyond controls by rewarding exclusion – the traditional experimental procedure for manipulating an SDT criterion of report.

Reflecting on this literature, early proponents of unconscious perception, Merikle and Reingold (1998:309) wrote, "We doubt that it will be possible ever to prove the existence of unconscious perception." If we examine this literature and invoke Bayesian relative credibility, we can ask which is more credible: These biased procedures produce spurious "unconscious perception" effects. Or evolution has prepared us to perceive unconsciously but it shows up only

in experiments that fail to replicate or embody biased procedures. Although introspective reports of conscious states may show exceptionally strong orderliness among themselves and to action (as in Dulany 1968; Carlson & Dulany 1988; Dulany 1997), they have an inherent degree and likelihood of error preventing their use for credible claims that some effect occurred "without awareness."

Unconscious perception with brain damage? With normal subjects, the results reveal small and transient dissociations within consciousness (or between consciousness and reports) following from experimental artifacts. With specific brain damage, however, studies have revealed dramatic, enduring, and conceptually significant dissociations – more credibly *within* consciousness, we can see, than between consciousness and unconscious perception. (Dulany 2000).

1. *"Blindsight"?* With damage to striate cortex, V1, some subjects, most famously DB and GY, show significant discrimination of various stimuli (e.g. direction of movement, nature of a figure) – what many have taken as revealing unconscious perception. What is ignored is that this is the familiar "direct objective" measure of first order awareness: No matter how low the confidence, he reports a conscious belief on every trial that the stimulus has one value rather than another. Under certain conditions, this discrimination is reported to be superior to that on the now standard "commentary key" in which he was "instructed that he was to press the 'yes' key whenever he had *any* experience whatsoever of the visual event" (Weiskrantz 1997: 64) – clearly a report of higher order awareness of first order seeing. Awareness of a stimulus too degraded for the usual categorization as "seeing", or even categorization as "awareness", is readily explained in these cases by residual or recovered function within V1 in DB and GY, and in these and other cases by transmission from the retina that by-passes the dorsal lateral geniculate route to V1, going instead via the superior colliculus to other parts of the visual system, V2, V3, V4 – alternative routes that Weiskrantz (1997: 128) also discusses. Indeed, an fMRI study has shown that visual stimuli may activate extra-striate cortex even when striate cortex is not activated (Stoerig, Kleinschmidt, & Frahm 1998). Despite the Oxonian oxymoron "blindsight," all this is significant evidence for a dissociation *within* consciousness, not between consciousness and unconscious perception (Dulany 2000, 2002).
2. *Unconscious perception in prosopagnosia*? With damage to the occipito-temporal cortices around the fusiform gyrus, persons may fail to recog-

nize faces while still able to respond in certain systematic ways to those faces – originally interpreted as a kind of unconscious perceptual recognition (e.g. Young 1994). For example, Tranel and Damasio (1985) found greater galvanic skin response (GSR) to familiar than to unfamiliar faces. Central to the mentalistic explanation is a neural network through which literal awareness of a face ordinarily activates identity awareness, but may still activate other associations despite that ordinary route being severed. With Farah (1994) I would agree that the visual recognition system is damaged, and that residual activation can produce the observed systematic response to faces despite not reaching temporal areas needed for recognition. What can still be represented in the recognition system, I would add, is literal awareness of facial form, as shown in ability to match facial photos whether familiar or not (DeHaan, Young, & Newcombe 1987). We also have no evidence that consciousness is a localized system in the brain that could be "disconnected" from an intact recognition system. Again this is evidence for a significant dissociation *within* consciousness, not perceptual recognition dissociated from consciousness (Dulany 2000, 2002).

3. *Unconscious perception within the "dorsal stream"?* Milner and Goodale (1995) have identified significant dissociations between psychological functions associated with a "ventral visual stream" (. . .occipital, temporal. . .) and a "dorsal visual stream" (. . .occipital, parietal. . .), as shown in an array of demonstrations. The extensively studied patient, DF, for example, suffers from an agnosia preventing immediate recognition of various ordinary objects or geometric forms, or the orientation of a slot, as a result of anoxia producing limited occipito-temporal damage – damage to the ventral stream. Nevertheless, she can draw from memory, reach out and grasp ordinary objects, move her hand with correct orientation for placing a card in the slot – a dissociated maintenance of the dorsal stream.

   Although "unconscious perception" in the dorsal stream has been claimed, given maintenance of "relatively normal low-level visual functions" (1995: 126), the authors disavow a firm conclusion to that effect (1995: 200–201) – and quite reasonably so. The accurate actions of grasping, orienting a card for a slot, etc. are well explained by the preserved "low-level visual function", the literal awareness that activates memory and participates in drawing of those objects and the guidance of previously learned, even automatized, actions. Preserved abilities in these "object agnosias" should no more require unconscious recognition than does prosopagnosia. This is another significant dissociation *within* conscious-

ness, between literal and identity awareness, with literal awareness activating residual capabilities in the undamaged dorsal stream.

Claims for other nonconscious states? I must now only point to methodological and conceptual critiques, containing references to still other critiques: For automaticity, change detection, concept learning, creativity, implicit learning, implicit memory, inattentional "blindness", metacognitive control, neglect, problem solving, syntactical learning, word extraction, etc., see e.g. Carlson (1997, 2002); Carlson and Dulany (1985); Dulany (1991, 1997, 2001, 2002, 2003); Ericsson and Simon (1993); Farah (1994); Mandler (1994); Mitroff, Simons, and Franconeri (2002); O'Brien and Opie (1999); Shanks and St. John (1994); Tzelgov (1997), etc.

In short – as I have written before (Dulany 1999), if claims for the power of a cognitive unconscious were correct, the experimental effects should be too strong and replicable, too methodologically and conceptually defensible, for these literatures even to be controversial. No one can claim that.

But could adoption of a HOR theory of consciousness actually predispose investigators to methodological error? If HOR theory is used to say that failure of an insensitive assessment simply means that the first order state was nonconscious, that would constitute one more spurious rationale in defense of a deeply entrenched metatheoretical commitment. . .And in a similar way, if ambient illumination impairs sensitivity of an optical microscope, the bacterium just wasn't there.

Upshot

First of all, I believe the rationales that evidently have motivated these higher order theories have been fundamentally misguided. Furthermore, I find no acceptable evidence or argument that first order mental states require any kind of relationship to any kind of higher order representations in order to be conscious mental states. First order conscious states not only have an evolutionary warrant, we have long been learning about the processes that produce them – processes unaided by their only occasional and fallible higher order representations. Perhaps the strongest impression I take away from this literature is that the theories have things backwards in the most fundamental way: It isn't a higher order representation, either actual or potential, that confers consciousness on a lower order mental state. It is what is intrinsic in lower order conscious states and their roles in establishing memory and inference that

sometimes leads to the their higher order conscious representations. What I have found most valuable in this literature is the sporadic concern for what is intrinsic – and unique – to conscious mental states.

References

Anderson, J. R. (2000). *Cognitive psychology: Its implications*. New York: Worth Publishers and W.H. Freeman.

Armstrong, D. (1968). *A materialist theory of the mind*. New York: Humanities Press.

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*, 227–247

Brentano, F. (1774, 1973) *Psychology from an empirical standpoint*. London: Routledge & Kegan Paul.

Carlson, R. A. (1997). *Experienced cognition*. Mahwah, NJ: Erlbaum.

Carlson, R. A. (2002). Conscious intentions in the control of skilled mental activity. *The Psychology of Learning and Motivation, 41*, 191–228.

Carlson, R. A. & Dulany, D. E. (1985). Conscious attention in abstraction and concept learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 11*, 45–58.

Carlson, R. & Dulany, D. (1988). Diagnostic reasoning with circumstantial evidence. *Cognitive Psychology, 20*, 463–492.

Carruthers, P. (2000). *Phenomenal consciousness: A naturalistic theory*. Cambridge, UK: Cambridge University Press.

Caston, A. (2002). Aristotle on consciousness. Mind, 111, 752–815.

Chalmers, D. (1996). *The conscious mind*. New York: Oxford University Press.

Churchland, P. M. (1993). Eliminative materialism and the propositional attitudes. In A. I. Goldman (Ed.), *Readings in philosophy and cognitive science* (pp. 255–270). Cambridge, MA: MIT Press.

Churchland, P. S. (1986). *Neurophilosophy*. Cambridge, MA: MIT Press.

Cronbach, L. J. & Meehl, P. (1955). Construct validation in psychological tests. *Psychological Bulletin, 52*, 281–382.

Debner, J. A. & Jacoby, L. L. (1994). Unconscious perception: Attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 304–317.

Dehaan, E. H. F., Young, A., & Newcombe, F. (1987). Face recognition without awareness. *Cognitive Neuropsychology, 4*, 385–415.

Dennett, D. (1991). *Consciousness explained*. Boston, MA: Little, Brown.

Dretske, F. (1995) *Naturalizing the mind*. Cambridge, MA: MIT Press.

Dulany, D. E. (1968). Awareness, rules, and propositional control: A confrontation with S-R behavior theory. In T. Dixon, & D. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 340–387). New York: Prentice Hall.

Dulany, D. E. (1984). A strategy for investigating consciousness. Psychonomic Society Meeting, San Antonio, TX.

Dulany, D. E. (1991). Conscious representation and thought systems. In R. S. Wyer & T. K. Srull (Eds.), *Advances in social cognition* Vol. 4, (pp. 97–120). Hillsdale, NJ: Erlbaum.

Dulany, D. E. (1997). Consciousness in the explicit (deliberative) and implicit (evocative). In J. Cohen & J. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 179–212). Mahwah, NJ: Lawrence Erlbaum Associates.

Dulany, D. E. (1999). Consciousness, connectionism, and intentionality. *Behavioral and Brain Sciences, 22*, 154–155.

Dulany, D. E. (2000). A mentalistic view of conscious unity and dissociation. Association for the Scientific Study of Consciousness, Brussels, Belgium.

Dulany, D. E. (2001), Inattentional awareness, *Psyche, 7(05)*, http://psyche.cs.monash.edu.au/v7/psyche-7-05-dulany.html.

Dulany, D. E. (2002) Mentalistic metatheory and strategies. *Behavioral and Brain Sciences, 24*, 337–338.

Dulany, D. E. (2003). Strategies for putting consciousness in its place. *Journal of Consciousness Studies, 10*(1), 33–43.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge: The MIT Press.

Farah, M. (1994). Visual awareness and visual perception after brain damage: A tutorial overview. In C. Umilta, & M. Moscovitch (Eds.), *Attention and performance XV. Conscious and nonconscious information processing* (pp. 37–76). Cambridge, MA: MIT Press.

Gardiner, J. M. & Richardson-Klavehn, A. (2000). In E. Tulving & F. I. M Craik (Eds.), *The Oxford Handbook of Memory.* New York: Oxford University Press.

Gennaro, R. (1996). *Consciousness and self-consciousness*. Amsterdam & Philadelphia, PA: John Benjamins Publishing Co.

Gennaro, R. (2002). Jean-Paul Sartre and the HOT theory of consciousness. *Canadian Journal of Philosophy, 32*, 293–330.

Graf, P, Squire, L., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 164–178.

Greenwald, A. G. & Klinger, M. R. (1990). Visual masking and unconscious processing: Differences between backward and simultaneous masking: *Memory & Cognition, 18*, 430–435.

Greenwald, A. G., Klinger, M. R., & Schuh, E. S. (1995). Activation by marginally perceptible ("subliminal") stimuli: Dissociation of unconscious from conscious cognition. *Journal of Experimental Psychology: General, 124*, 22–42.

Güzeldere, G. (1997). Is consciousness the perception of what passes in one's own mind? In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 789–806). Cambridge, MA: The MIT Press.

Haase, S. J. & Fisk. G. (2001). Confidence in word detection predicts word identification: Implications for an unconscious perception paradigm. *American Journal of Psychology, 114*, 439–468.

Haase, S. J. & Fisk, G., (in press). Valid distinctions between conscious and unconscious perception? *Perception & Performance.*

Holender, D. (1986). Semantic activation without conscious identification in dichotic listening, parafoveal vision and visual masking. *Behavioral and Brain Sciences, 9*.

Jackendoff, R. (1987). *Consciousness and the computational mind*. Cambridge, MA: MIT Press.

Jacoby, L. L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340.

Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Boston, MA: Houghton Mifflin Co.

Koriat, A. (2000). The feeling of knowing: some metatheoretical implications for consciousness and control. *Consciousness and Cognition, 9*, 149–171.

Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness & Cognition, 10*, 294–340.

Lycan, W. (1996). *Consciousness and experience*. Boston, MA: MIT Press.

Lurz, R. W. (2003). Neither "hot" nor "cold": An alternative account of consciousness. *Psyche, 9*(01), http://psyche.cs.monash.edu.au/v9/psyche-9-01-lurz.html

Mack, A. & Rock, I. (1998*). Inattentional blindness*. Cambridge, MA: The MIT Press.

MacKay, D. G. (1973). Aspects of the theory of comprehension, memory and attention. *Quarterly Journal of Experimental Psychology, 25*, 22–44.

Macmillan, N. A. & Creelman, D. C. (1990). Response bias: Characteristics of detection theory, and "nonparametric" indices. *Psychological Bulletin, 107*, 401–413.

Mandler, G. (1994). Hypermnesia, incubation, and mind popping: On remembering without really trying. In C. Umilta & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 4–33). Cambridge, MA: MIT Press.

Merikle, P. M. & Reingold, E. M. (1998). On demonstrating unconscious perception: Comment on Draine and Greenwald (1998). *Journal of Experimental Psychology: General, 127*, 304–310.

Metcalf, J. (2000). Feeling and judgments of knowing: Is there a special noetic state? *Consciousness and Cognition, 9*, 178–186.

Milner, A. D. & Goodale, M. A. (1995). *The visual brain in action.* New York: Oxford University Press.

Mitroff, S. R., Simons, D. J., & Franconeri, S. L. (2002). The siren song of implicit change detection. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 798–815.

Natsoulas, T. (2001). On the intrinsic nature of states of consciousness: Attempted inroads from the first-person perspective. *The Journal of Mind and Behavior, 22*, 219–248.

Nelson, T. O. (1996). Consciousness, self-consciousness, and metacognition. *Consciousness and Cognition, 9*, 220–223.

Newstead, S. E. & Dennis, I. (1978). Lexical and grammatical processing of unshadowed messages: A re-examination of the Mackay effect. *Quarterly Journal of Experimental Psychology, 31*, 477–488.

O'Brien, G. & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences, 22*, 127–196.

Perruchet, P. & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Science, 25*, 297–329.

Reingold, E. M. & Merikle, P. M. (1988). Using direct and indirect measures to study perception without awareness. *Perception & Psychophysics, 44*, 563–575.

Roediger, H. L. & Scrinivas, K. (1993). Specificity of operations in perceptual priming. In P. Graf & M. E. J. Masson (Eds.), *Implicit memory: New directions in cognition, development, and neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, D. (1986). Two concepts of consciousness. *Philosophical Studies, 94*, 329–359.

Rosenthal, D. (1998). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates.* (pp. 729–754). Cambridge, MA: The MIT Press.

Rosenthal, D. (2000a). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition, 9*, 203–214.

Rosenthal, D. (2000b). Metacognition and higher-order thoughts. *Consciousness and Cognition, 9*, 231–242.

Schooler, J. W. (2002). Re-representing consciousness: Dissociations between experience and meta-consciousness. *Trends in Cognitive Science, 6*, 339–334.

Shanks, D. R. & St. John, M. F. (1994) Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences, 17*, 367–447.

Snodgrass, M. (2002). Disambiguating conscious and unconscious influences: Do exclusion paradigms demonstrate unconscious perception? *American Journal of Psychology, 15*, 545–580.

Stephens, G. L. & Graham, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*. Cambridge, MA: MIT Press.

Stoerig, P., Kleinschmidt, A., & Frahm, J. (1998). No visual response in denervated V1: High-resolution functional magnetic resonance imaging of a blindsight patient. *Neuroreport, 9*, 21–25.

Tranel, D. & Damasio, A. R. (1985). Knowledge without awareness: An autonomic index of facial recognition by prosopagnosiacs. *Science, 228*, 1453–1454.

Tye, M. (1995). *Ten problems of consciousness.* Cambridge, MA: MIT Press.

Tzelgov, J. (1997). Specifying the relations between automaticity and consciousness: A theoretical note. *Consciousness and Cognition, 6*, 441–451.

Young, A. W. (1994). Conscious and nonconscious recognition of familiar faces. In C. Umiltà & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 153–178). Cambridge, MA: MIT Press.

Velmans, M. (1991). Is human information processing conscious?' *Behavioral and Brain Sciences*, *14*, 651–726.

Watson, J. B. (1924). *Behaviorism.* New York: W. W. Norton & Co.

Wegner, D. M. (2002). *The illusion of will.* Cambridge, MA: MIT Press.

Weiskrantz, L. (1997). *Consciousness lost and found.* Oxford: Oxford University Press.

Chapter 15

Ouch! An essay on pain*

Christopher S. Hill

I will be concerned in this paper with four topics. First, I will try to show that awareness of pain is deeply akin to such paradigmatic forms of perceptual awareness as vision and hearing, and that it is therefore appropriate to see it as fundamentally perceptual in character. Second, I will try to show that despite vivid intuitions to the contrary, it is possible to draw a distinction between its seeming to one that one is in pain and its being the case that one is in pain. In other words, I will attempt to show that it is possible to draw an appearance/reality distinction with respect to pain. Third, I will argue that it is appropriate to identify pains with certain physical process that occur in the body. To be more specific, I will urge that it is appropriate to identify pains with bodily disturbances that involve actual or potential damage. Finally, having developed a theory of pain and awareness of pain, I will consider this theory in relation to the "higher order perception" account of what it is for a pain to be conscious. Like the theory I will recommend, the higher order perception theory endorses a form of the view that there is perceptual awareness of pain. I will be concerned to chart the area of overlap between the theories, to identify the points of divergence, and to argue for the superiority of the former theory.

1.

Awareness of pain often takes the form of a judgment to the effect that one is in pain, but there are grounds for thinking that there is also a more fundamental form of awareness of pain, a form that is subdoxastic and subconceptual in nature. To see this, observe that when we are aware of a pain, we are often aware that it has a complex internal organization that cannot readily be described in terms of the general concepts that are available to us. Pains generally have parts of different intensities and different qualities (where by "qualities" I mean

characteristics like throbbing, burning, and piercing), and these parts tend to support complex relational structures like those of mosaics. We do not have general concepts that uniquely capture these complex relational structures, so, if we wish to describe a complicated pain, we are reduced to enumerating its parts, specifying the intensities and qualities of these parts, and then listing the relational facts in which the parts figure as constituents. Now it is generally impossible to complete such a description before the pain has changed in some fundamental respect, and this means that the task of describing it has to be abandoned, because we will be unable to remember all of its individuating details. Thus, we cannot be said to have a conceptual grasp of the internal organizations of complex pains. We cannot be said to grasp their natures by acts of judgment. But still, at least when we are being attentive, we do grasp their natures. There is a form of awareness of pains that enables us to take in their internal organizations all at once, much as a photograph can simultaneously do justice to all of the details of a complex situation.

It might be said that it is possible to achieve conceptual awareness of the structural complexities of pains by employing certain demonstrative concepts – specifically, concepts of the forms *the internal structure of this pain* and *this internal structure*. But this proposal suffers from disabling flaws. To mention only one, a demonstrative cannot refer to the internal structure of a pain independently of one's experiential awareness of the pain. On the contrary, if a demonstrative refers to the internal structure of a pain, this is because one is experientially aware of the pain – the demonstrative borrows its content from the state of awareness. Accordingly, it would be circular to attempt to account for the content of a state of experiential awareness by appealing to the content of a demonstrative that accompanies that state of awareness.[1]

In view of considerations of this sort, it seems appropriate to claim that there is a form of awareness of pain that is subdoxastic and subconceptual. I will now give two reasons for thinking that this form of awareness, which I will call *experiential awareness*, should be classified as fundamentally perceptual in character.

In the first place, when one reflects on the claims that folk psychology makes about the paradigmatic forms of perception, and considers them in relation to our everyday experience of pain, one comes to see that there are a number of striking similarities between the paradigmatic forms and awareness of pain. I will cite six of these similarities:

First similarity: The content of a state of perceptual awareness can transcend our conceptual repertoire. Consider vision: it is possible to be visually aware of

a color or a shape even though one has no concept that stands for that color or shape. As we just noticed, a similar principle applies to experiential awareness of pains: it is possible to be aware of types of pain that one is unable to describe adequately.

Second similarity: Unlike the contents of a conceptual representation, the content of a state of perceptual awareness is specific, representing not just determinable properties, but also determinate forms of those properties. Thus, unlike a judgment that represents an object as blue, a perceptual state that represents an object as blue will also represent the object as being of a particular shade of blue. A similar principle applies to pains: it is impossible to be experientially aware of a pain as intense without representing it as being of a particular level of intensity.

Third similarity: As with visual and other forms of perceptual awareness, awareness of pain is associated with mechanisms for directing and enhancing attention. Moreover, attentive awareness of pain has effects like those of attentive visual awareness. Experiments by Marisa Carrasco and her associates have shown that visual attentiveness increases spatial resolution, and that it also increases the contrast of a stimulus with its background. (See, e.g., Yeshuran and Carrasco (1998).) I know of no studies which show that attention to pain has similar effects, but introspection convinces me that it has such effects in my own case.

Fourth similarity: As in the case of standard forms of perceptual awareness, when one is experientially aware of a pain, one experiences the object of awareness as having a location in physical space. Moreover, as in the case of vision and touch, one experiences the object of awareness as having spatial properties in addition to location. For example, one normally experiences it as having a certain more or less determinate size.

Fifth similarity: Unlike cases in which one's awareness of a phenomenon takes the form of a conceptual representation, when one is visually aware of a phenomenon, one is also aware of a broad range of parts of that phenomenon, all the way down to the parts that count (relative to one's perspective) as minima sensibilia. In short, a visual experience that represents a whole also represents the parts of that whole. The same is true, mutatis mutandis, of experiential awareness of pain: if, e.g., you are experientially aware of a pain that runs from

your left wrist up to your elbow, you are ipso facto aware of the painfulness of all of the intermediate parts of your forearm.

Sixth similarity: When we are perceptually aware of an object as having one property, we are also aware of it as having other properties. For example, it is impossible to be aware of the pitch of a sound without also being aware of its volume. The same is true in the case of experiential awareness of pains. When we are aware of the quality of a pain (that is, when we are aware of it as being a stabbing pain, or as being a burning pain, etc.), we are also aware of the pain's intensity.

It is clear, then, that there are a number of points of correspondence between paradigmatic forms of perception and experiential awareness of pain. These points of correspondence strongly suggest that that it would be both legitimate and fruitful to regard experiential awareness of pain as being on a par with the paradigmatic forms. But there is also another reason for embracing this view.

This second reason has to do with certain developments in cognitive neuroscience. These developments indicate that the structures and processes that underlie and explain awareness of pain are fundamentally akin to the structures and processes that underlie paradigmatic forms of perceptual awareness. Thus, like vision and hearing, awareness of pain is supported by a complex system – the *nociperceptual system*, to give it a name – that involves high level representations and complex processing devices. To be more specific, like vision and hearing, the nociperceptual system makes use of hierarchical systems of feature detectors, maps of the perceptual environment, attention mechanisms, devices that enable form enhancement and "filling in," and devices that reduce noise. Also, again like vision and hearing, the nociperceptual system includes sense organs that are designed to pick up and transmit information about a domain that is of importance to the organism. For it is true that the C fibers and A-δ fibers perform the functions of sense organs. These tissues spread throughout the body, detecting conditions of actual and potential damage, and sending messages to the cortex that specify the kinds, degrees, and locations of such conditions.[2]

We have found two reasons for embracing the hypothesis that there is a form of awareness of pain is fundamentally perceptual in nature. I now observe that one of the main principles governing paradigmatic forms of perception is the principle that perception can fail as well as succeed in its representational endeavors. According to this principle, in any case of perceptual representation,

it is possible to distinguish between a perceptual appearance and the underlying reality, or in other words, it is possible to distinguish between its seeming to one that a certain state of affairs obtains and its being the case that the state of affairs obtains. Combining this principle with our hypothesis about the nature of experiential awareness of pain, we arrive at the corollary that it is always possible to distinguish between its seeming to one experientially that one is in pain and its being the case that one is in pain.

I do not expect this corollary to meet with immediate acceptance. On the contrary, it seems likely that the reader will at first be strongly inclined to reject it. I will now attempt to show that despite its initial implausibility, it merits sympathetic consideration.

2.

Most of us have a vivid and stable intuition to the effect that pain is *self presenting*. This intuition can be expressed simply by saying that it is impossible distinguish between its seeming to one that one is in pain and its being the case that one is in pain. But there are other ways of formulating the intuition. For example, it can be expressed by saying that if one has an experience of the sort that makes it natural and appropriate to judge that one is in pain, then one really is in pain. It seems that in one version or another, the intuition is acknowledged by most philosophers who have written about pain, and it seems to be shared by non-philosophers. It appears, for example, to command the respect of students who are being exposed to philosophy for the first time.

In the present section I will attempt to locate the sources of this intuition and to determine whether they provide grounds for thinking that it is correct. In the end, I will conclude that there are no considerations that make it epistemically necessary to embrace the intuition, and further, that it is possible to imagine situations in which one would be inclined to set the intuition aside.

As I see it, the intuition has three main sources.

In the first place, there is no form of experiential access to pains other than introspective awareness. It follows that there is no epistemically fundamental way of checking up on the accuracy of introspection, that is, of determining whether its deliverances concerning pain might be out of step with the reality of pain. Thus, we cannot observe pain directly by using vision or touch, and since it is impossible for others to be introspectively aware of one's pains, there is no body of third person testimony that can be used to determine whether introspection provides a trustworthy guide to pain. We have to take the de-

liverances of introspection at face value because there is no way of checking up on them.

It appears, however, that information derived from checking procedures generally plays a crucial role in providing motivation for a distinction between appearance and reality. Thus, it seems to be generally true that if we are prepared to draw such a distinction with respect to a mode of awareness A, then there is a second, independent mode of awareness A' such that (i) A' provides us with access to the same domain as A and (ii) we regard A' as being no less reliable than A. This is certainly true of the paradigmatic forms of perceptual awareness, and it is true of memory as well. To be sure, it is probably not absolutely essential for there to be independent access to a domain in order for us to be wiling to draw an appearance/reality distinction. One might, for example, be led to draw such a distinction if one found that a sense modality encouraged contradictory beliefs, or if one found that some deliverances of a modality were considerably more confused than others. But there can be no doubt that in actual practice, the motivation provided by considerations of this kind is secondary to the motivation provided by alternative forms of awareness.

It seems, then, that if we are to list the factors that are responsible for the intuition that pains are self presenting, it is appropriate to begin with the fact that there is no fundamental mode of access to pains other than introspection. But more needs to be said here, for what has been said so far fails to address the question of *why* introspection provides our only fundamental form of access to pains. After all, if we regarded pains as identical with bodily phenomena of some sort, such as disturbances involving actual or potential damage, we would hold that both vision and touch provide independent access to pains, and we would regard vision and touch as providing information that could be used to assess the credentials of introspection. Accordingly, if we wish to take the present explanation of the intuition that pains are self presenting to a reasonably deep level, we must ask why it is that we are not motivated to identify pains with bodily phenomena. In particular, we must ask why we are not motivated to think of them as identical with bodily disturbances involving actual or potential damage.

Part of the reason is surely that introspection fails to put us in touch with any of the properties that we regard as constitutive or symptomatic of bodily damage. It provides no foothold for the concept of a puncture wound, or for the concept of an open sore, or for the concept of a burn. Nor does it provide any foothold for such concepts as blistering, bleeding, and swelling. Further, what is probably much more important, introspective recognition of pains as pains appears to depend on the fact that pains occasion certain responses in us.

They cause various forms of psychological distress, such as anxiety, anger, and fear, and they also have a powerful ability to command attention – or as some writers say, to invade consciousness and take control of it. It is plausible, I think, that we rely on such responses in determining whether an object of awareness is a pain. That is to say, as I see it, these responses are partially constitutive of the recognition process itself. Now of course, visual and tactile awareness of bodily disturbances have no immediate tendency to produce such responses. It follows that crucial components of the recognitional system that is responsible for awareness of pain are not engaged by awareness of bodily disturbances.

In saying that our psychological responses to pain are partially constitutive of the process by which pains are recognized, I do not mean only to be endorsing the fairly weak claim that we are not inclined to judge that an object of awareness is a pain unless it produces a certain psychological response in us. Rather, I mean to be claiming, in addition, that we are inclined to judge that an object of awareness is *not* a pain unless it produces a certain psychological response. There are various reasons for thinking that this second, stronger claim is true. Thus, for example, it receives support from Ramachandran's explanation of the Capgras delusion. Victims of this delusion are able to determine that the people in their environment *look just like* certain people who were familiar to them prior to the onset of their delusion, but instead of recognizing these people as identical with their familiars, the victims take them to be mere duplicates – duplicates who have somehow supplanted their family members and friends and who are continuing to play their roles. Examining one such victim carefully, Ramachandran found that their neural pathways leading from the perceptual centers to the emotional centers were damaged, and went on to explain his patient's curious take on his parents by hypothesizing that the patient's inability to experience familiar emotions as a result of perceiving his parents made it impossible for him to recognize his parents. This hypothesis was confirmed by various considerations, not least of which was the fact that the patient regained his ability to recognize his parents when the damaged neural pathway regained its ability to carry signals.[3]

It appears, then, that there are features of the recognitional system that is involved in awareness of pain that tend to foster the perception that pain is ontologically independent of bodily disturbances and of physical phenomena generally. Of course, in fostering this perception, the features in question also foster the perception that we cannot use the modes of awareness that provide information about the physical world to check up on what introspection tells us about pain, and by the same token, they foster the perception that there is no basis for distinguishing between its seeming to one that one is in pain and its

being the case that one is in pain. But now we must ask: Does the fact that these perceptions are fostered by features of the recognitional system that subserves awareness of pain provide a good reason for thinking that the perceptions are *correct*? It is clear that the answer is "no." It is clearly possible to appreciate the special features of the recognitional system that subserves introspective awareness of pain while holding that is necessary to go beyond the deliverances of introspection to get at the truth about the nature of pain. By the same token, it is possible to appreciate the special features the recognitional system while holding that when one does go beyond the deliverances of introspection, one finds strong theoretical grounds for identifying pains with certain bodily disturbances, and also for thinking that such an identification could be used to ground and motivate a practice of distinguishing between the appearances of pains and their underlying natures.

I began this discussion by claiming that there are three main sources of the intuition that pains are self presenting. I turn now to the second source.

When we consider introspective awareness of pain from a commonsense perspective, it appears that pain is given to us without the mediation of a sense organ, and a fortiori, without any informational uptake by a perceptual system. We are aware of nothing that corresponds to an eye or an ear, and we are aware of no physical processes corresponding to the sensory sampling that is involved in seeing and hearing.

It is evident that our awareness of the existence of sense organs and their role in the perceptual process helps to generate and to sustain our sense that the paradigmatic forms of perception employ sensory representations that are independent of the objects of perception. Thus, for example, if it seems to one that one is visually aware of something on an occasion when one's eyes are closed, one is in possession of a compelling demonstration that there is a difference between visual experience and the objects of visual awareness. On the other hand, since awareness of pain appears to occur without being supported by a sense organ, there is no evident foothold for the idea that such awareness comes at the end of a process of information-pickup and processing, and by the same token, no apparent foothold for the idea that there is a representation at the end of the process that is ontologically independent of events at earlier stages.

As I see it, then, if we are to explain why commonsense does not draw a distinction between its seeming to one that one is in pain and its being the case that one is in pain, we must mention the fact that commonsense is not apprised of a sense organ that subserves awareness of pain. But of course, the fact that commonsense is not apprised of such an organ has no tendency to show that

an organ does not exist. And in fact, as we observed in section I, there is a vast system of nerve fibers that seems to play the role of a sense organ in the process of detecting and classifying bodily disturbances involving actual or potential damage. If it should turn out to be possible to make a case for identifying pains with bodily disturbances of this kind, we could say that, *pace* common sense, awareness of pain resembles other forms of perceptual awareness in that it depends essentially on a sensory pickup system and on perceptual information processing.

I turn now to what I take to be the third main source of the intuition that pains are self presenting.

This third source is the fact that in the case of awareness of pain, there is no counterpart of the shifts in perspective that are possible in the case of the paradigmatic perceptual systems. In the case of the paradigmatic systems, it is possible to change one's perspective on an object of awareness by changing the position or orientation of the relevant sense organ with respect to the object of awareness. This gives one a certain amount of control over the nature of the information that one's perceptual processes make available. Indeed, it makes perceiving the object a voluntary affair. In the case of awareness of pain, however, one is not able to change the relationship between a sensory system and an object of awareness. As result, the nature of one's information about one's pains is largely fixed by processes that lie outside one's control. And, tragically, it is not possible to make one's pains disappear by turning one's head or walking away.

This difference between the standard perceptual systems and the faculty that supports awareness of pain helps to explain why we are not motivated to draw a distinction between the appearance of a pain and the underlying reality in our everyday thought and talk. As I see it, however, it does not pose a logical obstacle to drawing such a distinction. The idea of a perceptual system is the idea of a system that satisfies certain conditions. These conditions include such things as having a consistent sensory pick-up system that is designed to respond to stimuli in a certain preferred range, employing representations whose contents bear on the environment in certain characteristic ways (some of which are described in Section I), and interacting in certain characteristic ways with other cognitive faculties, such as the ones that are responsible for constructing enduring cognitive maps of the environment and guiding voluntary behavior. It seems much less plausible that the conditions that are constitutive of a perceptual system include having a constituent sensory pick-up system whose position and orientation can change over time. A fortiori, it seems much less plausible that the conditions include having a sensory pick-up system that is

under voluntary control. These are conditions that are satisfied in the cases that are most familiar to us, but they seem not to be constitutive of perceptual systems in general.[4]

I have cited three sources of the intuition that it is impossible to draw an appearance/reality distinction with respect to pain, and I have urged that none of them provides a ground for thinking that the intuition is correct. It seems, however, that even if one feels that these considerations have a certain appeal, one might still have a strong attachment to the intuition. With a view to bringing the remaining bonds of attachment to the fore, and hopefully dissolving them, let us consider an agent, Fred, who is on his way to see his doctor about a persistent severe pain in his lower back. Let us suppose that the doctor examines Fred and determines that he is altogether free from damage and disease. Suppose further that the doctor addresses Fred as follows: "My dear fellow, you are quite fortunate! There is absolutely nothing wrong with you, and therefore, however it might appear to you, you are not in pain. Your sense that you are in pain is a mere hallucination." Will Fred have any inclination to accept this judgment? Surely not. Surely Fred will refuse to pay the doctor's bill and will move on to another physician. In doing so, he will in effect be endorsing that intuition that it is impossible to distinguish between its seeming to one that one is in pain and its being the case that one is in pain. Moreover, since we are strongly inclined to applaud Fred's behavior, it is clear that we are tempted to endorse this intuition as well.

There are, however, a few considerations that may on reflection lead us to take a somewhat different view of the matter.

First, in evaluating the claim that a pain is hallucinatory, it is important to keep in mind the fact that hallucinations can be utterly compelling. When, for example, one has a visual hallucination to the effect that there is a dagger before one, one may actually be led to reach for the dagger, and may feel astonishment when one's fingers close on empty air. Or consider the auditory hallucinations that schizophrenics experience. In such cases, the impression that one is hearing a voice may be so strong as to make it seem natural and appropriate to deny the testimony of one's eyes, and to postulate an invisible speaker. There is no reason to suppose that nociperceptual representations should be less vivid, or less capable of compelling belief, than hallucinations of these other kinds.

Second, the claim that certain impressions of pain are hallucinatory is entirely compatible with allowing that those who have such impression are in states that fit the functional profile of pain. That is to say, it is entirely compatible with allowing that they are in states that cause such psychological phenomena as anxiety, distress, and a desire that pain cease, and that they also cause

such behavioral phenomena as wincing, crying out, and nursing of the apparent site of the pain. Thus, it seems entirely possible, and in fact quite likely, that there are *two* types of state that fit the standard functional profile of pain: bodily disturbances involving actual or potential damage fit the profile, but so do nociperceptual representations of such damage. To be more precise, it seems quite likely there are two groups of closely related functional properties, one consisting of properties that are possessed by certain bodily disturbances, and another consisting of properties that are possessed by nociperceptual representations. The members of the first group are what might be called *distal* causal powers, and the members of the second group are what might be called *proximal* causal powers. Thus, bodily disturbances of the given sort have the power to cause nociperceptual representations, and nociperceptual representations have the power to serve as *proximal* causes of such phenomena as psychological distress and wincing. Because of these facts, bodily disturbances can be said to have the power to serve as *distal* causes of these same phenomena. The distal causal powers of bodily disturbances are powers to serve as distal causes and the proximal causal powers of nociperceptual representations are powers to serve as proximal causes.

Given that nociperceptual representations satisfy the functional profile of pain, and given also that states of seeming to be aware of pain supervene on nociperceptual representations, we can say that states of seeming to be aware of pain satisfy the functional profile of pain as well. But this means that when it seems to one that one is aware of a pain, one will be in a state that satisfies the functional profile of pain, whether or not one's state of apparent awareness is veridical.

Third, observe that Fred's doctor was unnecessarily abrupt in responding to Fred's complaint. Let us imagine a different doctor who provides Fred with the lines of thought that have been put forward thus far in the present paper. That is, let us suppose that the doctor lays out the full case for thinking that awareness of pain is perceptual, together with a systematic critique of the intuition that pains are self-presenting. I would be reluctant to speak for Fred on most other issues, but I feel confident that after hearing all of the arguments, Fred would at least have some inclination to replace his original judgment with the claim that he was experiencing a nociperceptual hallucination. And I feel confident that this inclination would be strengthened if Fred were to be given the rationale for identifying pains with certain bodily disturbances that I will provide in the next section.

3.

For better or worse, I have now finished with my discussion of awareness of pain. I wish to turn now to consider the nature of pain itself. I will defend a strongly reductionist view about pain. Such views are familiar. Unlike standard reductionist proposals, however, the one that I wish to defend does not identify pains with brain processes. Rather it claims that pains are identical with the phenomena that are represented by the nociperceptual system. Thus, according to the account I wish to defend, pains are in most cases peripheral phenomena, phenomena that can occur independently of the brain and even of the entire central nervous system.

My argument begins with propositions a. and b., which I take to be components of folk psychology:

a. Pains have various functional properties. Specifically, they have the power to cause such psychological phenomena as distress and aversion to stimuli that appear to produce them, and they have the power to cause such behavioral phenomena as withdrawal, wincing, crying out, and nursing of injured parts of the body.
b. Pains have bodily locations and other spatial properties.

I expect it will be readily acknowledged that a. is correct, but b. is more controversial, so I will say a few things in its defense.

I begin by noting that b. has considerable intuitive appeal. References to locations and other spatial properties pervade our ordinary thought and talk about pain. Thus, for example, you might tell a doctor that there is a burning pain in your left forearm. Elaborating, you might say that the pain has different areas of intensity in different parts of the forearm, with the greatest intensities occurring in the neighborhood of your wrist. The doctor would find your comments perfectly natural, and would indeed rely heavily on them in seeking the cause of your pain.

We must ask, however, whether it is appropriate to take such ascriptions of spatial properties literally. After all, it might be urged that there is an alternative to the literalistic view. Instead of taking claims of the form "I have a pain in bodily location L" at face value, it might be urged, we can construe them as in some sense equivalent to claims that attribute *representational* properties to pains. In particular, we can construe them as in some sense equivalent to claims of the form "I have a pain that represents a disturbance involving actual or potential damage in bodily location L." A review of the literature shows that there are a number of philosophers who favor views of this kind.

I am not one of them, however. As I see it, all such views face an insurmountable problem.

Sentences of the form "I have a pain in bodily location L" are intended to articulate what one is aware of when one is aware of a pain. Now when one is introspectively aware of a pain, one is aware of it as being in a certain bodily location. Pains are experienced as being out there in the body, at some distance from the center of awareness. But if sentences of the form "I have a pain in bodily location L" are used to articulate what one is aware of in being aware of a pain, and we are aware of pains as having locations, then sentences of the sort in question must be construed as ascribing locations to pains. If one interprets them as ascribing representational properties to pains, then one has failed to understand one or the other (or both) of two fundamental facts – either a fact concerning the role that sentences of the form "I have a pain in bodily location L" play in the language, or a fact concerning awareness of pain.

Part of what I am claiming here is that awareness of the locations of pains is, phenomenologically, on a par with awareness of the intensities of pains. When one is aware of the intensity of a pain, one is not aware of the pain as representing a certain condition. Rather, one is aware of it as having a certain non-representational property. The same is true, I suggest, of awareness of the locations of pains. To be aware of the location of a pain is not to be aware of the pain as having a representational property, but to be aware of it as being located, as being out there in the body, as being situated in space.

As I see it, then, attempts to explain statements of the form "I have a pain in bodily location L" in terms of representational properties are incompatible with the phenomenology of awareness of pain.

I have been defending the idea that statements of the form "I have a pain in bodily location L" should be taken literally, or in other words, the idea that their truth conditions are similar to the truth conditions of statements that ascribe locations in physical space to extra-mental particulars. In order to complete the defense, I need to take account of another opposing view. The view I have in mind here asserts that statements of the given form are implicitly concerned with correlations of a certain kind – specifically, with correlations between qualia and heightened activity in various regions of the body. In its full dress form, the view maintains that statements of the form "I have a pain in bodily location L" have truth conditions that are captured by claims of the form "There is a quality Q such that (i) I have a pain with quality Q, and (ii) Q is strongly correlated with heightened activity in location L of my body." I propose to call this the *correlational theory* of bodily locations.

The correlational theory faces several problems. First, it conflicts with the phenomenology of locatedness: contrary to what the theory implies, we experience pains as located, and not merely as correlated with things that are located. Second, the correlational theory fails to honor our intuition that the locations of our pains are unique. Suppose that I now have a pain in my left knee. The correlational theory claims that there is a qualitative property that is strongly correlated, in my internal economy, with heightened activity in my left knee, and it explains the fact that my current pain is located in my left knee in terms of the fact that the pain is an instance of the given property. Let us suppose that the first part of this story is correct – that there is in fact a qualitative property, φ, that is correlated with heightened activity in my left knee. Now reflection shows that this one correlation involving φ must be accompanied by others. Thus, when there is heightened activity in my left knee, there is also heightened activity in various other regions of my body – in certain parts of my spinal cord, in certain parts of my thalamus, and so on. Given that φ is correlated with heightened activity in my left knee, it must also be correlated with heightened activity in these other areas. But this means that the correlational theory is committed to claiming that my current pain is located in the other areas as well as in my left knee. For example, it is committed to claiming that the pain is located in my thalamus. Claims of this sort are clearly wrong. The pain is located in my left knee, period. (It might seem that we could rule out the unwanted additional locations by saying that an area cannot serve as the location of a pain unless there is damage in that location. In fact, however, if we suppose that there is a reference to damage in the content of the predicates we use to ascribe locations to pains, we will be committed to saying that the predicates in question have contents that are quite different than those of the predicates we use in ascribing bodily locations to sensations of other kinds, such as itches and sensations of heat. That is, we would be committed to saying that predicates of the form "is in bodily location L" are ambiguous. This view is clearly wrong. If I say that I have a pain in my left knee, and then, a bit later on, say that I now have an itch in my left knee, I am ascribing the same property to the itch as I ascribed to the pain.)

I will state one more objection. Observe that if the correlational theory is correct, then saying that my pain is in my knee is closely related to saying that the present paper, "Ouch!," is "in" a certain location L on the hard drive of my computer. After all, we say that "Ouch!" is "in" L because the paper appears on the screen of my monitor when and only when there is activity of an appropriate sort in L. It is pretty clear that something has gone wrong here. The sense of "in" that we use in attributing computer locations to philosophy

papers appears to be far removed from the sense of "in" that we normally use in talking about the locations of pains. To mention only one difference, it is clear that the second "in" is compatible with multiple locations.

In view of these considerations, it seems appropriate to conclude that like a., b. is a fundamental truth about the nature of pain:

a. Pains have various functional properties. Specifically, they have the power to cause such psychological phenomena as distress and aversion to stimuli that appear to produce them, and they have the power to cause such behavioral phenomena as withdrawal, wincing, crying out, and nursing of injured parts of the body.
b. Pains have bodily locations and other spatial properties.

At all events, I will assume that this is so in the sequel.[5]

Now in addition to being possessed by pains, the properties cited in a. and b. are possessed by certain bodily phenomena that I will call *disturbances*, meaning thereby to highlight the fact that the phenomena involve actual or potential bodily damage. Disturbances are the phenomena that are represented by nociperceptual representations, or in other words, the phenomena that give rise to signals in the damage-detecting C fibers and A-δ fibers. Given this definition, it is not hard to see that disturbances are capable of causing psychological and behavioral phenomena that parallel the effects of pain, and that disturbances therefore have the functional properties that are cited in proposition a. Further, it is clear that disturbances have bodily locations and other spatial properties, and in particular, that they have locations and other spatial properties that correspond perfectly, in most cases, to the locations and other spatial properties of pains.

Extending these observations and also refining them, I wish to claim that the following correlation thesis is true:

(CT) Where x is any pain, there is a disturbance y such that (i) y occurs in exactly the same bodily location as x, and (ii) y has exactly the same causal powers as x. Further, where y is any disturbance, there is a pain x such that (i) x occurs in exactly the same place as y, and (ii) x has exactly the same causal powers as y.

As is widely recognized, (CT) holds in an enormous number of cases.

It must be conceded, however, that (CT) appears to be called into question by a wide range of counterexamples. Thus, it appears to true that in many cases, chronic back pain is not accompanied by concurrent damage, but is rather due to a feedback loop in the spinal cord. In such cases, the loop may have been

brought into existence by an injury to the back, but other than the loop itself, the deleterious effects of the injury have disappeared. Again, it appears to be possible to produce an impression of pain by directly stimulating appropriately chosen points in the somatosensory cortex. In such cases, the experience of the stimulated subject may be indistinguishable from the experience of someone who is actually undergoing a trauma, and who is therefore unquestionably feeling real pain. Surely if an experience is indistinguishable from the experience of a subject who is feeling real pain, then the experience is itself an experience of real pain. Further, as is well known, soldiers and athletes often sustain serious injuries that appear not to be accompanied by pain, at least at the times when the injuries actually occur. Thus, the victims show none of the usual behavioral effects of pain, and they deny that there is any awareness or experience of pain.

Although it is natural to see phenomena of this sort as posing a fatal threat to the hypothesis that pains are identical with disturbances, I believe that the threat can in fact be dealt with successfully. In each of the last two sections I argued that it is possible to distinguish between its seeming to one that one is in pain and its being the case that one is in pain. If this is correct, then we have the means to explain away cases that are presented as counterexamples to (CT). Thus, for example, we can respond to the foregoing claims about chronic back pain by saying that the cases they cite are in fact only cases in which it appears to an agent that he or she is in pain, or in other words, that they are in fact only cases in which vivid nociperceptual representations give one the impression that that one is in pain. Further, we can respond to alleged cases in which bodily damage is not accompanied by pain by saying that they are cases in which, for one reason or another, damage is not accompanied by a nociperceptual representation of pain.

Before we can commit ourselves fully to this way of defending (CT), however, we must consider the question of whether there are grounds for applying the distinction between appearance and reality in cases of chronic back pain and in other problem-posing cases. That is to say, we must consider whether there is any reason to suppose that the apparent counterexamples to (CT) really are cases in which hallucinations or illusions occur. After all, science and philosophy depend for their integrity on showing a proper respect for data. We must be sure that we are not being disrespectful in the present case.

I think we can in fact be sure of this. Thus, reflection shows, I believe, that there are good reasons for thinking that the data that cause problems for (CT) are the products of hallucination or illusion. The experiences that are described as cases of chronic back pain appear to be due to pathologies of the nociperceptual

tual system, and the same is true of the experiences that are described as cases of phantom limb pain. Surely it is appropriate to think of experiences that are due to perceptual pathologies as hallucinatory. Further, the experiences that are produced by Penfieldian probing of the brain can legitimately be described as due to external interference with the normal course of nociperceptual activity. As with the products of pathologies, we are used to thinking of the products of external interference with perceptual systems as hallucinatory. Finally, while the nature of referred pain is a complex issue, it is plausible that referred pain results in many cases from design imperfections – specifically, imperfections having to do with the use of single channels to carry different kinds of information. Insofar as this is true, it seems appropriate to think of referred pains as illusory.

I hope that the reader now shares my enthusiasm for (CT). If not, perhaps the remainder of the discussion will make it seem more plausible.

Given (CT), it is easy to formulate an argument for the claim that pains are identical with disturbances. The argument I have in mind runs as follows: "When one encounters a lawful correlation, it is always appropriate to ask for an explanation of why the correlation holds; and it is no less appropriate to explain the correlation by claiming that the correlated phenomena are identical. So, in particular, it is appropriate to explain the truth of (CT) in terms of an identity hypothesis. The hypothesis that pains are identical with disturbances is graced by simplicity and elegance, and it is also characterized by considerable explanatory power. Surely this hypothesis provides the best possible explanation for the truth of (CT)! Now there is a familiar principle of inductive reasoning which counsels that the hypothesis that gives the best explanation of a set of facts is quite likely to be true. Accordingly, there is good reason to think that the pain/disturbance identity hypothesis is correct."

Before going on, we should pause to consider an objection to this line of thought. The objection acknowledges that pains and disturbances share a number of important properties, but it goes on to assert that pains also have a range of properties that disturbances lack. Thus, the objection maintains, pains have a variety of qualitative properties, such as *being very intense*, *having a throbbing feel*, and *having a burning feel*, that are not possessed by disturbances. In view of these differences, the objection continues, we are obliged to reject the hypothesis that pains are identical with disturbances.

I will not attempt here to provide a detailed answer to this objection. Rather, I will just observe that there is reason to believe that each of the qualitative characteristics of pains to which the objection refers corresponds to a physical characteristic of disturbances. Consider, for example, the intensities of

pains. It is plausible that each level of intensity is matched by a level of physical activity – an energy level, if you like – at the site of the pain. Now the need to explain the truth of the correlation thesis (CT) gives us a reason to identify intensity levels of pains with the corresponding energy levels of disturbances. Moreover, it appears that there is nothing that stands in the way of this identification. In asserting that the characteristics are identical we are making a move of a familiar and obviously acceptable sort, for our assertion is an instance of the familiar practice of identifying perceptually given phenomena, such as heat and mist, with underlying physical phenomena, such as molecular motion and suspended water droplets.

In addition to the foregoing (CT)-based argument for the pain/disturbance identity hypothesis, there is also a second argument for the hypothesis. This second argument has two main premises. The first premise is the proposition that it is the function of the nociperceptual system to detect disturbances. Since it is plausible that the high level representations of a perceptual system represent the phenomena that the system is designed to detect, it follows from this first premise that the high level representations of the nociperceptual system, the representations that consist of activity patterns in the somatosensory cortex, represent disturbances. The second premise is the proposition that states of awareness of pain supervene on high level nociperceptual representations. Now when one state supervenes on another, the properties of the former are determined by properties of the latter. Accordingly, the second premise implies that the representational contents of states of awareness of pain are determined by the representational contents of high level nociperceptual representations. But this implies in turn that the phenomena that are represented by states of awareness of pain supervene upon the phenomena that are represented by nociperceptual representations. Combining this result with the forementioned consequence of the first premise, we arrive at the conclusion that pains supervene upon disturbances. Now the best explanation of this supervenience relationship is the hypothesis that pains are identical with disturbances. Hence, this hypothesis is probably true.

4.

I have been concerned thus far to elaborate a theory of pain and awareness of pain that consists primarily of the following propositions:

1. Awareness of pain is perceptual.

2. Like all other cases of perceptual awareness, awareness of pain involves representations that are ontologically independent of the phenomena they represent. Accordingly, awareness of pain can be illusory and hallucinatory. Also, pains can exist without being objects of awareness.
3. Pains can be identified with bodily disturbances that involve actual or potential damage.

I would like now to consider this theory in relation to another one, an alternative account that shares some of its components but that also contains others that are altogether foreign to its letter and spirit. This second theory consists of 1. and 2. and also of 4.–6.:

4. Pains can be identified with certain perceptual representations – specifically, with perceptual representations of bodily disturbances that that involve actual or potential damage.
5. Hence (in view of 1., 2., and 4.), awareness of pain depends essentially on *higher order perceptual representations* – that is, on perceptual representations of perceptual representations.
6. Pains count as conscious when they are represented perceptually and as unconscious when they are not so represented. That is to say, a conscious pain is one that is the object of a higher order perceptual representation.

I do not know of any place in which this second theory has been articulated in exactly these words, but it appears to be very close in spirit to a theory that is defended by Lycan (1996), and it has substantial affinities to theories that have been put forward by Tye (1990: 113–114) and Dretske (1995: 102–103). I will call it the *higher order perception theory*.

One reason for considering the higher order perception theory is just that I wish to remind the reader that there is a substantial literature in which 1. and 2. are taken quite seriously and even embraced with enthusiasm. Since 1. and 2. have withstood scrutiny by others, the chance that they suffer from internal contradiction or other forms of incoherence is small. A second reason is that I hope to build support for the first theory by converting adherents of the higher order perception theory. The theories are sufficiently similar in spirit that it is natural to suppose that someone who abandoned the second theory would find it natural to come to embrace the first. I will try to show that there is a strong reason for abandoning the higher order perception theory.

The first theory claims that awareness of pain involves a certain perceptual faculty, P. It describes P as focused on certain disturbances in the body, and it identifies pains with these disturbances. The higher order perception

theory agrees that there is perceptual awareness of the disturbances in question. More specifically, it joins the first theory in recognizing P. Unlike the first theory, however, it identifies pains with the perceptual representations that are associated with P. It also departs from the first theory in postulating a second perceptual faculty P*. It describes P* as focused on the representations associated with P, and it maintains that awareness of pain occurs when P* produces a representation that represents one of the former representations. In short, it claims that awareness of pain occurs when a P-representation is represented by a P*-representation.

As this characterization makes clear, the existential claims of the first theory are properly included in the existential claims of the higher order perception theory: while the first theory is content to postulate P, the higher order perception theory postulates both P and P*. It follows that there is a substantial prima facie reason to prefer the first theory. To be sure, this is not by itself a decisive consideration, for reasons having to do with ontological simplicity can be trumped by considerations of other kinds. I will now argue, however, that there are no simplicity-neutralizing considerations that favor the higher order perception theory. There are no a priori reasons that strongly favor the higher order perception theory, and there are no empirical reasons, either.

It might seem at first that there are more or less a priori intuitions which show that the higher order perception theory is correct. Thus, for example, there is no doubt that we have intuitions to the effect that phantom limb pains are cases of *real* pain, not cases of hallucinatory pain. If taken at face value, these intuitions show that it is possible to be in pain even though one's body is free from disturbances involving actual or potential damage. For this reason, it might be thought that they provide decisive a priori support for the higher order perception theory. Thus, unlike the first theory, the higher order perception theory allows for the possibility of real pains that are not accompanied by bodily disturbances.

As I argued in section II, however, there are grounds for thinking that it is possible to draw an appearance/reality distinction with respect to pains. Assuming that this is correct, it is not necessary to take our intuitions about phantom limb pains at face value. Such pains can appropriately be dismissed as hallucinatory. Psychosomatic pains, referred pains, and pains produced by artificial stimulation of the brain can be explained away in similar fashion.

It seems reasonable to hope that these observations will have an ecumenical appeal, for the arguments offered in section II are relatively free from partisan assumptions. The observations should have a particular appeal, however, to philosophers who embrace the higher order perception theory. After all, these

philosophers are independently committed to the claim that it is possible to draw an appearance/reality distinction with respect to pains. Since the other observations follow trivially from this central claim, advocates of the higher order perception theory should find it easy to accept the entire package.

When we turn to consider the question of whether there is empirical evidence that favors the higher order perception theory, we find, I think, that the answer is "no." Surely there is no introspective evidence of this sort. When we examine awareness of pain introspectively, we find that such awareness resembles visual awareness of shapes and colors: it provides us with a sensuous grasp of phenomena that are located in physical space, and at some remove from the center of awareness. Introspection thus supports the idea that awareness of pain is perception of a bodily disturbance. Moreover, as far as I can tell, it supports this idea exclusively. In particular, it provides no support at all for the idea that two perceptual faculties are involved in awareness of pain, nor for the related idea that awareness of pain involves a sensuous grasp of a perceptual representation of a bodily disturbance.

Does science provide empirical support for the higher order perception theory? Well, as I have observed in earlier sections, neuroscience provides considerable support for the view that there is a perceptual faculty P whose function it is to monitor disturbances involving actual or potential damage, and to produce high-level representations of disturbances of this sort that pass a certain threshold in intensity. Thus, there is neuroscientific support for one of the assumptions of the higher order perception theory. This is, however, an assumption that it shares with the first theory. Is there any scientific support for the ontological assumption that distinguishes the higher order perception theory from the first – that is, for the assumption that there is a second perceptual faculty that monitors and registers the perceptual representations that are produced by P? At present, I believe, the answer is "no." To be sure, nociperceptual representations are registered elsewhere in the brain; but I know of no scientific grounds for thinking that the mechanisms that register these representations support distinctively perceptual faculties.

To amplify: Advocates of the higher order perception theory sometimes try to support their position by pointing out that there are monitors or scanning mechanisms in the brain that register the representations associated with P. Reflection shows, however, that arguments of this sort fall far short of their purpose. Thus, as we have seen in earlier sections, genuinely perceptual faculties have structural and functional features that distinguish them from representational faculties of other kinds. It follows that if a mechanism in the brain is to realize a perceptual faculty, it must have an appropriate structural and

functional profile. As far as I know, there is at present no evidence that there is a brain mechanism that both has an appropriate profile and is dedicated to registering nociperceptual representations.

It appears, then, that there are no considerations that mandate acceptance of the higher order perception theory. But this means that the considerations of simplicity that favor the first theory can reasonably be taken as decisive. The first theory should be preferred to the higher order perception theory.

I conclude with a remark about the claim that pains are conscious. Advocates of the higher order perception theory endorse this claim. They accept it as true and propose to explain the grounds of its truth. As we saw, their explanation comes roughly to this: a pain counts as conscious just in case it is an object of perceptual awareness.

My view of the matter is quite different. Thus, as I see it, it is at best highly misleading to say that a pain is conscious. On my view, pains are bodily phenomena – that is, physical entities out there in three dimensional space. Further, awareness of them is perceptual. If these views are right, then it is no more appropriate to apply "conscious" to pains than it is to apply "conscious" to such objects of visual awareness as cars, tables, and blades of grass. It is only mental events that can count as conscious, and the views that I have endorsed in this paper imply that pains are not mental.

To be sure, it is perfectly correct to say that we are conscious *of* pains, just as it is correct to say that we are conscious *of* objects of visual awareness. But this has no tendency to show that we can legitimately apply the adjective "conscious" to pains. No one would say that the fact that we are conscious of cars makes it true to say that cars are conscious. Equally, as I see it, no one should say that the fact that we are conscious of pains makes it true that pains are conscious.[6]

Notes

* I have benefited considerably from conversations with Jane Dyer, discussions with the students in a seminar at Brown University in the fall of 2002, questions following a presentation at a conference at the University of Magdeburg, questions following a talk at Cornell, and Sydney Shoemaker's comments on the penultimate draft.

1. One might try to avoid this objection by claiming that the content of a state of awareness of pain is due to a non-demonstrative indexical like *the internal structure of the pain that I am having now*. But it would not be rational to entertain such an indexical unless one was experientially aware of a pain when one entertained it.

2. For an illuminating review of the relevant facts, see Price (1999).

3. Ramachandran's explanation of the Capgras delusion is engagingly presented in a Nova video entitled *Secrets of the Mind* (Boston: WGBH Educational Foundation Video 2001).

4. D. A. Smith arrives at the same conclusion by a somewhat different route. See Smith (2002: 140–146).

5. There is a well known linguistic observation, due to Ned Block, which is sometimes taken to show that the meaning of the word "in" does not have a genuine spatial significance in claims of the form "I have a pain in bodily location L." (See Block (1983: 517).) The argument begins with the observation that when "in" is used in the sense of spatial containment, it is a transitive relation: necessarily, if my foot is in my sock and my sock is in my shoe, then my foot is in my shoe. On the other hand, when "in" is used in the sense it has in claims about sensations, transitivity seems to fail. Thus, it would be very odd, and probably mistaken, to reason as follows: there is a pain in my finger and my finger is in my pocket, so there is a pain in my pocket. Because of this contrast, the argument continues, the sense of "in" that is germane to claims about sensations is quite different from the sense it has in claims that in claims that assign spatial locations to objects. But then we must conclude that claims of the form "There is a pain in bodily location L" do not assign spatial locations to pains. (I have here modified Block's example so as to prescind from an aspect that is irrelevant to my purpose. Also, I do not mean to suggest that Block endorses this argument. The conclusion that he draws from his example is that "'in' as applied in locating pains differs in meaning systematically from the standard spatial enclosure sense." (p. 517) It does not follow from this that the "in" in question lacks spatial significance altogether. The argument I have cited is one that is suggested by Block's example, not one that Block has given himself.)

Block's example shows that the "in" that figures in claims about sensations is not the "in" of containment, but the foregoing argument goes wrong in inferring that statements involving the former "in" do not assign spatial locations to sensations. To see this, observe that it would be quite odd, and probably wrong, to reason as follows: there is arthritis in my finger, and my finger is in my pocket, so there is arthritis in my pocket. It would clearly be a mistake to infer from this observation that the claim that there is arthritis in my finger does not assign a spatial location to my arthritis. (The pain example has a different feel than the arthritis example, but this is because we are somewhat more inclined to think of pains as "objects," and therefore as being the kinds of things that can be contained in other things, than we are to think of cases of arthritis as objects. (Cases of arthritis are "conditions.") This does not affect the point at issue.)

I am grateful to Block and to Brian Weatherson for discussion concerning these matters.

6. Advocates of the higher order perception view are by no means the only theorists who endorse the claim that pains are conscious. The claim is also embraced by David Rosenthal and other friends of the "higher order thought" theory of consciousness (see, e.g., Rosenthal 1997), and by Ned Block and other friends of the view that there is a certain form of consciousness, usually called "phenomenal consciousness," that is possessed by all events with phenomenal character (see, e.g., Block 1995). If the position recommended above is correct, these other views are no less in need of correction than the higher order perception theory.

References

Block, N. (1983). Mental Pictures and Cognitive Science. *The Philosophical Review, XCII*, 499–541.

Block, N. (1995). On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences, 18*, 227–247.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Lycan, W. G. (1996). *Consciousness and Experience.* Cambridge, MA: MIT Press.

Price, D. D. (1999). *Psychological Mechanisms of Pain and Analgesia*. Seattle: ISAP Press.

Rosenthal, D. (1997). "A Theory of Consciousness." In N. Block, O. Flanagan, & G. Guzeldere (Eds.), *The Nature of Consciousness*. Cambridge, MA: MIT Press.

Smith, D. A. (2002). *The Problem of Perception.* Cambridge, MA: Harvard University Press.

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Yeshurun, Yaffa & Marissa Carrasco (1998). Attention Improves or Impairs Visual Performance by Enhancing Spatial Resolution. *Nature, 365*, 72–75.

Index of names

Index of topics

In the series ADVANCES IN CONSCIOUSNESS RESEARCH (AiCR) the following titles have been published thus far or are scheduled for publication:

1. GLOBUS, Gordon G.: *The Postmodern Brain.* 1995.
2. ELLIS, Ralph D.: *Questioning Consciousness. The interplay of imagery, cognition, and emotion in the human brain.* 1995.
3. JIBU, Mari and Kunio YASUE: *Quantum Brain Dynamics and Consciousness. An introduction.* 1995.
4. HARDCASTLE, Valerie Gray: *Locating Consciousness.* 1995.
5. STUBENBERG, Leopold: *Consciousness and Qualia.* 1998.
6. GENNARO, Rocco J.: *Consciousness and Self-Consciousness. A defense of the higher-order thought theory of consciousness.* 1996.
7. MAC CORMAC, Earl and Maxim I. STAMENOV (eds): *Fractals of Brain, Fractals of Mind. In search of a symmetry bond.* 1996.
8. GROSSENBACHER, Peter G. (ed.): *Finding Consciousness in the Brain. A neurocognitive approach.* 2001.
9. Ó NUALLÁIN, Seán, Paul MC KEVITT and Eoghan MAC AOGÁIN (eds): *Two Sciences of Mind. Readings in cognitive science and consciousness.* 1997.
10. NEWTON, Natika: *Foundations of Understanding.* 1996.
11. PYLKKÖ, Pauli: *The Aconceptual Mind. Heideggerian themes in holistic naturalism.* 1998.
12. STAMENOV, Maxim I. (ed.): *Language Structure, Discourse and the Access to Consciousness.* 1997.
13. VELMANS, Max (ed.): *Investigating Phenomenal Consciousness. Methodologies and Maps.* 2000.
14. SHEETS-JOHNSTONE, Maxine: *The Primacy of Movement.* 1999.
15. CHALLIS, Bradford H. and Boris M. VELICHKOVSKY (eds.): *Stratification in Cognition and Consciousness.* 1999.
16. ELLIS, Ralph D. and Natika NEWTON (eds.): *The Caldron of Consciousness. Motivation, affect and self-organization – An anthology.* 2000.
17. HUTTO, Daniel D.: *The Presence of Mind.* 1999.
18. PALMER, Gary B. and Debra J. OCCHI (eds.): *Languages of Sentiment. Cultural constructions of emotional substrates.* 1999.
19. DAUTENHAHN, Kerstin (ed.): *Human Cognition and Social Agent Technology.* 2000.
20. KUNZENDORF, Robert G. and Benjamin WALLACE (eds.): *Individual Differences in Conscious Experience.* 2000.
21. HUTTO, Daniel D.: *Beyond Physicalism.* 2000.
22. ROSSETTI, Yves and Antti REVONSUO (eds.): *Beyond Dissociation. Interaction between dissociated implicit and explicit processing.* 2000.
23. ZAHAVI, Dan (ed.): *Exploring the Self. Philosophical and psychopathological perspectives on self-experience.* 2000.
24. ROVEE-COLLIER, Carolyn, Harlene HAYNE and Michael COLOMBO: *The Development of Implicit and Explicit Memory.* 2000.
25. BACHMANN, Talis: *Microgenetic Approach to the Conscious Mind.* 2000.
26. Ó NUALLÁIN, Seán (ed.): *Spatial Cognition. Selected papers from Mind III, Annual Conference of the Cognitive Science Society of Ireland, 1998.* 2000.
27. McMILLAN, John and Grant R. GILLETT: *Consciousness and Intentionality.* 2001.

28. ZACHAR, Peter: *Psychological Concepts and Biological Psychiatry. A philosophical analysis.* 2000.
29. VAN LOOCKE, Philip (ed.): *The Physical Nature of Consciousness.* 2001.
30. BROOK, Andrew and Richard C. DeVIDI (eds.): *Self-reference and Self-awareness.* 2001.
31. RAKOVER, Sam S. and Baruch CAHLON: *Face Recognition. Cognitive and computational processes.* 2001.
32. VITIELLO, Giuseppe: *My Double Unveiled. The dissipative quantum model of the brain.* 2001.
33. YASUE, Kunio, Mari JIBU and Tarcisio DELLA SENTA (eds.): *No Matter, Never Mind. Proceedings of Toward a Science of Consciousness: Fundamental Approaches, Tokyo, 1999.* 2002.
34. FETZER, James H.(ed.): *Consciousness Evolving.* 2002.
35. Mc KEVITT, Paul, Seán Ó NUALLÁIN and Conn MULVIHILL (eds.): *Language, Vision, and Music. Selected papers from the 8th International Workshop on the Cognitive Science of Natural Language Processing, Galway, 1999.* 2002.
36. PERRY, Elaine, Heather ASHTON and Allan YOUNG (eds.): *Neurochemistry of Consciousness. Neurotransmitters in mind.* 2002.
37. PYLKKÄNEN, Paavo and Tere VADÉN (eds.): *Dimensions of Conscious Experience.* 2001.
38. SALZARULO, Piero and Gianluca FICCA (eds.): *Awakening and Sleep-Wake Cycle Across Development.* 2002.
39. BARTSCH, Renate: *Consciousness Emerging. The dynamics of perception, imagination, action, memory, thought, and language.* 2002.
40. MANDLER, George: *Consciousness Recovered. Psychological functions and origins of conscious thought.* 2002.
41. ALBERTAZZI, Liliana (ed.): *Unfolding Perceptual Continua.* 2002.
42. STAMENOV, Maxim I. and Vittorio GALLESE (eds.): *Mirror Neurons and the Evolution of Brain and Language.* 2002.
43. DEPRAZ, Natalie, Francisco VARELA and Pierre VERMERSCH.: *On Becoming Aware. A pragmatics of experiencing.* 2003.
44. MOORE, Simon and Mike OAKSFORD (eds.): *Emotional Cognition. From brain to behaviour.* 2002.
45. DOKIC, Jerome and Joelle PROUST: *Simulation and Knowledge of Action.* 2002.
46. MATHEAS, Michael and Phoebe SENGERS (ed.): *Narrative Intelligence.* 2003.
47. COOK, Norman D.: *Tone of Voice and Mind. The connections between intonation, emotion, cognition and consciousness.* 2002.
48. JIMÉNEZ, Luis: *Attention and Implicit Learning.* 2003.
49. OSAKA, Naoyuki (ed.): *Neural Basis of Consciousness.* 2003.
50. GLOBUS, Gordon G.: *Quantum Closures and Disclosures. Thinking-together post-phenomenology and quantum brain dynamics.* 2003.
51. DROEGE, Paula: *Caging the Beast. A theory of sensory consciousness.* 2003.
52. NORTHOFF, Georg: *Philosophy of the Brain. The 'Brain problem'.* 2004.
53. HATWELL, Yvette, Arlette STRERI and Edouard GENTAZ (eds.): *Touching for Knowing. Cognitive psychology of haptic manual perception.* 2003.

54. BEAUREGARD, Mario (ed.): *Consciousness, Emotional Self-Regulation and the Brain.* 2004.
55. PERUZZI, Alberto (ed.): *Mind and Causality.* 2004.
56. GENNARO, Rocco J. (ed.): *Higher-Order Theories of Consciousness. An Anthology.* 2004.
57. WILDGEN, Wolfgang: *The Evolution of Human Language. Scenarios, principles, and cultural dynamics.* n.y.p.
58. GLOBUS, Gordon G., Karl H. PRIBRAM and Giuseppe VITIELLO (eds.): *Brain and Being. At the boundary between science, philosophy, language and arts.* n.y.p.