

PDF version of the entry
Higher-Order Theories of Consciousness
<https://plato.stanford.edu/archives/fall2020/entries/consciousness-higher/>
from the FALL 2020 EDITION of the

STANFORD ENCYCLOPEDIA OF PHILOSOPHY



Edward N. Zalta Uri Nodelman Colin Allen R. Lanier Anderson
Principal Editor Senior Editor Associate Editor Faculty Sponsor

Editorial Board
<https://plato.stanford.edu/board.html>

Library of Congress Catalog Data
ISSN: 1095-5054

Notice: This PDF version was distributed by request to members of the Friends of the SEP Society and by courtesy to SEP content contributors. It is solely for their fair use. Unauthorized distribution is prohibited. To learn how to join the Friends of the SEP Society and obtain authorized PDF versions of SEP entries, please visit <https://leibniz.stanford.edu/friends/>.

Stanford Encyclopedia of Philosophy
Copyright © 2020 by the publisher
The Metaphysics Research Lab
Center for the Study of Language and Information
Stanford University, Stanford, CA 94305

Higher-Order Theories of Consciousness
Copyright © 2020 by the authors
Peter Carruthers and Rocco Gennaro

All rights reserved.

Copyright policy: <https://leibniz.stanford.edu/friends/info/copyright/>

Higher-Order Theories of Consciousness

First published Tue Apr 3, 2001; substantive revision Wed Sep 2, 2020

Higher-order theories of consciousness try to explain the difference between unconscious and conscious mental states in terms of a relation obtaining between the conscious state in question and a higher-order representation of some sort (either a higher-order perception of that state, or a higher-order thought about it). The most challenging properties to explain are those involved in *phenomenal* consciousness—the sort of state that has a *subjective* dimension, that has ‘feel’, or that it is *like something* to undergo. These properties will form the focus of this article.

- 1. Kinds of Consciousness
- 2. The Motivation for a Higher-Order Approach
- 3. Higher-Order Perception Theory
- 4. Higher-Order Thought Theory (1): Actualist
- 5. Higher-Order Thought Theory (2): Dispositionalist
- 6. Self-Representational Higher-Order Theories
- 7. Objections to a Higher-Order Approach
 - 7.1 Local objections
 - 7.2 Generic objections
- Bibliography
- Academic Tools
- Other Internet Resources
- Related Entries

1. Kinds of Consciousness

One of the advances made in the last few decades has been to distinguish between different questions concerning consciousness (see particularly: Rosenthal 1986; Dretske 1993; Block 1995; Lycan 1996). Not everyone agrees on quite *which* distinctions need to be drawn. But all are agreed that we should distinguish *creature* consciousness from *mental-state* consciousness. It is one thing to say *of an individual person or organism* that it is conscious (either in general or of something in particular); and it is quite another thing to say *of one of the mental states* of a creature that it is conscious.

It is also agreed that within creature-consciousness itself we should distinguish between *intransitive* and *transitive* variants. To say of an organism that it is conscious *simpliciter* (intransitive) is to say just that it is awake, as opposed to asleep or comatose. There don't appear to be any deep philosophical difficulties lurking here (or at least, they aren't difficulties specific to the topic of consciousness, as opposed to mentality in general). But to say of an organism that it is conscious *of such-and-such* (transitive) is normally to say at least that it is *perceiving* such-and-such, or *aware of* such-and-such. So we say of the mouse that it is conscious of the cat outside its hole, in explaining why it doesn't come out; meaning that it *perceives* the cat's presence. To provide an account of transitive creature-consciousness would thus be to attempt a theory of perception.

There is a choice to be made concerning transitive creature-consciousness, failure to notice which may be a potential source of confusion. For we have to decide whether the perceptual state in virtue of which an organism may be said to be transitively-conscious of something must itself be a conscious one (state-conscious—see below). If we say 'Yes' then we shall need to know more about the mouse than merely that it perceives the cat if we are to be assured that it is conscious of the cat—we shall need to

establish that its percept of the cat is itself conscious. If we say 'No', on the other hand, then the mouse's perception of the cat will be sufficient for the mouse to count as conscious of the cat; but we may have to say that although it is conscious of the cat, the mental state in virtue of which it is so conscious is not itself a conscious one! It may be best to by-pass any danger of confusion here by avoiding the language of transitive-creature-consciousness altogether. Nothing of importance would be lost to us by doing this.

Turning now to the notion of *mental-state consciousness*, the major distinction here is between *phenomenal consciousness*, on the one hand—which is a property of states that it is *like something* to be in, that have a distinctive 'feel' (Nagel 1974)—and various functionally-definable forms of *access consciousness*, on the other (Block 1995). Most theorists believe that there are mental states—such as occurrent thoughts or judgments—that are access-conscious (in whatever is the correct functionally-definable sense), but that are not phenomenally conscious. In contrast, there is considerable dispute as to whether mental states can be phenomenally-conscious without also being conscious in the functionally-definable sense—and even more dispute about whether phenomenal consciousness can be *reductively explained* in functional and/or representational terms.

It is plain that there is nothing deeply problematic about functionally-definable notions of mental-state consciousness, from a naturalistic perspective. For mental functions and mental representations are the staple fare of naturalistic accounts of the mind. But this leaves plenty of room for dispute about the form that the correct functional account should take. Some claim that for a state to be conscious in the relevant sense is for it to be poised to have an impact on the organism's decision-making processes (Kirk 1994; Dretske 1995; Tye 1995, 2000), perhaps also with the additional requirement that those processes should be distinctively *rational* ones (Block 1995). Others think that the relevant requirement for

access-consciousness is that the state should be suitably related to higher-order representations—experiences and/or thoughts—of that very state (Armstrong 1968, 1984; Rosenthal 1986, 1993, 2005; Dennett 1978a, 1991; Carruthers 1996, 2000, 2005; Lycan 1987, 1996; Gennaro 2012).

What *is* often thought to be naturalistically problematic, in contrast, is phenomenal consciousness (Nagel 1974, 1984; Jackson 1982, 1986; McGinn 1991; Block 1995; Chalmers 1996). And what is really and deeply controversial is whether phenomenal consciousness can be *explained* in terms of some or other functionally-definable notion. *Cognitive* (or *representational*) theories maintain that it can. *Higher-order* cognitive theories maintain that phenomenal consciousness can be reductively explained in terms of representations (either experiences or thoughts) that are higher-order. It is such theories that concern us here.

2. The Motivation for a Higher-Order Approach

Higher-order theories, like cognitive/representational theories in general, assume that the right *level* at which to seek an explanation of phenomenal consciousness is a cognitive one, providing an explanation in terms of some combination of *causal role* and *intentional content*. All such theories claim that phenomenal consciousness consists in a certain kind of intentional or representational content (*analog* or ‘fine-grained’ in comparison with any concepts we possess) figuring in a certain distinctive position in the causal architecture of the mind. They must therefore maintain that these latter sorts of mental property don’t already implicate or presuppose phenomenal consciousness. In fact, all cognitive accounts are united in rejecting the thesis that the very properties of *mind* or *mentality* already presuppose phenomenal consciousness, as proposed by Searle (1992, 1997) for example. The higher-order approach does not attempt to reduce consciousness directly to neurophysiology but rather its

reduction is in mentalistic terms, that is, by using such notions as thoughts and awareness.

The major divide amongst representational theories of phenomenal consciousness in general, is between accounts that are provided in purely first-order terms and those that implicate higher-order representations of one sort or another (see below). These higher-order theorists will allow that first-order accounts—of the sort defended by Dretske (1995) and Tye (1995), for example—can already make some progress with the problem of consciousness. According to first-order views, phenomenal consciousness consists in analog or fine-grained contents that are available to the first-order processes that guide thought and action. So a phenomenally-conscious percept of red, for example, consists in a state with the analog content *red* which is tokened in such a way as to feed into thoughts about red, or into actions that are in one way or another guided by redness. Now, the point to note in favor of such an account is that it can explain the natural temptation to think that phenomenal consciousness is in some sense *ineffable*, or *indescribable*. This will be because such states have fine-grained contents that can slip through the mesh of any conceptual net. We can always distinguish many more shades of red than we have concepts for, or could describe in language (other than indexically—e.g., ‘*That shade*’).

The main motivation behind higher-order theories of consciousness, in contrast, derives from the belief that all (or at least most) mental-state types admit of both conscious and unconscious varieties. Almost everyone now accepts, for example, (post-Freud) that beliefs and desires can be activated unconsciously. (Think, here, of the way in which problems can apparently become resolved during sleep, or while one’s attention is directed to other tasks. Notice, too, that appeals to unconscious intentional states are now routine in cognitive science.) And then if we ask what makes the difference between a conscious and an unconscious mental

state, one natural answer is that conscious states are states that we are *aware of*. And if *awareness* is thought to be a form of creature-consciousness (see section 1 above), then this will translate into the view that conscious states are *states of which the subject is aware*, or states of which the subject is creature-conscious. That is to say, these are states that are the objects of some sort of higher-order representation—whether a higher-order perception or experience, or a higher-order thought. This is similar to the widely referenced Transitivity Principle (TP) which says that a conscious state is a state whose subject is, in some way, aware of being in it. On the other hand, a mental state of which a subject is completely *unaware* is clearly an unconscious state. (See also Lycan's (2001b) 'related simple argument' for a higher-order representation account of consciousness.)

One crucial question, then, is whether perceptual states as well as intentional states admit of both conscious and unconscious varieties. Can there be, for example, such a thing as an unconscious visual perceptual state? Higher-order theorists are united in thinking that there can. 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 having been day-dreaming or engaged in intense conversation with a passenger. We were apparently not consciously aware of much 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 (Carruthers 1989, 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. whether 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 1986, 1997, for details and discussion.)

A powerful case for the existence of unconscious visual experience has also been generated by the *two-systems theory* of vision proposed and defended by Milner and Goodale (1995; see also Jacob and Jeannerod 2003; Glover 2004). 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 and object recognition. And only the perceptions generated by the temporal-lobe system are phenomenally conscious, on their account. (Note that this isn't the 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.)

To get the flavor of Milner and Goodale's hypothesis, consider just one strand from the wealth of evidence that they provide. 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 agnostic, she isn't 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, say); 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 when 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 these data without supposing that the sensorimotor perceptual system is functionally and anatomically distinct from the object-recognition/conscious system.

But what implications does this have for phenomenal consciousness? Must these unconscious 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

question, 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 theorists are united in thinking that the relevant difference consists in the presence of something *higher-order* in the first case that is absent in the second. The same would go for the difference between unconscious and conscious desires, emotions, pains, and so on. The core intuition, again, is that a phenomenally conscious state will be a state *of which the subject is aware*.

What options does a first-order theorist have to resist this conclusion? One is to deny that the data are as problematic as they appear (as does Dretske 1995). It can be said that the unconscious states in question lack the kind of fineness of grain and richness of content necessary to count as genuinely *perceptual* states. On this view, the contrast discussed above isn't really a difference between conscious and unconscious perceptions, but rather between conscious perceptions, on the one hand, and unconscious belief-like states, on the other. Another option is to accept the distinction between conscious and unconscious perceptions, and then to explain that distinction in first-order terms. It might be said, for example, that conscious perceptions are those that are available to *belief* and *thought*, whereas unconscious ones are those that are available to guide *movement* (Kirk 1994). A final option is to bite the bullet, and insist that blindsight and sensorimotor perceptual states are indeed phenomenally conscious while not being *access-conscious*. (See Block 1995; Tye 1995; and Nelkin 1996; all of whom defend versions of this view.) On this account, blindsight percepts are phenomenally conscious states to which the subjects of those states are blind. Higher-order theorists will argue, of course, that none of these alternatives is acceptable (see, e.g., Carruthers 2000; Rosenthal 2005).

Further, Lau and Rosenthal (2011) survey the empirical evidence pertaining to the difference between higher-order theories and first-order

ones. While much is equivocal, and many questions are left unanswered, they point to a pair of studies that support a higher-order account. One is Lau and Passingham (2006), who are able to demonstrate using carefully controlled stimuli that there are circumstances in which people's subjective reports of visual experience are impaired while their first-order discrimination abilities remain fully intact. They also find that visual consciousness in these conditions is specifically associated with activity in a region of dorsolateral prefrontal cortex. Then in a follow-up study Rounis et al. (2010) find that transcranial magnetic stimulation directed at this region of cortex, thereby disrupting its activity, also has a significant impact on people's meta-visual awareness, but again without impairing first-order task performance. The degree to which the prefrontal cortex is *required* for having a conscious state and the view that the prefrontal cortex is the likely site of *all* higher-order thoughts are also the subject of vigorous continuing debate (Block 1995; Gennaro 2012, chapter nine; Kozuch 2014; Odegaard, Knight, and Lau 2017).

Most generally, then, higher-order theories of phenomenal consciousness claim the following:

Higher Order Theory (In General):

A phenomenally conscious mental state is a mental state (of a certain sort—see below) that either is, or is disposed to be, the object of a higher-order representation of a certain sort (see below).

Higher-order theorists do agree that one must normally become aware of the lower-order state *non-inferentially* since mental states can sometimes become targets of higher-order representation via conscious inference without being phenomenally conscious. For example, if I become aware of my unconscious desire to kill my boss because I have consciously inferred it from a session with my psychiatrist, then the characteristic phenomenal feel of such a conscious desire may be absent.

Still, there are then two main dimensions along which higher-order theorists disagree amongst themselves. One concerns whether the higher-order states in question are perception-like, on the one hand, or thought-like, on the other. A thought is composed of or constituted by concepts. Those taking the former option are higher-order *perception* (often called 'inner-sense') theorists, and those taking the latter option are higher-order *thought* theorists. The two theories are therefore often abbreviated as HOP (higher-order perception) and HOT (higher-order thought) theory. The other general disagreement is internal to higher-order thought approaches, and concerns whether the relevant relation between the first-order state and the higher-order thought is one of *availability* or not. That is, the question is whether a state is conscious by virtue of being *disposed* to give rise to a higher-order thought, or rather by virtue of being the *actual target* of such a thought. These are the three main options that will now concern us. (A fourth will be considered in section 6.)

3. Higher-Order Perception Theory

According to this view, humans not only have first-order non-conceptual and/or analog perceptions of states of their environments and bodies, they also have second-order non-conceptual and/or analog perceptions of their first-order states of perception. And the most popular version of higher-order perception (HOP) theory holds, in addition, that humans (and perhaps other animals) not only have sense-organs that scan the environment/body to produce fine-grained representations, but they also have *inner* senses which scan the first-order senses (i.e. perceptual experiences) to produce equally fine-grained, but higher-order, representations of those outputs. A version of this view was first proposed by the British Empiricist philosopher John Locke (1690). In our own time it has been defended especially by Armstrong (1968, 1984) and by Lycan (1996, 2004).

A terminological point: ‘inner-sense theory’ should more strictly be called ‘higher-order-sense theory’, since we of course have senses that are physically ‘inner’, such as pain-perception and internal touch-perception, that aren’t intended to fall under its scope. For these are first-order senses on a par with vision and hearing, differing only in that their purpose is to detect properties of the body, rather than of the external world (Hill 2004). According to the sort of higher-order theory that is presently under discussion, these senses, too, will need to have their outputs scanned to produce higher-order analog contents in order for those outputs to become phenomenally conscious. In what follows, however, the term ‘inner sense’ will be used to mean, more strictly, ‘higher-order sense’.

We therefore have the following proposal to consider:

Inner-Sense Theory:

A phenomenally conscious mental state is a state with analog/non-conceptual intentional content, which is in turn the target of a higher-order analog/non-conceptual intentional state, via the operations of a faculty of ‘inner sense’.

On this account, the difference between a phenomenally conscious percept of red and the sort of unconscious percepts of red that guide the guesses of a blindsighter and the activity of the sensorimotor system, is as follows. The former is scanned by our inner senses to produce a higher-order analog state with the content *experience of red* or *seems red*, whereas the latter states aren’t—they remain *merely* first-order states with the analog content *red*; and in so remaining, they lack any dimension of *seeming* or *subjectivity*. According to inner-sense theory, it is the higher-order perceptual contents produced by the operations of our inner-senses that make some mental states with analog contents, but not others, available to their subjects.

One of the main advantages of inner-sense theory is that it can explain how it is possible for us to acquire *purely recognitional concepts* of experience. 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 grounded in those higher-order analog contents. (Compare the way in which first-order perceptual contents representing color and sound enable us to acquire first-order recognitional concepts for colors and sounds.) And this should be possible without those recognitional concepts thereby having any conceptual connections with our beliefs about the content of the states recognized, nor with any of our surrounding mental concepts. This is then how inner-sense theory will claim to explain the familiar philosophical thought-experiments concerning one’s own experiences, which are supposed to cause such problems for physicalist/naturalistic accounts of the mind (Kripke 1972; Chalmers 1996). (For discussion of this ‘phenomenal concept strategy’ see Carruthers and Veillet 2007.)

For example, I can think, ‘*R* [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), experience of type *R* might occur without representing *red*. Likewise I can think, ‘*R* might occur in someone without occupying the role of *experience*, but rather (say) of belief.’ In the same sort of way, I shall be able to think, ‘*P* [an experience of pain] might have occurred in me, or might occur in others, in the absence of any of the usual causes and effects of pain. There could be someone in whom *P* experiences occur but who isn’t bothered by them, and where those experiences are never caused by tissue damage or other forms of bodily insult. And conversely, there could be someone who behaves and acts just as I do when in pain, and in response to the same physical causes, but who is never subject to *P* types of experience.’ If we

possess purely recognitional concepts of experience such as *R* and *P*, then the thinkability of such thoughts is unthreatening to a naturalistic approach to the mind.

Inner sense theorists are thus well placed to respond to those who claim that there is an unbridgeable explanatory gap between all physical, functional, and intentional facts, on the one hand, and the facts of phenomenal consciousness, on the other (Levine 1983; Chalmers 1996). And likewise they can explain the conceivability of zombies without becoming committed to the existence of any non-physical properties of experience (*contra* Chalmers 1996). It is the conceptual isolation of our higher-order recognitional concepts of experience that explains how there can be no a priori entailment between physical, functional, and intentional facts and the occurrence of states of type *R* or *P* (where *R* and *P* express purely recognitional concepts).

Inner-sense theory does face a number of difficulties, however. One objection is as follows (see Dretske 1995; Güzelçere 1995). If inner-sense theory were true, then how is it that there isn't any phenomenology distinctive of inner sense, in the way that there is a phenomenology associated with each outer sense? Since each of the outer senses gives rise to a distinctive set of phenomenological properties, one might expect that if there *were* such a thing as inner sense, then there would also be a phenomenology distinctive of its operation. But there doesn't appear to be any.

This point turns on the so-called 'transparency' of our perceptual experience (Harman 1990). Concentrate as hard as you like on your 'outer' (first-order) experiences—you won't find any *further* phenomenological properties arising out of the attention you pay to them, beyond those already belonging to the contents of the experiences themselves. Paying close attention to your experience of the color of the

red rose, for example, just produces attention to the *redness*—a property of the rose. Put like this, however, the objection just seems to beg the question in favor of first-order theories of phenomenal consciousness. It assumes that first-order—'outer'—perceptions already have a phenomenology independently of their targeting by inner sense. But this is just what an inner-sense theorist will deny. And then in order to explain the absence of any kind of higher-order phenomenology, an inner-sense theorist only needs to maintain that our higher-order perceptions are never themselves targeted by an inner-sense-organ which might produce *third-order* analog representations of them in turn.

Another objection to inner-sense theory is as follows (see Sturgeon 2000). If there really were an organ of inner sense, then it ought to be possible for it to malfunction, just as our first-order senses sometimes do. And in that case, it ought to be possible for someone to have a first-order percept with the analog content *red* causing a higher-order percept with the analog content *seems-orange*. Someone in this situation would be disposed to judge, 'It's red', immediately and non-inferentially (i.e. not influenced by beliefs about the object's normal color or their own physical state). But at the same time they would be disposed to judge, 'It *seems* orange'. Not only does this sort of thing never apparently occur, but the idea that it might do so conflicts with a powerful intuition. This is that our awareness of our own experiences is *immediate*, in such a way that to *think* that you are undergoing an experience of a certain sort *is* to be undergoing an experience of that sort. But if inner-sense theory is correct, then it ought to be possible for someone to believe that they are in a state of *seeming-orange* when they are actually in a state of *seeming-red*. (The problem of misrepresentation will be addressed further below in sections 6 and 7.)

A different sort of objection to inner-sense theory is developed by Carruthers (2000). It starts from the fact that the internal monitors postulated by such theories would need to have considerable

computational complexity in order to generate the requisite higher-order experiences. In order to perceive an experience, the organism would need to have mechanisms to generate a set of internal representations with an analog or non-conceptual content representing the content of that experience, in all its richness and fine-grained detail. And notice that any inner scanner would have to be a physical device (just as the visual system itself is) which depends upon the detection of those *physical* events in the brain that are the outputs of the various sensory systems (just as the visual system is a physical device that depends upon detection of physical properties of surfaces via the reflection of light). For it is hard to see how any inner scanner could detect the presence of an experience *qua* experience. Rather, it would have to detect the physical *realizations* of experiences in the brain, and construct the requisite higher-order representation of the experiences that those physical events realize, on the basis of that physical-information input. This makes it seem inevitable that the scanning device that supposedly generates higher-order experiences of our first-order visual experience would have to be almost as sophisticated and complex as the visual system itself.

Given this complexity in the operations of our organs of inner sense, there should be some plausible story to tell about the evolutionary pressures that led to their construction (Pinker 1994, 1997). But there would seem to be no such stories on the market. The most plausible suggestion is that inner-sense might have evolved to subserve our capacity to think about the mental states of conspecifics, thus enabling us to predict their actions and manipulate their responses. (This is the so-called ‘Machiavellian hypothesis’ to explain the evolution of intelligence in the great-ape lineage. See Byrne and Whiten 1988, 1998; and see Goldman 2006, for a view of inner sense of this sort.) But this suggestion presupposes that the organism must *already* have some capacity for higher-order *thought*, since it is such thoughts that inner sense is supposed to subserve. And yet as we shall see shortly (in section 5), some higher-order theories can claim all of

the advantages of inner-sense theory as an explanation of phenomenal consciousness, but without the need to postulate any ‘inner scanners’.

Lycan no longer holds HOP theory (Sauret and Lycan 2014) mainly because he now thinks that some sort of *attention* to first-order states is sufficient for an account of conscious states and there is little reason to suppose that the attentional mechanism in question is a higher-order representational state (see also Prinz 2012).

4. Higher-Order Thought Theory (1): Actualist

Actualist higher-order thought (HOT) theory is a proposal about the nature of state-consciousness in general, of which phenomenal consciousness is but one species. Its main proponent has been Rosenthal (1986, 1993, 2005). The proposal is this: a conscious mental state *M*, of mine, is a state that is actually causing an activated thought (generally a non-conscious one) that I have *M*, and causing it non-inferentially. (The qualification concerning non-inferential causation will be discussed in a moment.) An account of phenomenal consciousness can then be generated by stipulating that the mental state *M* should have some causal role and/or content of a certain distinctive sort in order to count as an experience (e.g., with an analog content, perhaps), and that when *M* is an experience (or a mental image, bodily sensation, or emotional feeling), it will be phenomenally conscious when (and only when) suitably targeted. The HOT is typically of the form: ‘I am in mental state *M*.’

We therefore have the following proposal to consider:

Actualist Higher-Order Thought Theory:

A phenomenally conscious mental state is a state of a certain sort (e.g. with analog/non-conceptual intentional content, perhaps) which is the

object of a higher-order thought, and which causes that thought non-inferentially.

As noted earlier, Rosenthal interprets the non-inferential requirement as ruling out only *conscious* inferences in the generation of a consciousness-making higher-order thought. This enables him to avoid having to say that my unconscious motives become conscious when I learn of them under psychoanalysis, or that my jealousy is conscious when I learn of it by noticing and interpreting my own behavior. But Rosenthal (2005) thinks that *unconscious* self-interpretation is acceptable as a source of the conscious status of the states thereby attributed. So if I arrive at the thought that I am feeling cheerful by unconsciously noticing the spring in my own step and the smile on my own face and drawing an unconscious inference, my cheerfulness will thereby have been rendered conscious. This aspect of Rosenthal's actualist form of HOT theory would appear to be optional for a HOT theorist, however.

In addition, and more controversially, Rosenthal (2005) thinks that the occurrence of a suitably caused HOT is *sufficient* for consciousness, even in the absence of any targeted first-order state (usually called 'targetless' or 'empty' HOTs). So I am undergoing a conscious experience of red provided that I *think* that I am undergoing an experience of red, even if I am actually in no first-order perceptual state whatever. This aspect of Rosenthal's view, too, appears optional for an actualist HOT theorist. Such a theorist can—and perhaps should—insist that phenomenally conscious experience occurs when and only when a first-order perceptual state causes a higher-order thought in the existence of that state in a way that doesn't depend upon self-interpretation. In recent years, the twin problems of *misrepresentation* between HOTs and their first-order targets as well as *targetless* HOT cases has led to significant disagreement among HOT theorists (see section 7 below).

The actualist HOT account avoids some of the difficulties inherent in inner-sense theory, while retaining the latter's ability to explain the distinction between conscious and unconscious perceptions. (Conscious perceptions will be analog states that are targeted by a HOT, whereas perceptions such as those involved in blindsight or subliminal perceptions will be unconscious by virtue of *not* being so targeted.) In particular, it is easy to see a function for HOTs, in general, and to tell a story about their likely evolution. A capacity to entertain HOTs about experiences would enable a creature to negotiate the is/seems distinction, perhaps learning not to trust its own experiences in certain circumstances, and also to induce appearances in others, by deceit. And a capacity to entertain HOTs about mental states (such as beliefs and desires) would enable a creature to reflect on, and to alter, its own beliefs and patterns of reasoning, as well as to predict and manipulate the thoughts and behaviors of others. Indeed, it can plausibly be claimed that it is our capacity to target higher-order thoughts on our own mental states that underlies our status as rational agents (Burge 1996; Sperber 1996; Rolls 2004).

A common initial objection to HOT theory (or even HOP) is that they lead to an infinite regress. It might seem that an infinite regress results because a conscious mental state (M) must be accompanied by a HOT, which, in turn must be accompanied by another HOT and so on. However, the standard and widely accepted reply or explanation is that when M is conscious, the HOT is not itself conscious (Rosenthal 1986, 2005). M is a first-order world-directed conscious state, such as a desire or perception, accompanied by an unconscious HOT. But when the HOT is itself conscious, there is a yet another higher-order (or third-order) thought directed at the conscious HOT. This would be a case of *introspection* according to HOT theory such that one's attention is directed inward at M (such as introspecting my desire). When this crucial distinction is overlooked, it can lead to some misguided objections such as supposing

that, according to HOT theory, having any conscious state (even for animals and infants) requires the ability to introspect.

One objection to HOT theory is due to Dretske (1993). We are asked to imagine a case in which we carefully examine two line-drawings, say (or in Dretske's example, two patterns of differently-sized spots). These drawings are similar in almost all respects, but differ in just one aspect—in Dretske's example, one of the pictures contains a black spot that the other lacks. It is surely plausible that, in the course of examining these two pictures, one will have enjoyed a conscious visual experience of the respect in which they differ—e.g. of the offending spot. But, as is familiar, one can be in this position while not knowing *that* the two pictures are different, or in what *way* they are different. In which case, since one can have a conscious experience (e.g. of the spot) without being aware that one is having it, consciousness cannot require higher-order awareness.

Replies to this objection have been made by Seager (1994), Byrne (1997), and Rosenthal (2005), among others. They point out that it is one thing to have a conscious experience of the aspect that differentiates the two pictures, and quite another to consciously experience that the two pictures are differentiated by that aspect. That is, consciously seeing the extra spot in one picture needn't mean seeing that this is the difference between the two pictures. So while scanning the two pictures one will enjoy conscious experience of the extra spot. A HOT theorist will say that this means undergoing a percept with the content *spot here* that forms the target of a HOT that one is undergoing a perception with that content. But this can perfectly well be true without one undergoing a percept with the content *spot here in this picture but absent here in that one*. And it can also be true without one forming any HOT to the effect that one is undergoing a perception with the content *spot here* when looking at a given picture but not when looking at the other. In which case the purported counter-example isn't really a counter-example.

Another objection to actualist HOT theory is epistemological, and is due to Goldman (2000). It turns crucially on the fact that the consciousness-making higher-order thoughts postulated by the theory are, themselves, characteristically *unconscious*. The objection goes like this. When I undergo a conscious mental state *M*, I generally know, or have good reason to believe, that *M* is conscious. But how can this be, if what *makes M* conscious is the existence of an *unconscious* HOT targeted on *M*? Since I don't know that this thought exists, it seems that I shouldn't be able to know that *M* is conscious, either. As Goldman himself acknowledges, however, this argument can only really work on the assumption that actualist HOT theory is supposed to be some sort of analytic or logical truth. Rosenthal has always made clear, however, that the theory isn't intended to be a piece of conceptual analysis, but is rather an account of the properties that *constitute* the property of being conscious (see Rosenthal 1986, as well as his 2005). And the epistemological argument gets no traction against this sort of view.

A different sort of problem with the actualist version of higher-order thought theory relates to the huge number of thoughts that would have to be caused by any given phenomenally conscious experience. (This is the analogue of the 'computational complexity' objection to inner-sense theory, sketched in section 3 above). Consider just how rich and detailed a conscious experience can be. It would seem that there can be an immense amount of which we can be consciously aware at any one time. Imagine looking down on a city from a window high up in a tower-block, for example. In such a case you can have phenomenally conscious percepts of a complex distribution of trees, roads, and buildings; colors on the ground and in the sky above; moving cars and pedestrians; and so on. And you can—it seems—be conscious of all of this simultaneously. According to actualist HOT theory, then, it seems you would need to have a distinct activated HOT for each distinct aspect of your experience—either that, or just a few such thoughts with immensely complex contents. Either way,

the objection is the same. For it seems implausible that all of this higher-order activity should be taking place (albeit non-consciously) every time someone is the subject of a complex conscious experience. What would be the point? And think of the amount of cognitive/neural space that these thoughts would take up! (In contrast, we know that neural tissue and activity are expensive; see Aiello and Wheeler 1995; and we also know that as a result of such constraints, the wiring diagram for the brain is about as efficient as it is possible for it to be; see Cherniak *et al.* 2004.)

This objection to actualist forms of HOT theory is considered at some length in Carruthers (2000), where a variety of possible replies are discussed and evaluated. Perhaps the most plausible and challenging such reply would be to deny the main premise lying behind the objection, concerning the rich nature of phenomenally conscious experience. The theory could align itself with Dennett's (1991) conception of consciousness as highly fragmented, with multiple streams of perceptual content being processed in parallel in different regions of the brain, and with no stage at which all of these contents are routinely integrated into a phenomenally conscious perceptual manifold. Rather, contents become conscious on a piecemeal basis, as a result of internal or external *probing* that gives rise to a HOT about the content in question. This serves to convey to us the mere *illusion* of riches, because wherever we direct our attention, there we find a conscious perceptual content. (For a related reply, see Gennaro 2012, chapter six).

It is difficult to know whether this sort of 'fragmentist' account can really explain the phenomenology of our experience, however. For it still faces the objection that the objects of attention can be immensely rich and varied at any given moment, hence requiring there to be an equally rich and varied repertoire of HOTs tokened at the same time. Think of immersing yourself in the colors and textures of a Van Gogh painting, for example, or the scene as you look out at your garden—it would seem that

one can be phenomenally conscious of a *highly* complex set of properties, which one could not even begin to describe or conceptualize in any detail. However, since the issues here are large and controversial, it cannot yet be concluded that actualist forms of HOT theory have been refuted. This is particularly the case when one considers such phenomena as change and inattentional blindness where subjects often do not even *notice* somewhat significant changes occurring in an image or video even within one's focal visual field (Simons 2000; Simons and Chabris 1999).

Another difficulty for actualist forms of HOT theory takes the form of a puzzle: how can the targeting of a perceptual state by HOT make the former 'light up', and acquire the properties of 'feel' or *what it is like-ness*? Suppose, for example, that I am undergoing an unconscious perception of red. How could such a percept then acquire the properties distinctive of phenomenal consciousness merely by virtue of me coming to think (in non-inferential fashion) that I am undergoing an experience of red?

Rosenthal (2005) replies to this objection by pointing to cases in which (he says) the acquisition and application of novel higher-order concepts to our experience transforms the phenomenal properties of the latter. Thus a course in wine-tasting can lead me to have experiences of the wine that are phenomenally quite distinct from any that I enjoyed previously (see also Siegel 2010; Gennaro 2012, chapter six). And a course in classical music appreciation might lead to changes in my experience of the sound of the orchestra, perhaps distinguishing between the sounds of the oboes and the clarinets for the first time. Since changes in higher-order concepts can lead to changes in phenomenal consciousness, Rosenthal thinks, it is plausible that it is the presence of higher-order thoughts targeting our perceptual states that is responsible for the latter's phenomenal properties *tout court*.

In response, an opponent of the theory might observe that some of the concepts that one acquires in such cases do not appear to be higher-order ones at all. Thus the concepts *oaky* and *tanniny* that one acquires when wine-tasting pick out secondary qualities of *the wine* (which are first-order), not higher-order properties of our experience of the wine. And likewise the concept *oboe* when applied in an experience is a first-order concept of a sound type, not a higher-order concept of one's experience of sound. The phenomenon here is quite general: acquiring and applying new concepts in one's perception can transform the similarity spaces and organization of one's perceptual states. (Think here of the familiar duck/rabbit.) But it appears to be a first-order phenomenon, not a higher-order one. At any rate, there is considerable work for a HOT theorist to do here in making out the case to the contrary.

5. Higher-Order Thought Theory (2): Dispositionalist

According to the dispositionalist HOT theory, the conscious status of an perceptual state consists in its *availability* to higher-order thought (Dennett 1978a; Carruthers 1996, 2000, 2005). As with the non-dispositionalist version of the theory, in its simplest form we have here a quite general proposal concerning the conscious status of any type of occurrent mental state, which becomes an account of phenomenal consciousness when the states in question are experiences (or images, emotions, etc.) with analog content. The proposal is this: a conscious mental event *M*, of mine, is one that is disposed to cause an activated thought (generally a non-conscious one) that I have *M*, and to cause it non-inferentially.

The proposal before us is therefore as follows:

Dispositionalist Higher-Order Thought Theory:

A phenomenally conscious mental state is a state of a certain sort

(perhaps with analog/non-conceptual intentional content, and perhaps held in a special-purpose short-term memory store) which is available to cause (non-inferentially) higher-order thoughts about itself (or perhaps about any of the contents of the memory store).

In contrast with the actualist form of theory, the higher-order thoughts that render a percept conscious are not necessarily actual, but potential. So the objection now disappears, that an unbelievable amount of cognitive space would have to be taken up with every conscious experience. (There need not *actually* be *any* HOT occurring, in order for a given perceptual state to count as phenomenally conscious, on this view.) So we might be able to retain our belief in the rich and integrated nature of phenomenally conscious experience—we just have to suppose that all of the contents in question are simultaneously *available* to higher-order thought. (Such availability might be realized by the 'global broadcast' of perceptual representations to a wide range of conceptual systems in the brain, for drawing inferences, for forming memories, and for planning, as well as for forming higher-order beliefs. See Baars 1988, 1997, 2002.) Nor will there be any problem in explaining why our faculty of higher-order thought should have evolved, nor why it should have access to perceptual contents in the first place—this can be the standard sort of story in terms of Machiavellian intelligence.

It might well be wondered how their mere *availability* to higher-order thoughts could confer on our perceptual states the positive properties distinctive of phenomenal consciousness—that is, of states having a *subjective* dimension, or a distinctive subjective *feel*. The answer may lie in the theory of content. Suppose that one agrees with Millikan (1984) that the representational content of a state depends, in part, upon the powers of the systems that *consume* that state. That is, suppose one thinks that *what* a state represents will depend, in part, on the kinds of inferences that the cognitive system is prepared to make in the presence of that state, or on

the kinds of behavioral control that it can exert. In that case the presence of first-order perceptual representations to a consumer-system that can deploy a ‘theory of mind’, and that is capable of recognitional applications of theoretically-embedded concepts of experience, may be sufficient to render those representations *at the same time* as higher-order ones. This would be what confers on our phenomenally conscious experiences the dimension of subjectivity. Each experience would at the same time (while also representing some state of the world, or of our own bodies) be a representation that we are undergoing just such an experience, by virtue of the powers of the ‘theory of mind’ system. Each percept of green, for example, would at one and the same time be an analog representation of *green* and an analog (non-conceptual) representation of *seems green* or *experience of green*. (Consumer semantics embraces not only a number of different varieties of *teleosemantics*, but also various forms of *inferential role semantics*. For the former, see Millikan 1984, 1986, 1989; and Papineau 1987, 1993. For the latter, see Loar 1981, 1982; McGinn 1982, 1989; Block 1986; and Peacocke 1986, 1992).

As an independent illustration of how consumer systems can transform perceptual contents, consider prosthetic vision (Bach-y-Rita 1995; Bach-y-Rita and Kerzel 2003). Blind subjects can be fitted with a device that transduces the output from a hand-held or head-mounted video-camera into patterns of electrically-induced tactile stimulation across the subject’s back or tongue. Initially, of course, the subjects just feel patterns of gentle tickling sensations spreading over the area in question, while the camera scans what is in front of them. But provided that they are allowed to control the movements of the camera themselves, their experiences after a time acquire three-dimensional distal intentional contents, representing the positions and movements of objects in space. (Note that the patterns of tactile simulations themselves become imbued with spatial content. The subjects in question say that it has come to *seem* to them that there is a spherical object moving towards them, for example.) Here everything on

the input side remains the same as it was when subjects first began to wear the device; but the planning and action-controlling systems have learned to interpret those states differently. And as a result, the subjects’ first-order intentional perceptual contents have become quite different. Likewise, according to dispositional HOT theory, when the ‘theory of mind’ system has learned to interpret the subject’s perceptual states *as* perceptual states: they all acquire a dimension of *seeming* or subjectivity.

Proponents of this account hold that it achieves all of the benefits of inner-sense theory, but without the associated costs. (Some potential draw-backs will be noted in a moment.) In particular, we can endorse the claim that phenomenal consciousness consists in a set of higher-order perceptions. This enables us to explain, not only the difference between conscious and unconscious perception, but also how analog states come to acquire a subjective dimension or ‘feel’. And we can also explain how it can be possible for us to acquire some purely recognitional concepts of experience (thus explaining the standard philosophical thought-experiments concerning zombies and such-like). But we don’t have to appeal to the existence of any ‘inner scanners’ or organs of inner sense (together with their associated problems) in order to do this. Moreover, it should also be obvious why there can be no question of our higher-order contents misrepresenting their first-order counterparts, in such a way that one might be disposed to make recognitional judgments of *red* and *seems orange* at the same time. This is because the content of the higher-order experience is parasitic on the content of the first-order one. Carruthers, therefore, also refers to this view as *dual content* theory.

On the downside, the account isn’t neutral on questions of semantic theory. On the contrary, it requires us to reject any form of pure input semantics, in favor of some sort of consumer semantics. We cannot then accept that intentional content reduces to informational content, nor that it can be explicated purely in terms of causal co-variance relations to the

environment. So anyone who finds such views attractive will think that the account is a hard one to swallow. (For discussion of various different versions of input semantics, see Dretske 1981, 1986; Fodor 1987, 1990; and Loewer and Rey 1991.)

Moreover, Rosenthal (2005) has objected that dispositional HOT theory can't account for our actual *awareness* of our conscious mental states, since mere dispositions to entertain thoughts doesn't make us aware of anything. Two replies can be made (see Carruthers 2000, 2005). One is that, in virtue of our disposition to entertain higher-order thoughts about it, a perceptual state will *already* possess an analog higher-order content. It is this content that makes us aware of the experience in question. But the second reply is that there does, in any case, seem to be a perfectly good dispositional sense of 'know' and 'aware'. As Dennett pointed out long ago (1978b), I can be said to know, or to be aware, that zebras in the wild don't wear overcoats, even though I have never actually considered the matter, because I am *disposed* to assent to that proposition in light of what I occurrently know.

In addition, Rowlands (2001) and Jehle and Kriegel (2006) have objected that dispositional HOT theory can't explain the sense in which the phenomenal properties of experience are *categorical*. For the higher-order analog intentional contents that our conscious perceptual states possess—and that are identified with the 'feel' of experience—are said to be constituted by the dispositional property that such states have, of giving rise to HOTs about themselves. This objection, however, appears to beg the question in favor of irreducible and intrinsic qualia as an account of the distinctive properties of phenomenally conscious states. In any case it doesn't seem to be an objection against dispositional HOT theory as such, since it will count equally against any representationalist theory of consciousness. (For example, Tye 1995, explains consciousness in terms of the *poisedness* of perceptual states to have an impact on belief and

reasoning, which is a dispositional notion.) Any theory that proposes to reductively explain phenomenal consciousness in terms of some combination of intentional content and causal role will be explaining consciousness in terms that are at least partly dispositional.

A well-known objection to dispositionalist higher-order thought theory, however, is that it may have to deny phenomenal consciousness to most species of non-human animal. This objection will be discussed, among others, in section 7, since it can be raised against *any* form of higher-order theory.

Carruthers no longer holds dispositional HOT theory or, for that matter, any form of higher-order theory and actually defends a version of first-order representationalism instead (Carruthers 2017). He responds to his own previous two main lines of argument against first-order representationalism and then finds it unnecessary to propose a higher-order theory in order to explain the difference between unconscious and conscious states. Still, Carruthers thinks that dispositional HOT theory is preferable to actualist HOT theory.

6. Self-Representational Higher-Order Theories

The two most familiar forms of higher-order theory postulate the existence of a pair of distinct mental states: a first-order perceptual or quasi-perceptual state with a given content, and a HOT or HOP representing the presence of that first-order state, thereby rendering it conscious. Either one of these states can occur without the other, although there may be a reliable causal relation between them, such that certain types of first-order perception (e.g. attended outputs of the temporal-lobe visual system) regularly cause higher-order representations of themselves to be formed. In recent years, however, a cluster of different proposals have been made that would reject this independent-state assumption. Rather, the

relationship between the conscious state in question and the higher-order state is said to be *constitutive*, or *internal*. To some extent, this view is inspired by Brentano (1874/1973) and the phenomenological tradition, including Sartre (1956). (See Kriegel 2006, 2018; Kriegel and Williford 2006; Zahavi 2004; Miguens et al. 2016.) We can refer to these as ‘self-representational’ higher-order theories. (Kriegel initially coined the term ‘same-order monitoring theory’ but this was potentially misleading).

We therefore have the following proposal to consider:

Self-Representational Theory:

A phenomenally conscious mental state is a state of a certain sort (perhaps with analog/non-conceptual intentional content) which also, at the same time, possesses an intentional content, thereby in some sense representing *itself* to the person who is the subject of that state.

There are two basic types of self-representational theory, depending on whether the constitutive relation between the conscious state and the higher-order state is one of *identity*, on the one hand, or *part-whole*, on the other. According to the former type of account, it is one and the same perceptual state that is both first-order (representing the world to us) and higher-order (presenting itself to us). (Caston 2002, argues that Aristotle had a theory of conscious perception of this sort.) Kriegel (2006) claims that such accounts are rather mysterious from a naturalistic perspective, but Carruthers (2000, 2005) and perhaps also Van Gulick (2001, 2004) purport to provide naturalistic explanations of just this sort of view. According to Carruthers, a first-order perceptual state with analog content acquires, at the same time, a higher-order analog content by virtue of its availability to a ‘theory of mind’ faculty, together with the truth of some suitable form of consumer semantics (as explained in section 5 above). Van Gulick can be interpreted as defending a similar view, which likewise relies on a form of consumer semantics/functional role semantics, which

he labels a ‘Higher-Order Global State (HOGS) theory’. On this account, globally broadcast first-order perceptual states acquire at the same time a higher-order *seeming* dimension though their availability to, and incorporation into, higher-order models of the self and its relation to the perceived environment. (What isn’t entirely clear is whether Van Gulick thinks that the resulting perceptual state *is* the HOG state, or is rather a component *part* of the HOG state—in which case he would be advocating a kind of part-whole self-representational account.)

Kriegel’s (2009) eventual view emphasizes and argues for the claim that there is a ubiquitous conscious (but inattentive or peripheral) self-awareness which accompanies all first-order (attentive and outer-directed) conscious states. Gennaro (2012, chapter five) rejects this view by, among other things, arguing that it is difficult to make sense of such alleged pervasive peripheral self-awareness especially when one is focused on outer-directed tasks. It is at least not as clearly present as, say, outer-directed peripheral vision in normal visual perception. At minimum, it is notoriously difficult to settle these sorts of disagreements between competing phenomenological claims.

Some varieties of part-whole self-representational theory take the same general form as actualist kinds of HOT theory, in which a first-order perceptual state with the content *analog-red* (as it might be) gives rise to a higher-order thought that one is experiencing red. But rather than claiming that it is the first-order perception that becomes phenomenally conscious because of the presence of the higher-order thought, what is said that the complex state made up of *both* the first-order perception *and* the higher-order thought becomes conscious. Gennaro (1996, 2008, 2012) defends such a view which he calls the *wide intrinsicality view* such that the HOT is better thought of as belonging to the same overall complex state as its target. It is, however, not always clear how this theory could offer any substantive benefits not already obtainable from actualist HOT theory.

Rather, the claim is merely that a conscious state is one that contains two parts, one of which is an awareness of the other. Kriegel himself (2003, 2006, 2009) and (as Kriegel interprets him) Van Gulick (2001, 2004) emphasize that the first-order perception and the higher-order judgment need to be *integrated* with one another in order for the resulting complex state to be phenomenally conscious. Kriegel argues that there needs to be a kind of integration resulting from a psychologically real process (as opposed to a theorist's definition) in order for the resulting state to have causal powers that differ from those of the first-order state/higher-order state pair.

Kriegel and Van Gulick do not give fully developed accounts of just *why* the integration of first-order perceptions with higher-order judgments should give rise to the properties that are distinctive of phenomenal consciousness. But one plausible reconstruction is as follows, modeled on the way that the conceptualization of analog (non-conceptual) first-order perceptual content can transform the latter's properties. Consider, for example, the familiar duck/rabbit. When someone sees this figure for the first time she may just experience a complex of curved lines, representing nothing. But when she comes to see it *as* a rabbit, those lines take on a certain distinctive organization (the figure now has both a front and a back, for example), thereby transforming the represented properties of the figure. Arguably what happens in such cases is that the conceptual systems succeed in deploying a recognitional template for the concept *rabbit*, finding a 'best match' with the incoming non-conceptual representations. Indeed, there is reason to think that just such a process routinely takes place in perception, with conceptual systems seeking matches against incoming data, and with the resulting states possessing contents that integrate both conceptual and non-conceptual (analog) representations (Kosslyn 1994; Carruthers 2000). The result is a single perceptual state that represents *both* a particular analog shape *and* a rabbit. Now suppose that when such states are globally broadcast and are made available to the

systems responsible for higher-order thought, a similar process takes place. Those systems bring to bear the concept *experience* or the concept *seeing* to produce a further integrated perceptual state. This single state will not only have first-order contents representing the lines on the page, and representing a rabbit, they will also have a higher-order content representing that one is *experiencing* something rabbit-like. Hence the perceptual state in question becomes 'self-presenting', and acquires, as part of its content, a dimension of *seeming* or *subjectivity*. (See also Gennaro 2005, 2012, for a related line of argument in response to this sort of challenge.)

Picciuto (2011) points out, however, that Kriegel's form of self-representational theory still permits a mismatch between the first-order and higher-order components of the integrated state. For there seems to be nothing in the structure of the account to rule out the possibility of a first-order analog content *green* becoming integrated with the higher-order judgment *I am experiencing yellow*, for example. In order to avoid this difficulty, Picciuto (2011) proposes an alternative form of part-whole self-representational theory. (See also Coleman, 2015; Timpe, 2015.) He does so by appropriating, and deploying for a novel purpose, the idea of a *quotational* phenomenal concept, originally introduced by Papineau (2002) and Balog (2009) as part of their defense of physicalism against the arguments of Chalmers (1996) and others. Picciuto's idea is that the relevant sort of complex self-representational state will consist of a first-order perceptual content combined with a higher-order concept like *experience* that embeds, or 'quotes' that very perceptual content. Given this structure, it will be impossible that there should be a mismatch between the two. For the higher-order component of the complex state is not a *judgment about* the experience component (which would permit them to be mismatched) but rather a concept that *quotes* the experience component.

All part-whole self-representational accounts differ from the dual-content theory of Carruthers (2000, 2005) in the following way, however: on Carruthers' account, the end-product can be entirely non-conceptual. And in particular, the higher-order content possessed by a conscious percept is a non-conceptual one, representing a *seeming* of the first-order content of the state by virtue of its availability to higher-order consumer systems. On all of the part-whole accounts sketched above, in contrast, a conscious perception is always partially conceptual, containing the higher-order concept *experience* (or something similar) as part of its content. There are probably multiple dimensions along which these two sorts of theory could be compared, and each may have its own advantages.

7. Objections to a Higher-Order Approach

There have, of course, been a whole host of objections raised against higher-order theories of phenomenal consciousness over the years. (See, e.g., Aquila 1990; Jamieson and Bekoff 1992; Dretske 1993, 1995; Goldman 1993, 2000; Güzeldere 1995; Tye 1995; Chalmers 1996; Byrne 1997; Siewert 1998; Levine 2001; Rowlands 2001; Seager 2004; Block, 2011.) Many of these objections, although perhaps intended as objections to higher-order theories as such, are often framed in terms of one or another particular version of such a theory. A general moral to be taken away from the present discussion should then be this: the different versions of a higher-order theory of phenomenal consciousness need to be kept distinct from one another, and critics should take care to state which version of the approach is under attack, or to frame objections that turn merely on the *higher-order character* of all of these approaches. I shall discuss a few 'local' objections first, before discussing some generic ones.

7.1 Local objections

A good many objections against specific versions of higher-order theory have already been discussed above. Thus in section 3 we discussed Dretske's (1995) 'lack of any higher-order phenomenology' objection to inner sense theory (which *only* targets inner sense theory). And in section 4 we discussed Dretske's (1993) 'spot' objection to actualist higher-order thought theory, as well as Goldman's (2000) epistemological objection, each of which appears to apply only to HOT theories. Of course some of the objections discussed above target more than one version of higher-order theory, while still not being fully general in scope. Thus the cognitive/computational complexity objections discussed in sections 3 and 4 apply to inner sense theories and to actualist HOT theories, but not to dispositionalist HOT or to some self-representational theories.

Another 'local' objection (which is actually a generalization of a variant of the misrepresentation problem discussed in connection with inner sense theory in section 3 above) is the targetless higher-order representation problem (Byrne 1997; Neander 1998; Levine 2001). This is confronted by both inner sense theory and actualist HOT theory (but not by either dispositionalist HOT or self-representational theories, according to which the relevant higher-order state can't exist in the absence of the targeted state). For in each case it seems that a higher-order experience of a perception of red, say, or a HOT about a perception of red, might exist in the absence of any such perception occurring. So it would *seem* to the subject that she is experiencing red, or she might *think* that she is experiencing red, in the absence of any such experience. (Note that the point isn't just that she might undergo such a seeming in the absence of anything really red. Rather, the point is that she might not really be undergoing any sort of visual experience *as of red* at all.) In which case, does the subject have a phenomenally conscious experience *as of* red, or not?

Both Lycan (1996) and Rosenthal (2005) are sanguine in the face of this objection. Each allows that targetless higher-order representations are a possibility (albeit rare, perhaps), and each opts to say that the subject in such a case is phenomenally conscious. But each denies that this is a problem for their account. Lycan, for example, insists that it is surely possible that it might *seem* to someone that she is feeling pain when really no relevant first-order representation of pain is present. (He suggests that the effects of morphine, which leaves patients saying that their pain feels just as it was, but that they no longer care, might constitute such a case.) And surely such a person would have a phenomenally conscious experience *as of* pain. Rosenthal, likewise, uses pain as an example. He points to cases of dental patients who initially experience pain in the dentist's chair despite the fact that the relevant nerves are completely destroyed. It seems that their fear, combined with the noise and vibration of the drill, causes them to mistakenly think that they are feeling pain. (When the drilling is stopped, and their dead nerves are explained to them, they thereafter experience only the sound and the vibration.) So this would be a case in which a HOT about experiencing pain is alone sufficient to induce a phenomenally conscious experience *as of* being in pain. A critic, however, might respond that the illusion is caused, instead, by a vivid *imagining* of pain, rather than by a HOT about feeling pain. Alternatively, if a HOT is causally involved it might be that a top-down expectation of pain *causes* a first-order experience of pain, as opposed to being *constitutive* of the feeling of pain. It might be that *introspective* anticipation of pain causes the pain in the first-place. (Note that this seems perfectly possible, since it is the opposite of well-known placebo effects of expectation in reducing pain.)

The targetless HOT problem has recently become a very significant topic of debate among HOT theorists as well as some critics (Block 2011; Rosenthal 2005, 2011; Weisberg 2011; Gennaro 2012, chapter four; Wilberg 2010; Berger 2014, 2017; Brown 2015; Lau and Brown 2019)

which has also led some to advocate for other variants of HOT theory or to clarify their own theories. Gennaro (2012) argues, for example, that since the HOTs in question are typically themselves unconscious, it makes little sense to suppose that these HOTs are phenomenally conscious in the context of HOT theory, especially since a conscious HOT would be an introspective state instead. Maintaining that an unconscious HOT would yield the same subjective experience without any target state seems to run counter to a central initial motivation of HOT theory, namely, to explain what makes a *first-order* state conscious. Thus, the first-order state must exist first in order to be rendered conscious by an appropriate and accompanying HOT with some sort of corresponding conceptual content. If both aren't present, then no relevant conscious state occurs (see also Wilberg 2010). Berger (2014), however, argues that consciousness is not a property of states at all; instead, it is a property of individual persons (that is, how my mental states appear to *me*). And Brown (2015) challenges the very assumption that HOT theory is best interpreted as a relational theory at all; instead, it is better construed as a HOROR theory, that is, higher-order representation of a representation, regardless of whether or not the target representation exists. In this sense, HOT theory is better understood as *non-relational* which also seems to be Rosenthal's considered view in recent decades. The debate continues (Rosenthal 2018).

In response specifically to Block (2011), Rosenthal (2011) and Weisberg (2011) stress that the mere *seeming* of, say, being in pain (provided by the HOT that one is in pain in the absence of first-order pain) is sufficient for phenomenally conscious pain. Consciousness is about mental *appearance*. It is not clear that this fully addresses Block's point, which is that one would not expect the mere thoughts that one feels pain to *matter* to us in all the ways that pain matters. Block develops this point with respect to the *awfulness* of pain. It would be remarkable (indeed, mysterious) if a higher-order thought should have all of the causal powers of the mental state that the thought is about. And in particular, there is no reason to expect that a

HOT that one is in pain should possess the negative valence and high-arousal properties of pain itself. But the latter are surely crucial components of phenomenally conscious pain. If so, then a HOT that one feels pain in the absence of first-order pain will *not* be sufficient for the conscious feeling of pain. It is also again important not to conflate introspection (conscious HOTs) with mere unconscious HOTs. (See also Shepherd 2013.)

Yet another ‘local’ objection is targeted against higher-order thought theories in particular (whether actualist or dispositionalist). It presents such theories with a dilemma: either they are attempts to explicate the *concept* of consciousness, in which case they are circular; or they are attempts to provide a reductive explanation of the *property* of being conscious, in which case they generate a vicious regress (Rowlands 2001). The first horn can be swiftly dismissed. For as Rosenthal (2005) and many others have made clear, higher-order theories aren’t in the business of conceptual analysis. Rather, their goal is to provide a reductive explanation of *what it is* for a state to be phenomenally conscious. Our discussion will therefore focus upon the second horn.

Rowlands thinks that HOT theories face a vicious regress because they explain state-consciousness in terms of HOT, and because (Rowlands claims) only *conscious* thoughts make us aware of the things that those thoughts concern. He gives the example of coming to believe that his dog is seriously ill. If he (Rowlands) thinks and behaves in ways that are best explained by attributing to him the thought that his dog is ill, but if that thought isn’t entertained consciously, then surely this won’t be a case in which he is *aware* that his dog is ill. So if we are to become aware of our conscious states by entertaining higher-order thoughts about them, then these thoughts will have to be conscious ones, requiring us to be aware of them, in turn, via further higher-order thoughts that are also conscious; and so on.

HOT theorists might respond in several ways: One is to challenge the intuition that only conscious thoughts make us aware of things. Thus it seems that Rowlands, when reflecting back on his dog-nurturing behavior of recent days, could surely conclude something along the lines of, ‘It seems that I have been aware of my dog’s illness all along; that is why I have been behaving as I have.’ Another response would be to allow that there is *a* way of understanding the concept of awareness such that a person only counts as aware of something if the mental state in virtue of which they are aware of that thing is itself a conscious one, but to deny that this is the relevant sense of ‘awareness’ which is put to work in HOT theories. A third option would be to stress the distinction between *phenomenal* consciousness and state consciousness more generally, claiming that there need be no regress involved in explaining the former in terms of the latter, provided that some separate account can be provided for the latter.

7.2 Generic objections

One generic objection, which can probably be recast in such a way as to apply to any higher-order theory (although it is most easily expressed against inner sense theory or actualist HOT theory), is the so-called ‘rock’ objection (Goldman 1993; Stuenkel 1998). We don’t think that when we become aware of a rock (either perceiving it, or entertaining a thought about it) that the rock thereby becomes conscious. So why should our higher-order awareness of a mental state render that mental state conscious? Thinking about a rock doesn’t make the rock ‘light up’ and become phenomenally conscious. So why should thinking about my perception of the rock make the latter phenomenally conscious, either?

An initial reply to this objection involves pointing out that my perception of the rock is a *mental* state, whereas the rock itself isn’t (Lycan 1996). Since phenomenal consciousness is a property that (some) mental states

possess, we can then say that the reason why the rock isn't rendered phenomenally conscious by my awareness of it is that it isn't the right *sort* of thing to *be* phenomenally conscious, whereas my perception of the rock is. This reply may be apt to strike the objector as trite. But perhaps more can be said from the perspective of inner sense theory, at least. Notice that my perception of the rock does, in one sense, confer on the latter a subjective aspect. For example, the rock is represented from one particular spatial perspective, and only some of its properties (e.g. color) and not others (e.g. mass) are represented. Likewise, then, with my perception of the rock.

Similar replies to the rock objection are given by Van Gulick (2001) and Gennaro (2005). Both point out that a rock isn't the kind of thing that can be incorporated *into* a complex mental state that involves higher-order representations, in the sort of way required by a self-representational or HOT theory. In contrast, whether actualist HOT theory can reply adequately to the rock objection will depend on whether or not there is an adequate reply to the problem considered in section 4, which challenges the actualist HOT theorist to say why targeting a mental state with a HOT about that state should cause the latter to 'light up' and acquire a subjective dimension or *feel*.

Another generic objection is that higher-order theories, when combined with plausible empirical claims about the mental abilities of non-human animals, will conflict with our common-sense intuition that such animals enjoy phenomenally conscious experience (Jamieson and Bekoff 1992; Dretske 1995; Tye 1995; Seager 2004). This objection can be pressed most forcefully against higher-order *thought* theories, of either variety, and against self-representational theories; but it is also faced by inner-sense theory (depending on what account can be offered of the evolutionary function of organs of inner sense). Are cats and dogs really capable of having such apparently complex HOTs which presumably contain mental

state concepts? Since there has been considerable dispute as to whether even chimpanzees (and other primates) have the kind of sophisticated 'theory of mind' to enable them to entertain thoughts about experiential states as such (Byrne and Whiten 1988, 1998; Povinelli 2000), it seems implausible that many other species of mammal (let alone reptiles, birds, and fish) would qualify as phenomenally conscious, on these accounts (Carruthers 2000, 2005). Yet the intuition that such creatures enjoy phenomenally conscious experiences is a powerful one, for many people. (Witness Nagel's classic 1974 paper, which argues that there must be something that it is like to be a bat.)

Many higher-order theorists have attempted to resist the claim that their theory has any such entailment (e.g. Gennaro 1996, 2004; Van Gulick 2006). In each case, a common strategy is to claim that the relevant higher-order representations are somehow *simpler* than those tested for by those who do comparative 'theory of mind' research, hence leaving it open that these simpler representations might be widespread in the animal kingdom. Gennaro (1996), for example, suggests that although animals might lack the concept *experience*, they can nevertheless be capable of higher-order indexical thoughts of the form '*this* is different from *that*' (where 'this' and 'that' might refer to experiences of red and of green, respectively). The trouble here, however, is to explain what makes these indexicals higher-order in content without attributing concepts like *experience of green* to the animal.

Gennaro (2004) takes a somewhat different tack. While allowing that animals lack the concept *experience of green*, he thinks that they might nevertheless possess the (simpler) concept *seeing green*. But here he faces a dilemma. There is, indeed, a simpler concept of seeing, grounded in the capacity to track eye-direction and line of sight. But this isn't necessarily a higher-order concept. To say, in this sense, that someone sees green in just to say that there is some green in the line in which their eyes are pointed—

no mental state needs to be attributed. In contrast, it appears that any concept of seeing that is genuinely higher-order will be one that it would be less plausible to attribute to most species of animal (given the comparative evidence). Perhaps a first-order explanation of experimental observations of animals is virtually always possible (see, for example, Carruthers 2008). But Gennaro (2012, chapter eight) ultimately argues that there is plenty of evidence that many animals are capable of metacognition (thinking about their own mental states) as well as mindreading (thinking about other minds). For example, in the case of mindreading, rhesus monkeys seem to attribute visual and auditory perceptions to others in more competitive paradigms (Flombaum and Santos 2005) and crows and scrub jays return alone to caches seen by other animals and recache them in new places (Emery and Clayton 2001). Any evidence of deception or empathy in animals would also seem to indicate some kind of mindreading ability. In addition, many animals seem capable of metacognition (including possessing self-concepts) as evidenced by the presence of episodic memory, for example (Dere et al. 2006; see also the essays in Terrace and Metcalf 2005; Hurley and Nudds 2006). A related debate takes place with respect to infant consciousness and the capacity of infants to have metacognitive and mindreading abilities (see Gennaro 2012, chapter seven, for some discussion).

Van Gulick (2006), in contrast, suggests that all of the higher-order representing sufficient to render an experience phenomenally conscious can be left merely *implicit* in the way that the experience enters into relationships with other mental states and the control of behavior. So animals that lack the sorts of explicit higher-order concepts tested for in comparative ‘theory of mind’ research can nevertheless be phenomenally conscious. The difficulty here, however, is to flesh out the relevant notion of implicitness in such a way that not every mental state, possessed by every creature (no matter how simple), will count as phenomenally conscious. For since mental states can’t occur singly, but are always part

of a network of other related states, mental states will always carry information about others, thus implicitly representing them. It is implicit in the behavior of any creature that drinks, for example, that it is thirsty; so the drinking behavior implicitly represents the occurrence of the mental state of thirst.

Of course, the basis for the common-sense intuition that animals possess phenomenally conscious states can even be challenged. (How, after all, are we supposed to *know* whether it is like something to be a bat?) And that intuition can perhaps be explained away as a mere by-product of imaginative identification with the animal. (Since our *images* of their experiences are phenomenally conscious, we may naturally assume that the experiences *imaged* are similarly conscious (Carruthers 1999, 2000). But there is no doubt that one major source of resistance to higher-order theories will lie here, for many people, especially given various moral considerations about animal pain and suffering. (For one set of attempts to defuse this resistance, arguing that a higher-order account need have few if any implications for our moral practices or for comparative psychology, see Carruthers 2005, chapter nine; 2019, chapter eight.) Of course, some will point out that there are also enough *neurophysiological similarities* between (at least some parts of) human and animal brains to justify attributions of, say, pains, desires, and basic perceptual states. It is worth emphasizing here that HOT theory does *not* say that having conscious states requires having *introspective* states, that is, having conscious HOTs. Conflating introspection with having mere unconscious HOTs (and therefore simply having first-order conscious states) may lead some to put forth a misguided straw man argument against HOT theory.

A third generic objection is that higher-order approaches cannot really *explain* the distinctive properties of phenomenal consciousness (Chalmers 1996; Siewert 1998; Levine 2006). Whereas the argument from animals is that higher-order representations aren’t *necessary* for phenomenal

consciousness, the argument here is that such representations aren't *sufficient*. It is claimed, for example, that we can easily conceive of creatures who enjoy the postulated kinds of higher-order representation, related in the right sort of way to their first-order perceptual states, but where those creatures are wholly *lacking* in phenomenal consciousness.

In response to this objection, higher-order theorists will join forces with first-order theorists and others in claiming that these objectors pitch the standards for explaining phenomenal consciousness too high (Block and Stalnaker 1999; Tye 1999; Carruthers 2000, 2005; Lycan 2001). They will insist that a reductive explanation of something—and of phenomenal consciousness in particular—doesn't have to be such that we cannot conceive of the *explanandum* (that which is being explained) in the absence of the *explanans* (that which does the explaining). (Indeed, we can also *explain why* no such explanation can be forthcoming, in terms of our possession of purely recognitional concepts of experience.) Rather, we just need to have good reason to think that the explained properties are *constituted by* the explaining ones, in such a way that nothing *else* needed to be added to the world once the explaining properties were present, in order for the world to contain the target phenomenon. But this is hotly contested territory. And it is on this ground that the battle for phenomenal consciousness may ultimately be won or lost.

Before we close, it is worth considering a variant of the third generic objection that we have just been discussing, which need involve no commitment to the latter's demanding standards of explanation. For it might be said that self-representing mental states (or indeed any of the theoretically-relevant kinds of pairing of first-order with higher-order representations) might occur within the unconscious mind, without (of course) thereby becoming conscious (Rey, 2008). Suppose that some version of Freudian theory is true, for example. Might there not be higher-order thoughts about the subject's experiences occurring within the

unconscious mind, formed while the latter tries to figure out how to get its messages past the 'censor' and expressed in speech? So here again, just as with the third generic objection, the claim is that the occurrence of the sorts of representations postulated by higher-order theories isn't *sufficient* for phenomenal consciousness.

One sort of response to this objection would be to deny that such purely unconscious higher-order cognition ever *actually* occurs. Indeed, one might deny that it is even possible, given the constraints provided by the evolution of cognitively demanding mental functions (Carruthers 2000). But note that this reply will be unavailable to any higher-order theorist who has opted to downplay the cognitive demands of the capacity for higher-order representation in response to the problem of animal consciousness. For if higher-order representation is easily evolved, and is rife within the animal kingdom, then there doesn't appear to be any reason why it shouldn't evolve within unconscious sub-systems of the mind as well. And in any case it is doubtful whether the mere *natural* impossibility of higher-order representing within the unconscious mind would be enough to rebut the objection. Since higher-order theories claim that phenomenal consciousness is to be *identified* with, or is *constituted by*, the relevant sorts of higher-order representing, we would need to show that the imagined occurrence of the latter within the unconscious mind is *metaphysically* impossible, not just that it is naturally so.

Other responses to the objection remain (Carruthers 2000, 2005). One would be to allow that unconscious phenomenal consciousness is possible, and to appeal to the distinction between *phenomenal* consciousness and *access* consciousness to explain away the seeming contradiction involved. Remember, phenomenally conscious states are those that it is *like* something to be in, and that possess a subjective *feel*; whereas access-conscious states are those that are available to interact with some specified cognitive processes (for example, they might be those that are reportable

in speech). So all we would be saying is that states with *feel* can occur in ways that aren't (for example) reportable by the subject. There is no contradiction here. An alternative possible response would be to extend the higher-order theory in question to include the relevant sort of access-consciousness as a further component. A dispositional HOT theorist, for example, might say that a phenomenally conscious state is one that *both* possesses the right sort of dual content *and* that is reportable by the subject. I shall not attempt to adjudicate between these possibilities here.

(Objections have also been raised as to how (or if) HOT theory and self-representational theories can account for various pathologies of self-awareness or 'depersonalization disorders,' such as somatoparaphrenia and thought insertion in schizophrenia. See the essays in Gennaro 2015 for some discussion.)

Bibliography

- Aiello, L. and Wheeler, P., 1995. 'The expensive tissue hypothesis,' *Current Anthropology*, 36: 199–221.
- Aquila, R., 1990. 'Consciousness as higher-order thoughts: two objections,' *American Philosophical Quarterly*, 27: 81–87.
- Armstrong, D., 1968. *A Materialist Theory of the Mind*. London: Routledge.
- , 1984. 'Consciousness and causality,' in D. Armstrong and N. Malcolm (eds.), *Consciousness and Causality*, Oxford: Blackwell.
- Baars, B., 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
- , 1997. *In the Theatre of Consciousness*. Oxford: Oxford University Press.
- , 2002. 'The conscious access hypothesis: origins and recent evidence,' *Trends in Cognitive Sciences*, 6: 47–52.
- Bach-y-Rita, P., 1995. *Non-Synaptic Diffusion Neurotransmission and*

- Late Brain Reorganization*. New York: Demos Press.
- Bach-y-Rita, P. and Kercel, S., 2003. 'Sensory substitution and the human-machine interface,' *Trends in Cognitive Sciences*, 7: 541–546.
- Balog, K. 2009. 'Phenomenal concepts,' in B. McLaughlin, A. Beckermann, and S. Walter (eds.), *Oxford Handbook in the Philosophy of Mind*, Oxford: Oxford University Press.
- Berger, J., 2014. 'Consciousness is not a property of states: a reply to Wilberg,' *Philosophical Psychology*, 27: 829–842.
- Berger, J., 2017. 'How things seem to higher-order thought theorists,' *Dialogue*, 56: 503–526.
- Block, N., 1986. 'Advertisement for a semantics for psychology,' *Midwest Studies in Philosophy*, 10: 615–678.
- , 1995. 'A confusion about a function of consciousness,' *Behavioral and Brain Sciences*, 18: 227–247.
- , 2011. 'The higher-order approach to consciousness is defunct,' *Analysis*, 71, 419–431.
- Block, N. and Stalnaker, R., 1999. 'Conceptual analysis, dualism and the explanatory gap,' *Philosophical Review*, 108: 1–46.
- Brentano, F., 1874/1973. *Psychology From an Empirical Standpoint*. New York: Humanities.
- Brown, R., 2015. 'The HOROR theory of phenomenal consciousness,' *Philosophical Studies*, 172: 1783–1794.
- Burge, T., 1996. 'Our entitlement to self-knowledge,' *Proceedings of the Aristotelian Society*, 96: 91–116.
- Byrne, A., 1997. 'Some like it HOT: consciousness and higher-order thoughts,' *Philosophical Studies*, 86: 103–129.
- , 2004. 'What phenomenal consciousness is like,' in R. Gennaro (ed.) 2004, pp. 203–226.
- Byrne, R. and Whiten, A. (eds.), 1988. *Machiavellian Intelligence*. Oxford: Oxford University Press.
- , (eds.), 1998. *Machiavellian Intelligence II: Evaluations and*

- extensions*. Cambridge: Cambridge University Press.
- Carruthers, P., 1989. 'Brute experience,' *Journal of Philosophy*, 86: 258–269.
- , 1996. *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.
- , 1999. 'Sympathy and subjectivity,' *Australasian Journal of Philosophy*, 77: 465–482.
- , 2000. *Phenomenal Consciousness: a naturalistic theory*. Cambridge: Cambridge University Press.
- , 2005. *Consciousness: essays from a higher-order perspective*. Oxford: Oxford University Press.
- , 2008. 'Meta-cognition in animals: a skeptical look,' *Mind and Language*, 23: 58–89.
- , 2017. 'In defence of first-order representationalism,' *Journal of Consciousness Studies*, 24: 74–87.
- , 2019. *Human and Animal Minds*. Oxford: Oxford University Press.
- Carruthers, P. and Veillet, B., 2007. 'The phenomenal concept strategy,' *Journal of Consciousness Studies*, 14 (9–10).
- Caston, V., 2002. 'Aristotle on consciousness,' *Mind*, 111: 751–815.
- Chalmers, D., 1996. *The Conscious Mind*. Oxford: Oxford University Press.
- Cherniak, C., Mokhtarzada, Z., Rodriguez-Esteban, R., and Changizi, B., 2004. 'Global optimization of cerebral cortex layout,' *Proceedings of the National Academy of Sciences*, 101: 1081–1086.
- Coleman, S., 2015. 'Quotational higher-order thought theory,' *Philosophical Studies*, 172: 2705–2733.
- Dennett, D., 1978a. 'Toward a cognitive theory of consciousness,' in C. Savage (ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Minneapolis: University of Minnesota Press. (Reprinted in Dennett 1978b.)
- , 1978b. *Brainstorms*. Cambridge, MA: MIT Press.
- , 1991. *Consciousness Explained*. London: Allen Lane.
- Dere, E., Kart-Teke, E., Huston, J., and Silva, D., 2006. 'The case for episodic memory in animals,' *Neuroscience and Biobehavioral Reviews*, 30: 1206–1224.
- Dretske, F., 1981. *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.
- , 1986. 'Misrepresentation,' in R. Bogdan (ed.), *Belief*, Oxford: Oxford University Press.
- , 1988. *Explaining Behavior*, Cambridge, MA: MIT Press.
- , 1993. 'Conscious experience,' *Mind*, 102: 263–283.
- , 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.
- Emery, N. and Clayton, N., 2001. 'Effects of experience and social context on prospective caching strategies in scrub jays,' *Nature*, 414: 443–446.
- Flombaum, J. and Santos, L., 2005. 'Rhesus monkeys attribute perceptions to others,' *Current Biology*, 15: 447–452.
- Fodor, J., 1987. *Psychosemantics*. Cambridge, MA: MIT Press.
- , 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.
- Gennaro, R., 1996. *Consciousness and Self-Consciousness*. Amsterdam: John Benjamins.
- , 2004. 'Higher-order thoughts, animal consciousness, and misrepresentation,' in R. Gennaro (ed.) 2004, pp. 45–66.
- , (ed.) 2004, *Higher-Order Theories of Consciousness*, Philadelphia: John Benjamins.
- , 2005. 'The HOT theory of consciousness: between a rock and a hard place,' *Journal of Consciousness Studies*, 12: 3–21.
- , 2006. 'Between pure self-referentialism and the (extrinsic) HOT theory of consciousness,' in Kriegel and Williford (ed.) 2006.
- , 2012. *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*, Cambridge, MA: MIT press.

- , (ed.) 2015. *Disturbed Consciousness: New Essays on Psychopathology and Theories of Consciousness*, Cambridge, MA: MIT press.
- Glover, S., 2004. ‘Separate visual representations in the planning and control of action,’ *Behavioral and Brain Sciences*, 27: 3–24.
- Goldman, A., 1993. ‘Consciousness, folk-psychology, and cognitive science,’ *Consciousness and Cognition*, 2: 364–382.
- , 2000. ‘Can science know when you are conscious?’ *Journal of Consciousness Studies*, 7 (5): 3–22.
- , 2006. *Simulating Minds: the philosophy, psychology, and neuroscience of mind-reading*. Oxford: Oxford University Press.
- Graziano, M., 2013. *Consciousness and the social brain*, Oxford: Oxford University Press.
- Güzeldere, G. 1995. ‘Is consciousness perception of what passes in one’s own mind?’ in T. Metzinger (ed.), *Conscious Experience*, Paderborn: Ferdinand Schöningh; reprinted in N. Block, O. Flanagan, and G. Güzeldere (eds.), *The Nature of Consciousness*, Cambridge: MIT Press, 1997.)
- Harman, G., 1990. ‘The intrinsic quality of experience,’ *Philosophical Perspectives*, 4: 31–52.
- Hellie, B., 2007. ‘Higher-order intentionalism and higher-order acquaintance,’ *Philosophical Studies*, 134: 289–324.
- Hill, C., 2004. ‘Ouch! An essay on pain,’ in R. Gennaro (ed.) 2004, pp. 339–362.
- , 2006. ‘Perceptual consciousness: how it opens directly onto the world, preferring the world to the mind,’ in U. Kriegel and K. Williford (eds.), *Self-Representational Approaches to Consciousness*, Cambridge, MA: MIT Press.
- Hurley, S. and Nudds, M., (eds.) 2006. *Rational Animals?* New York: Oxford University Press.
- Jacob, P. and Jeannerod, M., 2003. *Ways of Seeing*. Oxford: Oxford University Press.
- Jackson, F., 1982. ‘Epiphenomenal qualia,’ *Philosophical Quarterly*, 32: 127–136.
- , 1986. ‘What Mary didn’t know,’ *Journal of Philosophy*, 83: 291–295.
- Jamieson, D. and Bekoff, M., 1992. ‘Carruthers on non-conscious experience,’ *Analysis*, 52: 23–28.
- Jehle, D. and Kriegel, U., 2006. ‘An argument against dispositional HOT theory,’ *Philosophical Psychology*, 19: 463–476.
- Kirk, R., 1994. *Raw Feeling*. Oxford: Oxford University Press.
- Kosslyn, S., 1994. *Image and Brain*. Cambridge, MA: MIT Press.
- Kozuch, B., 2014. ‘Prefrontal lesion evidence against higher-order theories of consciousness,’ *Philosophical Studies*, 167: 721–746.
- Kriegel, U., 2003. ‘Consciousness as intransitive self-consciousness: two views and an argument,’ *Canadian Journal of Philosophy*, 33: 103–132.
- , 2006. ‘The same-order monitoring theory of consciousness,’ in U. Kriegel and K. Williford (eds.) 2006.
- , 2009. *Subjective Consciousness*. Oxford: Oxford University Press.
- , 2018. ‘Brentano’s dual-framing theory of consciousness,’ *Philosophy and Phenomenological Research*, 97: 79–98.
- Kriegel, U. and Williford, K. (eds.), 2006. *Self-Representational Approaches to Consciousness*. Cambridge, MA: MIT Press.
- Kripke, S., 1972. ‘Naming and necessity’, in G. Harman and D. Davidson (eds.), *Semantics of Natural Language*, Dordrecht: Reidel. (Revised version printed in book form by Oxford: Blackwell, 1980.)
- Lau, H. and Brown, R., 2019. ‘The emperor’s new phenomenology? The empirical case for conscious experience without first-order representations’, in A. Pautz and D. Stoljar (eds.) *Blockheads! Essays on Ned Block’s Philosophy of Mind and Consciousness*. Cambridge, MA: MIT Press, pp. 171–198.
- Lau, H. and Passingham, R., 2006. ‘Relative blindsight in normal





- observers and the neural correlate of visual consciousness,' *Proceedings of the National Academy of Sciences*, 103: 18763–18768.
- Lau, H. and Rosenthal, D., 2011. 'Empirical support for higher-order theories of conscious awareness,' *Trends in Cognitive Sciences*, 15: 365–373.
- Levine, J., 1983. 'Materialism and qualia: the explanatory gap,' *Pacific Philosophical Quarterly*, 64: 354–361.
- , 2001. *Purple Haze*. Cambridge, MA: MIT Press.
- , 2006. 'Conscious awareness and (self-)representation,' in U. Kriegel and K. Williford (eds.) 2006.
- Locke, J., 1690. *An Essay Concerning Human Understanding*. (Many editions now available.)
- Loewer, B. and Rey, G. (eds.), 1991. *Meaning in Mind: Fodor and his critics*. Oxford: Blackwell.
- Lycan, W., 1987. *Consciousness*. Cambridge, MA: MIT Press.
- , 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.
- , 2001a. 'Have we neglected phenomenal consciousness?' *Psyche*, 7. Available from the ASSC depository
- , 2001b. 'A simple argument for a higher-order representation theory of consciousness,' *Analysis*, 61: 3–4.
- , 2004. 'The superiority of HOP to HOT,' in R. Gennaro (ed.) 2004, pp. 93–114.
- McGinn, C., 1982. 'The structure of content,' in A. Woodfield (ed.), *Thought and Object*, Oxford: Oxford University Press.
- , 1989. *Mental Content*. Oxford: Blackwell.
- , 1991. *The Problem of Consciousness*. Oxford: Blackwell.
- Mehta, N., 2013. 'Is there a phenomenological argument for higher-order representationalism?,' *Philosophical Studies*, 164: 357–370.
- Miguens, S., Preyer, G., and Morando, C., (eds.) 2016. *Pre-Reflective Consciousness: Sartre and Contemporary Philosophy of Mind*. London: Routledge Publishers.
- Millikan, R., 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.
- , 1986. 'Thoughts without laws: cognitive science with content,' *Philosophical Review*, 95: 47–80.
- , 1989. 'Biosemantics,' *Journal of Philosophy*, 86: 281–297.
- Milner, D. and Goodale, M., 1995. *The Visual Brain in Action*. Oxford: Oxford University Press.
- Nagel, T., 1974. 'What is it like to be a bat?' *Philosophical Review*, 83: 435–456.
- , 1986. *The View from Nowhere*. Oxford: Oxford University Press.
- Neander, K., 1998. 'The division of phenomenal labor: a problem for representational theories of consciousness,' in J. Tomberlin (ed.), *Language, Mind, and Ontology*, Oxford: Blackwell.
- Nelkin, N., 1996. *Consciousness and the Origins of Thought*. Cambridge: Cambridge University Press.
- Odegaard, B., Knight, R., and Lau, H. 2017. "Should a few null findings falsify prefrontal theories of conscious experience?" *The Journal of Neuroscience*, 37: 9593–9602.
- Papineau, D., 1987. *Reality and Representation*. Oxford: Blackwell.
- , 2002. *Thinking about Consciousness*. Oxford: Oxford University Press.
- , 1993. *Philosophical Naturalism*. Oxford: Blackwell.
- Peacocke, C., 1986. *Thoughts*. Oxford: Blackwell.
- , 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.
- Phillips, B., 2014. 'Indirect representation and the self-representational theory of consciousness,' *Philosophical Studies*, 167: 273–290.
- Picciuto, V., 2011. 'Addressing higher-order misrepresentation with quotational thought,' *Journal of Consciousness Studies*, 18(3–4): 109–136.
- Pinker, S., 1994. *The Language Instinct*. London: Penguin Press.

- , 1997. *How the Mind Works*. London: Penguin Press.
- Povinelli, D., 2000. *Folk Physics for Apes*. Oxford: Oxford University Press.
- Prinz, J., 2012. *The Conscious Brain*. New York: Oxford University Press.
- Rey, G., 2008. '(Even higher-order) intentionality without consciousness,' *Revue Internationale de Philosophie*, 62: 51–78.
- Rolls, E. 2004. 'A higher-order syntactic thought (HOST) theory of consciousness', In R. Gennaro (ed.) 2004, pp. 137–172.
- Rosenthal, D., 1986. 'Two concepts of consciousness,' *Philosophical Studies*, 49: 329–359.
- , 1993. 'Thinking that one thinks,' in Davies and Humphreys (eds) 1993.
- , 2004. 'Varieties of higher-order theory,' in R. Gennaro (ed.) 2004, pp. 17–44.
- , 2005. *Consciousness and Mind*. Oxford: Oxford University Press.
- , 2011. 'Exaggerated reports: reply to Block,' *Analysis*, 71: 431–437.
- , 2018. 'Misrepresentation and mental appearance,' *TransFormAcao*, 41: 49–74.
- Rounis, E., Maniscalco, B., Rothwell, J., Passingham, R., and Lau, H., 2010. 'Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness,' *Cognitive Neuroscience*, 1: 165–175.
- Rowlands, M., 2001. 'Consciousness and higher-order thoughts,' *Mind and Language*, 16: 290–310.
- Sartre, J.P., 1956. *Being and Nothingness*. New York: Philosophical Library.
- Sauret, W. and Lycan, W., 2014. 'Attention and internal monitoring: a farewell to HOP,' *Analysis*, 74: 363–370.
- Searle, J., 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.
- , 1997. *The Mystery of Consciousness*. New York: New York Review of Books.
- Seager, W., 1994. 'Dretske on HOT theories of consciousness,' *Analysis*, 54: 270–276.
- , 2004. 'A cold look at HOT theory,' in R. Gennaro (ed.) 2004, pp. 255–276.
- Shepherd, J., 2013. 'Why Block can't stand the HOT,' *Journal of Consciousness Studies*, 20: 183–195.
- Siegel, S., 2010. *The Contents of Visual Perception*. New York: Oxford University Press.
- Siewert, C., 1998. *The Significance of Consciousness*. Princeton: Princeton University Press.
- Simons, D., 2000. 'Current approaches to change blindness,' *Visual Cognition*, 7: 1–15.
- Simons, D. and Chabris, C., 1999. 'Gorillas in our midst: sustained inattentive blindness for dynamic events,' *Perception*, 28: 1059–1074.
- Sperber, D., 1996. *Explaining Culture*. Oxford: Blackwell.
- Stubenberg, L., 1998. *Consciousness and Qualia*, Amsterdam: John Benjamins.
- Sturgeon, S., 2000. *Matters of Mind: consciousness, reason and nature*. London: Routledge.
- Terrace, H. and Metcalfe, J., (eds.) 2005. *The Missing Link in Cognition: Origins of Self-Reflective Consciousness*. New York: Oxford University Press.
- Timpe, K., 2015. 'Quotational higher-order thought theory' *Philosophical Studies*, 172: 2705–2733.
- Tye, M., 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
- , 1999. 'Phenomenal consciousness: the explanatory gap as cognitive illusion,' *Mind*, 108: 705–725.
- , 2000. *Color, Consciousness, and Content*. Cambridge, MA: MIT

Press.

- Van Gulick, R., 2001. 'Inward and upward: reflection, introspection, and self-awareness,' *Philosophical Topics*, 28: 275–305.
- , 2004. 'Higher-order global states (HOGS): an alternative higher-order model of consciousness,' in R. Gennaro (ed.) 2004, pp. 67–92.
- , 2006. 'Mirror, mirror: is that all?' in Kriegel and Williford (eds.) 2006.
- Weisberg, J., 2011. 'Abusing the notion of what-it's-like-ness: A response to Block,' *Analysis*, 71: 438–443.
- Weiskrantz, L., 1986. *Blindsight*. Oxford: Oxford University Press.
- , 1997. *Consciousness Lost and Found*. Oxford: Oxford University Press.
- Wilberg, J., 2010. 'Consciousness and false HOTs,' *Philosophical Psychology*, 23: 617–638.
- Williford, K., 2006. 'The self-representational structure of consciousness,' in Kriegel and Williford (eds.) 2006.
- Zahavi, D., 2004. 'Back to Brentano'? *Journal of Consciousness Studies*, 11 (10–11): 66–87.

Academic Tools

-  How to cite this entry.
-  Preview the PDF version of this entry at the Friends of the SEP Society.
-  Look up this entry topic at the Internet Philosophy Ontology Project (InPhO).
-  Enhanced bibliography for this entry at PhilPapers, with links to its database.

Other Internet Resources

- Association for the Scientific Study of Consciousness E-print Archive
- Bibliography on Higher-Order Theories of Consciousness, at PhilPapers.
- Cognitive Science E-print Archive

Related Entries

animal: consciousness | consciousness | consciousness: and intentionality | consciousness: representational theories of

Copyright © 2020 by the authors
Peter Carruthers and Rocco Gennaro