The Southern Journal of Philosophy (2008) Vol. XLVI

Higher-Order Thought and the Problem of Radical Confabulation

Timothy Lane
National Chengchi University, Taiwan

Caleb Liang
National Taiwan University, Taiwan

Abstract

Currently, one of the most influential theories of consciousness is Rosenthal's version of higher-order-thought (HOT). We argue that the HOT theory allows for two distinct interpretations: a one-component and a two-component view. We further argue that the two-component view is more consistent with his effort to promote HOT as an explanatory theory suitable for application to the empirical sciences. Unfortunately, the two-component view seems incapable of handling a group of counterexamples that we refer to as cases of radical confabulation. We begin by introducing the HOT theory and by indicating why we believe it is open to distinct interpretations. We then proceed to show that it is incapable of handling cases of radical confabulation. Finally, in the course of considering various possible responses to our position, we show that adoption of a disjunctive strategy, one that would countenance both one-component and two-component versions, would fail to provide any empirical or explanatory advantage.

According to David Rosenthal's influential higher-order thought (HOT) theory of consciousness, what makes a mental state conscious is that it is accompanied by a suitable higher-order

---

Tim Lane has studied broadly, working at various times in three different PhD programs. Problems pertaining to mind and consciousness have always held his interest, especially those problems that afford the possibility of integrating conceptual analysis with empirical work. He currently works in the Office of Research and Development at National Chengchi University in Taiwan.

Caleb Liang is assistant professor of philosophy at National Taiwan University. His research interests are in the philosophy of mind and perception. He recently published "Phenomenal Character and the Myth of the Given (Journal of Philosophical Research, 2006).

thought (Rosenthal 1991, 2002, 2004, and 2005).[1] More specifically, Rosenthal (e.g., 2002, 408–11) hypothesizes that mental states are conscious just in case they are the objects of occurent, assertoric, seemingly noninferential thoughts to the effect that self is in said state. The relevant kinds of higher-order thoughts need not themselves be conscious, but to be suitable in Rosenthal's terms, they should not seem to be mediated,[2] not be those that we are merely disposed to have, and not be those that are just imagined, wondered, hoped, or desired.[3]

Rosenthal has repeatedly made it clear that his theory is proposed as an explanation of, among other things, phenomenal consciousness[4] (henceforth, p-consciousness). More specifically, his principal goal is to *explain* "what it is in virtue of which conscious states differ from mental states that aren't conscious" (2005, 3). He has also emphasized that HOTs, just like other kinds of thoughts, can misrepresent. But several authors (e.g., Byrne 1997; Neander 1998; Rowlands 2001; Seager 1999; and Levine 2001) have expressed doubts about many claims made on behalf of HOT theory, including the claims that concern p-consciousness and misrepresentation. One worry is that, since the HOTs themselves are usually unconscious, how could it be that they are able to give rise to there being *something-it-is-like* for the subject to be in a sensory state? A rather hyperbolic way of expressing this concern has sometimes been called the "problem of the rock." On Goldman's (1993, 366) version: "A rock does not become conscious when someone has a belief about it. Why should a first-order psychological state become conscious simply by having a belief about it?" A standard response from the HOT theorists is that, in order to be a conscious state, the object of HOT must be a mental state (Rosenthal 2005; Gennaro, 2005). Thoughts about rocks don't make them conscious, because rocks aren't mental states.

Another worry combines the claims made concerning p-consciousness and misrepresentation. This is the suspicion that if the content of the first-order state can be represented with little or no fidelity by HOTs, then the first-order state seems irrelevant to p-consciousness. Yet more vexing, it seems possible that HOTs can be the source of p-conscious states, even when the first-order states are completely lacking; thus, we are left with the seemingly contradictory claims that first-order sensory states must be accompanied by HOTs in order for there to be states that are p-conscious, along with the claim that HOTs can be sufficient for p-consciousness.[5] That the sensory qualities of first-order states are sometimes unnecessary is sometimes referred to as "the problem of targetless higher-order thought" (cf. also Kriegel 2003; Janzen Forthcoming).

Rosenthal partially addresses these worries in his discussion of such cases as wine tasting, parafoveal vision, the "filling in" of blind spots, dental fear, the-subject-as-target, and confabu-

lation. But we believe that his responses remain inadequate. In this paper, we argue that the problem of targetless higher-order thoughts does pose a serious problem for the HOT theory. We differ from other critics, however, in that while they focus principally on conceptual issues,[6] our focus is on explanatory adequacy, a failing that we regard as more damaging to Rosenthal because he regards HOTs as posits of an empirical theory, and it is a theory that has been adopted, with modifications, and promoted by members of the scientific community.[7] Our choice of focus is motivated by a desire to engage Rosenthal on that field of endeavor for which his theory was specially designed. Below, we proceed as follows: section 1 briefly describes how Rosenthal's theory attempts to explain p-consciousness. Section 2 presents what we call the *problem of radical confabulation*, a condition that can be manifest either when first-order content is completely misrepresented or when there simply is no relevant first-order state. We intend to show that this is an *empirical* issue for the HOT theory, an issue that the theory—in light of its stated purpose—cannot dismiss lightly. Section 3 considers some possible responses.

1.

Rosenthal's version of HOT is intended as an empirical theory that can, among other things, explain p-consciousness, the *what-it-is-like* aspect of conscious experience.[8] First, he emphasizes that not all mental states are conscious; it is possible for someone to undergo a qualitative or sensory state without being conscious of it (Rosenthal 1997, 2002, 2005). Here is a recent formulation of the claim:

> To be qualitative, a property need not always occur consciously; it must simply be able to occur consciously. Nonmental, physiological properties, by contrast, are never conscious. Qualitative properties are potentially conscious, not invariably or essentially conscious. (2005, 177)

Sensory qualities—understood as the properties of certain states that enable us "to discern similarities and differences" (e.g., Rosenthal 2005, 202)—as evidenced by subliminal perception, peripheral vision, blindsight, the cocktail party effect, headaches that come and go during the day, and other such phenomena—seem to be able to occur without p-consciousness (Rosenthal 2002, 411).[9] They seem to be mental states, not mere physiological ones, because their functional roles parallel the functional roles of conscious sensory states. The distinction applies to both perception and to bodily sensations. As for the latter, just as a conscious pain might cause me to adjust my posture, so too would an unconscious pain cause me to adjust

my position while I sleep.[10] Rosenthal's preferred pain example is of having headaches for extended periods, even though intermittent distractions often make us seemingly unaware of the pain. Rosenthal would say that during those nonconscious periods, the headache continues to play a functional role that befits a mental state.[11] As these examples show, sensory qualities can be teased apart from what-it-is-like.

What makes a sensory state p-conscious, according to Rosenthal, is that it is accompanied by a suitable HOT. He says: "accompanying HOTs do result in there being something it is like for one to be in states with those sensory qualities" (2002, 413). A recent elaboration of the same claim is: "How I represent to myself the sensation I have determines what it's like for me to have it. Differences in my HOTs result in differences in what it's like for me to have my qualitative states" (2005, 187).

These statements seem to show that Rosenthal's theory attempts to explain p-consciousness in terms of two components. First, there must be a first-order mental state. In the case of perceptual experience, the mental state would be a sensory state, one that could perform perceptual functions, even were it not conscious. Second, there must be a suitable HOT that makes the subject conscious of it. By Rosenthal's theory, the HOTs can't carry all the weight in explaining p-consciousness. As he says, "Strictly speaking, having a HOT cannot of course result in a mental state's being conscious if that mental state does not even exist" (1997, 744). This claim seems to imply that for there to be p-consciousness, the HOTs must target a sensory state. When they do so, "they make us *conscious of ourselves as being in certain qualitative states*, which results in the subjective impression of conscious mental qualities" (Rosenthal 2002, 413). We believe this is Rosenthal's view, properly understood. Let us call this account a *two-component view* of p-consciousness.

By way of adducing support for this view, Rosenthal asserts that there is a "striking connection" between what HOTs "we are able to have and what sensory qualities we are able to be aware of" (2002, 413–14; also see 1997, 745). Reasoning by inference-to-the-best-explanation, he claims that those HOTs are what enable us to be in conscious states with just those sensory qualities. If this line of reasoning is correct, it would then seem to follow that learning new concepts for, say, gustatory and olfactory experiences as when one is learning to be a wine taster, usually enables us to be conscious of more fine-grained differences, among sensory qualities.[12]

Rosenthal further observes that what is true for finely differentiated qualities is equally true for crudely individuated qualities. Taking the sound of an oboe as an example, he hypothesizes that one can construct a scale of HOTs from the finely differentiated to the crudely undifferentiated: we can, for

example, move from classification of sensations as "the sound of an oboe," to classification as "a woodwind," to some yet more generic type of sound, and then—yet more crudely undifferentiated—to something not even distinctive of sound (perhaps just indiscriminate sensory experience). If we continue to extrapolate beyond this point, "peel away" the least differentiated HOTs so to speak, then we would be left without any conscious experience whatsoever (2002, 413–14). The point is that, without a suitable HOT, all we have would be unconscious, auditory, mental states.

An important implication of this apparent two-component view is that it allows for the possibility that a HOT can misrepresent the sensory state one is in. And perhaps this should not come as a surprise for, after all, our HOTs, our occurent assertions, particularly when they seem to be isolated from inferential access to the external world, can go seriously astray. Thus, we can form "erroneous" HOTs that make it seem, from the first-person perspective, as though we are in sensory states we aren't actually in. One example of this is provided by the capacity of HOTs to compensate for the low-resolution sensations of parafoveal vision; HOTs can make those sensations seem clear and focused (see Rosenthal 1997, 744). In such cases, the low-level resolutions of first-order states are misrepresented; the first-order sensory state is there, but it is made to seem more distinct and focused than it actually is.

One among the other examples provided by Rosenthal (2002, 415) is the phenomenon sometimes known as dental fear. In such cases, dental patients seem to be experiencing pain even when nerve damage or anesthetic prevents them from being in the actual sensory state. The usual explanation of this phenomenon is that a fearful or anxious reaction to vibrations caused by the drill is what leads one to feel the pain, pain that is subjectively indistinguishable from pains felt when the actual sensory state obtains. According to the HOT theory, what happens is that one occurently believes that one is in a pain state. If the patient is later told that, for example, an anesthetic has already been administered, the patient will cease feeling pain. Information thus provided can lead to inferences that alter the subsequent subjective experiences. But that information doesn't change the patient's sense of what the prior experience was like, for conscious experiences are not altered by inferences based upon subsequent information. Rosenthal claims that the HOT theory effectively explains this phenomenon: the vibrations sensed by the patient serve as first-order states, while the subjective feeling of pain is brought about by the HOTs having misrepresented those first-order states in a way consistent with the patient's beliefs about the common consequences of dentistry. This proposed explanation seems to be compatible with a two-component interpretation of HOT.

But we must hasten to add that, strictly speaking, Rosenthal's account is open to another interpretation.[13] He claims that not only can HOTs misrepresent their targets, "they may even be about something that does not exist at all" (2005, 210). Here Rosenthal might be understood as suggesting that even though first-order sensory states are frequent causal antecedents for HOTs, they are not essential. Wishful thinking, self-deception, confabulation, and other like phenomena can also serve as causal antecedents and allow for the possibility of HOTs that are about notional perceptions or sensations, perceptions or sensations that don't exist. Rosenthal is careful to emphasize various constraints on the degree to which HOTs can diverge from targets (212), but, in the end, he allows that, "there can be something it's like for one to be in a state with particular mental qualities, even if no such state occurs" (211). HOTs have "the last word." Call this a *one-component* interpretation of the HOT theory; it will be discussed below, in the final section.

Others have criticized Rosenthal's position concerning the claim that HOTs can target nonexistents and still bring about the what-it-is-like of p-consciousness. As Kriegel observes, the claim that a person can "be under the impression that she is in a conscious state when in reality she is not" is, to say the least, "highly counterintuitive" (2007, 49). Rosenthal does not deny that his claim is counterintuitive; instead, he treats this reaction as just a pretheoretic[14] limitation on our intuitions, a limitation that is not as troublesome as it might seem. Pretheoretic experience, after all, is not a good guide to the causal antecedents of consciousness of self as being in a qualitative state, nor should we expect it to be.

We stand with Rosenthal on the claim that pretheoretic intuitions should not be taken as cause for excessive worry. Moreover, as we have indicated above, our concern resides primarily with explanatory adequacy, not with alleged problems of conceptual coherence. It is for this reason that we have devoted special attention to Rosenthal's examples, for it is here that explanatory adequacy can be better assessed. We claim that Rosenthal's examples (e.g., parafoveal vision, tasting wine, or listening to an oboe), at least those that can more clearly been seen to serve an explanatory role, presuppose a two-component view that suffers from explanatory deficiencies.

This completes our brief introduction to Rosenthal's account of how p-consciousness can be explained. Despite its brevity, we believe it accurately identifies the critical features of his account. In the next two sections, we argue that his view fails to provide a satisfactory explanatory account of the phenomenal character of consciousness. Below we will also further consider Rosenthal's claim that his theory does not, strictly speaking, require two components.

2.

In allowing for cases of misrepresentation, perhaps an allowance that an empirical theory that aspires to appropriate explanatory scope must make, Rosenthal has created a problem for himself. On the one hand, as mentioned in the previous section, aspects of his theory coupled with certain telling examples tend to imply that HOTs cannot, by themselves, account for p-consciousness. But on the other hand, as we will see, in confabulatory cases the HOTs seem, at least sometimes, to have a free hand. In this section, we challenge Rosenthal's account of confabulation. We distinguish between nonradical and radical cases of confabulation and suggest that his view faces what we call *the problem of radical confabulation*. We argue, first, that some of the cases of confabulation that Rosenthal has discussed—including parafoveal vision and the "filling-in" of blind spots—are nonradical. For others—for example, dental fear—his explanations are problematic in ways that he seems not to recognize. Second, we argue that there are radical cases of confabulation that the two-component view does not adequately explain. Our contention is that, with regard to p-consciousness, in particular the what-it-is-like aspect, Rosenthal's HOT theory is not empirically adequate.

By nonradical cases of confabulation we mean cases that fall under one (or more) of the following three categories: (a) those for which a relevant first-order target state clearly exists, (b) those for which only a less obviously relevant first-order target state exists, and (c) those that are perhaps better understood as instances of epistemic overconfidence. As for (a), consider parafoveal vision. It might be said that the HOTs still faithfully represent a first-order sensory state, but they extrapolate erroneously, exaggerating its coverage. The same might be said for such phenomena as the "filling-in" of blind spots. No content alien to the target state need be added.

As for (b), consider Rosenthal's account of dental fear once again. According to the two-component view, the vibrations sensed by the patient are taken to be first-order states, which are misrepresented by the HOTs as pain. (The role of anxiety here, while not intuitively or medically surprising, does not comfortably fit into the HOT model of misrepresentation.) This case is not as straightforward as either parafoveal vision or unnoticed blind spots. Misrepresentation seems to be doing more here than it is in either of the previous cases. Here it can be said that the HOT has a target, but the content of that target state seems less relevant to what is represented. Unlike the treatment given to parafoveal vision or blind spots, the HOT can't simply extrapolate from first-order content; the HOT must *actively, creatively* misrepresent. Vibrations cum fear are not pain,[15] but the HOT misrepresents them as pain. Or so it seems, on Rosenthal's description.

As for (c), what we call epistemic overconfidence, no sufficiently distinct example can be found within Rosenthal's work, so we here cite a common textbook case: according to Nisbett and Wilson (1977), when subjects were invited to choose from a variety of panty hose displayed in a row, they preferred those on their right. But their preference explanations made no reference to spatial orientation; instead, the subjects spoke of such (seemingly phenomenal) qualities as color and texture. Even after being told that the panty hose were identical and they were simply showing a preference for those on their right-hand sides, subjects tended to defend their confabulations.

The reason for treating (c) as distinct from (a) and (b) is that although subjects may report phenomenal qualities, it is somewhat less clear that they are actually experiencing phenomenal qualities. Many empirical studies of confabulation do not strongly or consistently suggest that what is confabulated is consciously experienced. Subject reports might seem consistent with a HOT explanation—that is, the HOTs might be said to misrepresent visual and haptic continuity, leading the subject to be conscious of discontinuity. But belief construction of this sort need not imply anything at all about p-consciousness, as various clinical and technical definitions of confabulation seldom include specific reference to p-consciousness (Hirstein 2005, 187–203).[16]

So as not to be misunderstood, we should make it clear that we agree that Rosenthal should try to explain misrepresentation as it might apply to p-consciousness; after all, as exemplified by confabulation, it is a common human phenomenon, exhibited in both mundane, nonpathological cases, and in exotic, pathological ones. It might just be a basic feature of the way we form and report beliefs, a form of epistemic overconfidence (Hirstein 2005; Lane 2006). As for the panty hose case, if we put the worries mentioned above aside, the subjects might be reporting judgments that are grounded in p-consciousness in a way that is consistent with what Rosenthal says about HOT theory: despite being wrong about the cause of their behavior, they could be said to be conscious of color in virtue of having HOTs that target certain sensory states, but said states seem to be radically misrepresented, for they are making distinctions where there are no distinctions to be made. In this case, the first-order state does not seem to carry *any* of the content that is being represented by the HOT, at least none of the content that enables the making of distinctions. So it is unlike parafoveal vision that simply builds upon what is already available to it. It "sees" differences where there are none to be seen. It doesn't misrepresent by indicating "more of the same"; rather, it misrepresents by indicating "difference despite sensory continuity." As with dental fear, we seem to have a case wherein HOTs can be said to have targets, but if we presuppose

that the subjects' reports are reliable and literal, their p-consciousness includes sensory qualities (i.e., the discontinuities in color and texture) that don't allow for the type of nonradical extrapolation available to the cases considered above.

We have suggested that the two-component view, when applied to dental fear and to the panty hose cases, misrepresents in ways that create a constructivist burden for HOTs, a burden not shared by blind spot or parafoveal vision cases. But we will not press the point closely here because we believe that neither represents a substantial challenge for HOT theory. For example, dental fear, since there is a target state, and since "fear" and "vibration" bear significant conceptual similarities to "pain," a plausible, relatively straightforward defense of HOT can certainly be managed: the relevant HOT does target first-order sensory states, but these states are misnamed. Fear cum vibration is unpleasant, as is pain; perhaps patients are just misnaming their unpleasant experiences in the way that subjects who can distinguish among colors readily enough might incorrectly name them. As for the Nisbett and Wilson panty hose case, perhaps an argument could be made that despite references to p-consciousness, there is no evidence that the subjects actually had such experiences. Perhaps their behavior can be entirely treated at the doxastic level.[17] We are not fully convinced that such is the case, but we choose to avoid these complexities for now. Instead, we use these constructivist examples as a bridge or transition to examples that we believe serve as yet more serious challenges to the explanatory ambitions of HOT.

Accordingly, we now turn to radical confabulation, by which we mean cases of radical misrepresentation, radical in that (a) there is good reason to say that there simply is no target, yet (b) there are robust grounds for claiming that p-consciousness actually obtains. Rosenthal's theory not only allows for misrepresentation by extrapolation from target content and by creative additions to target content, as we have indicated above, it also allows for the possibility of *radical* misrepresentation: that is, cases in which those first-order sensory states simply don't exist (e.g., 2004, 31). However, he does not believe that this poses serious problems for his theory. Addressing this issue, Rosenthal writes:

> Suppose my higher-order awareness is of a state with property P, but the target isn't P, but rather Q. We could say that the higher-order awareness misrepresents the target, but we could equally well say that it's an awareness of a state that doesn't occur. The more dramatic the misrepresentation, the greater the temptation to say the target is absent; but it's plainly open in any such case to say either. The two kinds of cases, moreover, should occasion the same kinds of phenomenological perplexities, if any. A higher-order

> awareness of a *P* state without any *P* state would be subjectively the same whether or not a *Q* state occurs. The first-order state can contribute nothing to phenomenology apart from the way we're conscious of it (2004, 32).[18]

Rosenthal seems to be anticipating our distinction, or like distinctions, between nonradical and radical cases of confabulation and downplaying its significance. On one reading of this passage, it might seem that Rosenthal is trying to distance himself from a two-component interpretation of HOT. But we believe his attempt to downplay the possibility of targetless HOTs is significant. Note that he writes: "We could say that the higher-order awareness misrepresents the target, but we could equally well say that it's an awareness of a state that doesn't occur. The more dramatic the misrepresentation, the greater the temptation to say the target is absent, but it's *plainly open* in any such case to say either" (Rosenthal 2004, 32; emphasis added). We believe that Rosenthal can make such a claim only because he deals with nonradical cases. The radical cases, we argue, are importantly different: they show that it is not "*plainly open* ... to say either."[19]

Case 1.

Consider the case of Anton's Syndrome,[20] in particular, the denial of blindness (Hirstein 2005, 145–46). Typically those who suffer from Anton's Syndrome (AS) have two brain lesions: one that affects vision and another that affects the ability to know whether or not they can see. The result is that although their neural mechanism cannot process visual stimulation, they believe they can, and they can skillfully confabulate when their failure to see things accurately is called to their attention. Confidence in confabulations is not shaken even when the patients are interacting with the environment; naturally they bump into things and encounter other problems, but they confabulate anyway, for example, by complaining that the room is too dark, or that they're not wearing their glasses, or that they're not familiar with the environment.

Here, although we seem to have an example of HOTs without sensory experience, a defender of HOT might, as with the case cited above, attempt to restrict explanation to the doxastic level. But a yet more striking characteristic of AS is that patients "often have visual hallucinations (that can be either simple or complex)." Typically, hallucinations, as opposed to *mere* confabulations, are said to have three characteristics:[21] (a) they are vivid and immediate, like perceptions; (b) they are experienced as though external to the person; and (c) they are not reassessed as imagery, just because evidence suggesting they should be so reassessed is provided. These three charac-

teristics might also be applicable to confabulatory cases wherein relevant first-order target states exist; they might, arguably, even be applicable to such cases wherein a less obviously relevant first-order target state exists.[22] But clearly they are not applicable to cases of epistemic overconfidence. And the problem posed for HOT is that (a), (b), and (c) seem to obtain, as clear indicators of p-consciousness, even when the subject suffers from AS.

In one case, for example, a patient with AS not only reported hallucinations (while suffering from delirium tremens),[23] he also believed his vision had been restored to the extent that he could provide descriptions of his environment (Swartz and Brust 1984). Unlike those cases for which defenders of HOT might be better able to restrict explanation to doxastic problems, here we have better evidence that p-consciousness, in the form of substantial visual experiences actually occurs. Here then we seem to have a case of radical confabulation for which it is difficult to deny that p-consciousness obtains, but for which the HOT seems to be doing all the work. In a word, the HOT seems to be *sufficient* for p-consciousness. This violates the two-component interpretation of Rosenthal's view; it does not seem "plainly open" to interpretation as target misrepresentation.

The problem here is that not only are the HOTs misrepresenting and confabulating, as in the case of dental fear or in the case of choosing panty hose, there seem to be grounds for saying there is no mental state to be targeted; yet the subject still seems capable of p-consciousness. In the cases of AS accompanied by visual hallucinations, although the patient claims to have visual experiences, apparently the HOTs *cannot be targeting* the sensory states in question.

A defender of HOT theory might want to claim that here too we have insufficient evidence for the existence of p-consciousness. Perhaps explanation can once again be handled by reference to intentional states alone. But we would regard such a response as ad hoc, because the diagnosis of hallucination presupposes p-consciousness, and that diagnosis is made on independent, medical grounds. Notice we are not suggesting that the reports of empirical scientists should be taken at face value. But we are claiming that the burden to correct a diagnosis rests with the critic who encounters an anomalous case that does not conform to theory expectations. Moreover, HOT is designed to explain empirical phenomena, so unless a principled reason can be given to reject the standard scientific description, the diagnosis and its conceptual implications should stand.

What we are claiming is that *if* Rosenthal's version of HOT is both intended as an explanation of p-consciousness and as allowing for the possibility of radical confabulation, then it flirts with inconsistency. On the one hand, it seems to imply that doxastic states alone—even HOTs—are insufficient as an

explanation of consciousness. The other mental states, in this instance the first-order sensory states, seem to be necessary. But on the other hand, the HOT theory also seems anxious to allow for radical confabulation, and this makes the first-order sensory states unnecessary. We further contend that radical cases are not ambiguous in such a way as to support Rosenthal's claim that they could "plainly" be reinterpreted as instances of target misrepresentation.

This is not a result that Rosenthal could want, for even he seems to share the widespread intuition that intentional states, alone, simply don't have the resources for enabling phenomenal consciousness (1997, 740; 2002, 413). It is "being able to form intentional states about certain sensory qualities" that results in "our being able to experience those qualities consciously" (2002, 413). But here there are no sensory qualities that can be intentionally targeted. At least for the radical cases, it seems to be inconsistent to allow for the possibility of erroneous HOTs alone giving rise to *something-it-is-like* for the subject to be in a sensory state.

One immediate response would be to treat it as an exceptional case, something that is beyond the reach of the existing version of the theory.[24] But actually this might not be such an exceptional case. There are other examples in the literature on pathologies that suggest that p-consciousness can occur despite the absence of the sort of target that is required by the HOT theory. To cite just two more cases:

Case 2.

Congenitally deaf patients who suffer from schizophrenia often claim to hear voices (see Atkinson 2006). Schizophrenia is as common among the deaf as it is among the general population, and approximately half of these patients report hearing voices. The voices are described as being similar to regular speech, in that they can vary along multiple dimensions, including loudness, pitch, content, and complexity. Moreover, voices are often personified such that the patients can detect accent, gender, and degree of familiarity. It must be admitted that there are many contentious issues concerning the proper interpretation of these data; for example, it is not altogether clear how we should understand hallucination descriptions that use signs glossed in English as "Heard," or "Voices." But at least this is an indication that cases of AS accompanied by hallucinations are not so exceptional as they might seem. In some respects, the phenomenon of deaf-hearing might be even more problematic for Rosenthal because these patients don't believe that they can hear. They claim to hear, despite not believing that they have the capacity for hearing. On the face of it, they seem to be lacking an essential prerequisite for formation of the relevant

HOT—the belief that they are capable of hearing. Nevertheless, they do, with the onset of schizophrenia, begin to hear voices.

Contrast this with Rosenthal's wine tasting example. In that case, it is alleged that learning new concepts enables one to be conscious of fine-grained differences among sensory qualities. Concept learning takes a leadership role. The same can be said concerning Rosenthal's thought-experiment wherein it is alleged that by systematically stripping away concepts, "the sound of an oboe" can be reduced to indiscriminate sensory experience. In these cases, concepts are enablers. But when the deaf suddenly start to hear, the concepts don't lead or enable, at least not in the way that it seems to be with wine or oboes. When the deaf begin to hear, their concepts seem capable of nothing more than merely responding to phenomena that abruptly assert themselves onto the scene. And this is especially odd because the phenomena are in direct conflict with previously held beliefs concerning what the agent is capable of experiencing.

Case 3.

Ramachandran discusses the cases of phantom limb patients who experience involuntary clenching spasms, such that nails can be felt digging into the palm, spasms that inflict great pain that the patient cannot relieve (see Ramachandran 1998; Ramachandran and Rogers-Ramachandran 1996). In order to provide therapeutic assistance to these patients, Ramachandran designed a mirror-box, a box that contains a centrally placed, vertical mirror, a box without a top or front. Patients can place their existing hands into the box, thereby creating the illusion that the phantom has been resurrected. As a consequence of this "resurrection," six out of ten patients claimed they could now feel the phantoms move and four of the ten were able to unclench their the hands, thereby relieving the pain.

On at least one description of such cases, patients who don't believe in the resurrection of the phantom hand are suddenly able to exercise control over it and eliminate select p-conscious experience (i.e., the nails digging into the palm). That is, there seems to be no intentional target, at least no obvious sensory state available for targeting. And there is no belief that the hand has been resurrected. Yet p-consciousness is nonetheless transformed.[25]

Once again, contrast this with Rosenthal's wine and oboe examples: were occurent, assertoric, noninferential thoughts playing a significant explanatory role in the generation of p-consciousness, one might reasonably expect conceptual addition or subtraction to be leading the way. But here the phenomenon —the illusion that the hand has been resurrected—trumps. At best, if there are accompanying HOTs, they seem to just follow the illusion, no matter where it leads.[26]

Although it might be possible for the HOT theory to accommodate such phenomena, still these cases differ strikingly from Rosenthal's paradigmatic cases. And those paradigm cases are what provide the HOT theory with much of its initial plausibility. They suggest a significant explanatory role for HOTs. In our cases, however, it seems difficult, or at least pointless, to leave explanatory room for HOTs.

On our view, in light of the cases described above, Rosenthal would seem to be hard pressed to continue to proclaim that HOT is a promising *empirical* theory. The point is not that good theories don't encounter anomalies; they do. The point is that HOT has not as yet established any significant empirical credentials. Examples like that of wine tasting are suggestive, but suggestive evidence of this sort is a far cry from the sort of evidence that growing empirical theories require. Such theories require the evidence provided by confirmed predictions or evidence provided when attempts to falsify fail. Given that existing evidence is no more than weakly suggestive, these anomalous cases would seem to constitute a serious threat to all attempts to invoke HOTs as part of the explanans for any given instance of p-consciousness.

In the next section, we consider some further possible responses to our criticisms of HOT theory. We argue that none of them succeed in alleviating the worries raised by the problem of radical confabulation.

3.

In this section, the following objections are considered: (1) the confabulatory cases can be reinterpreted; (2) the counterexamples exaggerate the relevance of peripheral or input mechanisms; (3) HOTs might be targeting something mental, even if that something is not a first-order mental state; (4) p-consciousness might fail to occur in the proposed counterexamples; and, (5) Rosenthal could opt for a disjunctive strategy, one that allows for the possibility that HOTs are sufficient. We argue that each of these fails.

(1) Reinterpreting cases of confabulation? A first response that might be attempted on behalf of Rosenthal is to try to reinterpret some cases of confabulation. For example, as for dental fear, a case that we treat as problematic, a defender of the HOT theory might try to argue that it is significantly analogous to standard instances of misrepresentation, like the manner in which similarity of shape, impaired vision, and distance might cause one to mistake a cow for a horse. Fear and vibration—or perhaps a philosopher should write "fear and trembling"—might be relevantly analogous to shape and impaired vision.

But even if this attempt succeeds, it still leaves us to wonder on just what grounds this can be called an account of p-

consciousness. For the less contentious cases, like parafoveal vision, at least there is a relevant, substantial, first-order sensory state with which to work. Perhaps, as we have suggested above, some similar account could be developed for dental fear: the relevant HOTs in this instance would actually be about fear and vibration, but these sensations would then be misnamed. Because fear cum vibration is unpleasant, just as is pain, such misnaming might be as common as is the misnaming of similar colors or sounds. We submit that dental patients, generally speaking, are not proficient at naming the different types of experienced unpleasantness that they undergo, so they tend to use a catchall term like "pain."

AS accompanied by hallucinations, however, presents a much more substantial worry by eliminating even the possibility of an analogue to shape, impaired vision, misnaming, and so forth. In some respects it seems that Rosenthal has not yet put to rest the "problem of the rock." Our concern is not that "beliefs" don't seem to be the sort of entity that could explain consciousness; actually, we suspect that worries about explanatory gaps and hard problems are excessively motivated by the desire to "feel" that one understands, and this is an *in*appropriate expectation to hold concerning empirical explanations. As Hempel (1965, 256–58) once observed in a similar context, the desire to feel that we "understand" is just to confuse "empathetic familiarity" with theory-based or cognitive understanding.

Doxastic states may well have all that is required to cause what-it-is-like, but once we leave behind paradigmatic cases like wine tasting and oboe listening—cases that imply a two-component view—it is not at all clear that such states are able to play an explanatory role. Although at first glance it might appear that AS accompanied by hallucinations might lend support to a one-component view, such is not necessarily the case. One can describe the phenomenon in terms consistent with the view that HOTs alone are sufficient for what-it-is-like, but one can just as easily describe it in terms compatible with the view that HOTs are either irrelevant or an outcome of events over which they yield no control—for example, despite the belief that one can't see, one suddenly starts to see, an effect that sets the confabulatory machinery of HOTs into motion. Concerning the relationship between conceptual consistency and explanatory adequacy, we will have more to say below.

Part of the confusion concerning the explanatory role of HOTs is, we submit, that in his more recent discussions of HOT Rosenthal tends to minimize talk of causal connections,[27] preferring to use the term "accompany." But if "accompany" is interpreted as "correlation," something that is not uncommon for Rosenthal (e.g., 2002, 413), then it is not at all clear just what kind of explanation can be forthcoming. Accidental correla-

tions are cheap; hence, explanatorily useless. They are the type of thing that statistics faculty like to entertain their undergraduates with, for example, as with citing instances of the correlation between stock market trends and women's skirt lengths. Unfortunately, Rosenthal has not yet clearly indicated just what kind of *non*accidental correlations are involved here,[28] and that is what we would need were "accompany" to play an explanatory role.

More substantial interpretations may well be forthcoming, but at least at this stage in the development of the HOT theory we have good grounds for asking in what sense it could be that HOTs "accompany" when either they wholly misrepresent what they are apparently about, or when they target nonexistents.[29] What is, at minimum, required is a more systematic account of "accompany," either when we should expect the two components to correlate, and how or when we should expect that they won't, and what consequences will follow. Rosenthal doesn't intend that "accompany" should apply to cases wherein HOTs intentionally target nonexistents,[30] but the worry here, for those of us concerned with explanation that is applicable to the empirical sciences, is that "accompany" can be dismissed too easily, in an ad hoc fashion. And if this is the case, then Rosenthal's attempt to avoid conceptual issues pertaining to the hard problems and explanatory gaps fails because his avoidance is justified by the claim that he is concerned with scientific explanation, but scientific explanation can't get off the ground if it allows for the ad hoc and relies upon the post hoc.

(2) Exaggerated emphasis on peripheral or input mechanisms? A defender of HOT might also proclaim that our cases of radical confabulation only seem to challenge HOT because we focus unduly on sensory input mechanisms. The argument might go that people can be blind or deaf in different ways. Perhaps it could be said that although they have lost sensory contact with the world, their internal sensory capabilities remain intact. Perhaps it could be argued that these internal sensory mechanisms are mental states that can serve as appropriate HOT targets.

To take the cases of AS as examples, perhaps it could be said that patients are just mistaking visual imagery for actual vision. It might then be open to a defender of HOT theory to claim that visual imagery provides the target for a HOT. But most clinical neurologists believe this is not the case because the damaged areas are precisely those areas that are needed for both imagery and actual vision.[31] Still, clinical cases vary greatly, so perhaps the general point does need to be addressed.

Rosenthal says, "To be qualitative, a property need not always occur consciously; it must simply be able to occur consciously. Nonmental, physiological properties, by contrast,

are never conscious" (2005, 177). Rosenthal defines sensory states as possessing properties that enable us to make distinctions and form classifications, states that have the potential to be conscious. "A state's having qualitative character ... is solely a matter of the role that state plays in perceiving" (Rosenthal 2205, 13). The first thing to observe here is that even if certain nonconscious, neurophysiological states obtain independently of sensory contact with the world, since they are playing no perceptual role, it is by no means obvious that those states can count as mental, something that would be required on Rosenthal's view.[32] What's more, identifying something as mental because it has the potential to be conscious seems to risk flirtation with higher-order dispositional views,[33] views that Rosenthal rejects.[34]

A third thing to observe is that, according to Rosenthal,[35] if relevant perceptual roles never existed, we would have no grounds for the ascription of mental qualities. It would then seem to follow that at least for the congenitally blind victims of AS, we would lack adequate grounds for the ascription of relevant, nonconscious, sensory states. Homomorphism theory doesn't require the persistent function of perceptual roles, but it does require that those roles were previously performed.

A fourth thing to observe is that once sensory contact with the world has been lost, the sense in which it can be said that we make distinctions and form classifications is less clear. During dream states, for example, when sensory input is blocked, although internal perception proceeds and emotions might be enhanced, cognitive functions are greatly altered (Hobson 2007, 106–7). Perhaps it could still be argued that first-order sensory distinctions and classifications are still being made, but when one hallucinates (as in sensory deprivation) or when one dreams, since cognitive functions are altered, it is not clear just what role HOTs are able to play in the generation of conscious states. At the very least we would need further discussion concerning the nature and role of HOTs when the senses are deprived or when the agent is asleep. Even worse for the HOT theorist, if internal perception is unimpaired but cognition is distorted, it might be that Rosenthal and others would be pushed toward a higher-order *perception*, or inner sense, view of consciousness. And that is a view that Rosenthal (2005, 340) adamantly rejects.

(3) Might HOTs still be targeting something mental, even if there is no first-order state? After all, on Rosenthal's account, a state's being conscious consists in one's being conscious of self as being in that state. Rosenthal writes: "Each HOT represents its target state as belonging to some individual" (2005, 347). Perhaps then p-consciousness could still be said to result from the targeting of something mental. But instead of a state serving as the target, a subject would so serve.

But even were we to grant that either state or subject could suffice, it is not clear that the theory can stand, at least not on Rosenthal's terms. Concerning the central notion of self he writes: "A minimal concept of self will suffice for reference to oneself; no more is needed than a concept that allows distinguishing between oneself and other things. Such a minimal concept need not specify what sort of thing the self is. Thus it need not imply that the self has some special sort of unity, or is a center of consciousness, or is transparent, or *even that it has mental properties*" (1997, 741; emphasis added). Rosenthal also writes that "the single subject which we're conscious of our conscious states as belonging to may not actually exist" (2005, 348). In other words, according to HOT, neither targeted state nor subject need actually exist.

If there is no target sensory state, nor any self with mental properties, then just what is left to "accompany" the HOT? We know that a rock would be insufficient. But when we consider Rosenthal's minimalist, even eliminativist, understanding of self, it isn't clear that there is enough substance here to enable p-consciousness, that is, unless the HOT really is doing all the work, unless HOTs really are sufficient for what-it-is-like.

Alternatively, he might try to incorporate a more robust sense of self. But doubtless such a response would resurrect old worries about HOT, particularly those raised by Dretske (1995) and others who are concerned that HOT theory excludes prelinguistic infants, feral children, and nonhuman animals from the community of creatures capable of phenomenal consciousness. Even if Rosenthal were to bite this bullet, it would by no means be an easy task to formulate a theoretically feasible concept of self that has the resources to play the role that he hopes for within HOT theory.

(4) Might it be that HOTs are sufficient because p-consciousness doesn't occur in the alleged counterexamples? Rosenthal might concede that for cases of radical confabulation the first-order targets are indeed missing, while at the same time denying that p-consciousness occurs. He could claim that subjects' first-person reports about p-consciousness reflect nothing but epistemic overconfidence.

But, as we observed above, the problem here would be that Rosenthal would have to deny the neuroscience and clinical standards for interpreting patient behavior (including verbal behavior) as indicating the actual experience of p-consciousness. This might be possible, but the burden for rejection of standard empirical characterizations rests with proponents of the new explanatory theory; in this instance, with the proponents of HOT. Presumably, were this to succeed, it would require a significant deflationary view of p-consciousness, a view that some others have already found appealing.

Especially noteworthy in this regard is Dennett (e.g., 1991, 369–411), who has invested much effort toward deflating standard characterizations of p-consciousness. But, despite the many similarities between their two positions, Dennett (e.g., 1991, 362–68) believes intentional states alone (understood as beliefs, or interpretations, or as memeplexes) are sufficient to explain consciousness. He sees no need to posit two distinct mental levels; even when he turns his attention to consciousness, he takes beliefs to be the "pretheoretical data, the *quod erat explicatum*" (2005, 44–45).

Were Rosenthal to recast HOT in this form, he might be better positioned to seek an explanatory scope that does not omit cases of radical confabulation.[36] Simplifying somewhat, if HOTs carried all of the load, all of the time, then there would be no need to mark distinctions between radical and nonradical, no need to mark distinctions between mere epistemic overconfidence and cases in which robust p-consciousness actually occurs. However, moving in that direction would mean giving up any role for first-order mental states, which would imply that we are never really in certain folk-psychological states that we take ourselves to be in; it would commit Rosenthal to denial of a distinction between the appearance and reality of consciousness. Obviously, such moves would mark a substantial departure from Rosenthal's current position (e.g., 2005, 231–56). Still, the advantage of such moves would be that Rosenthal would not be troubled by differences between ordinary instances wherein HOTs make self aware of being in a certain state and instances of radical confabulation. On Dennett's view, there simply is no substantial difference—in a sense, all cases are cases of radical confabulation because we are never in the states that we take ourselves to be in.

But a move in this direction would also be rife with problems. Recall, for example, the manner in which a simple trick with mirrors can lead to the "resurrection" of a phantom hand. What might a theory that puts all explanatory weight onto HOTs be able to say about this? Patients don't report occurent beliefs that their hands are resurrected. Quite the opposite. Nevertheless, the therapy works. It seems that Rosenthal (or Dennett) would have to posit a nonconscious belief in the hand's resurrection, a belief that conflicts with the conscious belief that no such thing is possible. We won't pursue this matter any further here, beyond just to say that Rosenthal would owe us a principled account of how conflicting HOTs are, or can be, adjudicated. Although according to the HOT theory, most HOTs are nonconscious, still the theory does allow for HOTs to be conscious, when they themselves are targeted by yet higher-order thoughts. Accordingly, what we need is an account that explains why the nonconscious HOT—which is nothing more than a theoretical posit—wields more influence than the con-

scious, reportable HOT—which is something that we have theory-independent evidence for.

One additional, related problem that we will mention, but reserve full development for another time, is what we term the problem of "conceptual inadequacy." Consider the case of people who suffer from pain asymbolia syndrome: although they claim to be experiencing pain, they say they don't care, that they seem to "float above" the pain.[37] Our worry, baldly stated, is that no society's folk psychology contains the concept of pains that one can "float above." For those people who know nothing of pain studies, who suddenly find themselves afflicted with pain asymbolia, the phenomenon just asserts itself without warning. These people wouldn't have the conceptual resources with which to form the relevant HOT. They are dramatically unlike people who cultivate an appreciation of wine, color, or music. The worry then for those who would adopt a move in this direction—that is, placing yet greater explanatory weight on beliefs or on the HOT—is that in cases like pain asymbolia, the higher-order thought seems irrelevant, for it lacks the conceptual resources whereby such a distinction could be made. Folk psychology doesn't distinguish between pains that we must wallow in and pains that we can float above.

(5) Finally, one might defend the HOT theory by suggesting a *disjunctive* strategy. Earlier, in section 1, we noted that Rosenthal's theory is open to a one-component interpretation, according to which, as Rosenthal says, "there can be something it's like for one to be in a state with particular mental qualities, even if no such state occurs" (2005, 211). On this interpretation, contrary to the two-component view, the HOTs alone are taken to be sufficient for p-consciousness. A disjunctive strategy would allow Rosenthal to employ a two-component version to account for the ordinary cases and nonradical confabulation cases, along with a one-component version to account for radical confabulation cases. The idea is that, on the one hand, Rosenthal can agree that in radical cases there are no first-order mental states, since the one-component view allows HOTs to play a sufficient role in explaining p-consciousness; on the other hand, he can maintain the two-component view for the rest of the cases.

The first problem is that this disjunctive strategy seems ad hoc. Rosenthal's (e.g., 2005, 187–88) paradigm cases—for example, the wine taster who becomes conscious of fine-grained differences among sensory qualities by learning new concepts—are all bound to the two-component version. Even the cases of misrepresentation that he cites, as we have argued above, are bound to the two-component version: recall, typically misrepresentations are extrapolations from first-order sensory states. So these cases do not motivate the strategy.

Moreover, the empirical scientists who invoke Rosenthal's HOT theory in constructing explanations also work with a two-

component view. Weiskrantz (1997, 71–76; 1997, 203–4), for example, in proposing explanations of various neuropsychological deficits, invokes HOT and sees it as largely consistent with his views on conscious experiences, but Weiskrantz does not treat HOTs as sufficient; instead, he sees HOTs—or something similar to HOTs—as being only necessary for the sort of conscious experience that is lacking in cases of blindsight, unilateral neglect, amnesia, aphasia, and so forth. Likewise for the research of Dienes and Perner (1999, 2001, and 2002) who, despite noting that Rosenthal claims HOTs can be sufficient, only apply his analysis to cases in which first-order states actually obtain. And Rolls (2001, 245–65) develops a view dubbed HOLT (higher-order *linguistic* thought), a view that differs from Rosenthal only in placing greater emphasis on syntactic manipulation. But nothing in Rolls's discussion or application suggests that he regards HOLTs as sufficient; his hypothesis "is that consciousness is the state which arises by virtue of having the ability to think about one's own thoughts, which has the adaptive advantage of enabling one to correct long multistep syntactic plans" (2001, 258). This is a version of the two-component view. He nowhere considers the possibility of thinking about nonexistent thoughts, nor does he construct explanations of this sort; instead, his focus is clearly on the manner in which the higher-order and lower-order interact.

Rosenthal (2005, 179) also reports on the research of Frith and Frith (1999), research that includes reference to brain imaging studies of subjects asked to report on their mental states. Frith and Frith record that, despite "wide variation in the nature of the states reported on, activity was observed in all these studies ... along the border between rostral anterior cingulated cortex and medial prefrontal cortex" (1999, p. 1693). Subjects were asked to monitor and report on pain, emotions, spontaneous thoughts, actions, and tickling. Rosenthal speculates that the brain area activated in all these instances might be that which subserves HOTs. But even if this is true, it does not lend support to the sufficiency claim because the Frith and Frith data all concern monitoring of actual first-order states.

What these empirical cases suggest is that a disjunctive strategy would be ad hoc. Paradigm cases—cases used to motivate serious consideration of HOT that actually offer some explanatory, predictive leverage—invoke two components. Even the cases of nonradical misrepresentation are best explained by a two-component view. What's more, the empirical scientists who invoke HOTs in their explanations, or whose research is regarded by Rosenthal as being suggestive of the role that HOTs can play in explanation, are working with a two-component view. The one-component view has not been shown to be empirically well motivated; therefore, the disjunctive strategy seems ad hoc.

A second problem, one intimately related to the ad hoc problem, is that the disjunctive strategy does not obviously provide any explanatory or predictive advantage. Although on the surface it may seem that AS accompanied by hallucination or schizophrenia in the congenitally deaf can be explained by the one-component version of HOT theory, perhaps reminiscent of the way in which Newtonian theory predicted the existence of Neptune and Pluto, such an interpretation would be misleading. The one-component version—markedly unlike the case with Newton—leaves absolutely no specific expectations concerning when targetless HOTs that enable p-consciousness should occur.[38] Wishful thinking, self-deception, and other standard cases of confabulation—that is, typical causes of misrepresentation (e.g., Rosenthal 2005, 125–26)—provide no reason for expecting AS accompanied by hallucinations or schizophrenic voices in the congenitally deaf. Most people who are blind know that they are blind and they have no p-conscious visual experiences to suggest otherwise; standard forms of wishful thinking or self-deception provide no relief from their blindness, although delirium tremens might. The same can be said for the congenitally deaf: neither wishful thinking nor self-deception help them to hear, only schizophrenia does. Accordingly, not even our examples can be taken to suggest that HOTs play a sufficient explanatory role.

In other words, the most that could be claimed on behalf of the disjunctive strategy when applied to our cases is that the one-component view is conceptually consistent with them. But conceptual consistency can be purchased cheaply. The failure to indicate just when we should expect wishful thinking, self-deception, or some other factor to give rise to what-it-is-like indicates that the one-component view does not obviously contribute anything substantial to the explanatory status of the HOT theory.

A third problem is that the disjunctive strategy seems not to be falsifiable. A virtue, albeit an ironic one, of the two-component view is that it is falsifiable. We say "ironic" because, if we are correct, it now stands as falsified by our counterexamples.[39] Employment of the disjunctive strategy, even if ad hoc in the sense of being unmotivated by paradigm cases or empirical research and even if it provides no obvious grounds for expecting the phenomena that occur in the counterexamples, can at least contribute to the development of an explanatory theory, if it is cast in such a form as to be potentially falsifiable. But the one-component disjunct of the disjunctive strategy is not so cast. What we need, and do not yet have, is a reasonably clear statement of the conditions under which the disjunctive strategy could be empirically challenged.

As we have said above, critics (e.g., Kriegel 2007) have found some of Rosenthal's claims to be counterintuitive, especially his

claim that people can be under the impression of being in conscious states that they are not actually in. Importantly, Rosenthal agrees that such claims are counterintuitive. Because he cannot use intuition to justify these claims, he must justify them in some other way. The way he chooses is justification in terms of a theory that he believes to have sound empirical, explanatory credentials. And this is a way that he needs in order to fend off worries pertaining to the alleged hard problem and the explanatory gap. But we are here showing that the HOT theory is not justified in claiming such credentials for itself. Therefore, the HOT theory is confronted with a serious challenge.

In sum, although it might seem that a disjunctive strategy would enhance the HOT theory's explanatory adequacy, such is not the case. Explanatory adequacy is not enhanced when the added option—in this case, the one-component option—is not empirically motivated by the paradigm cases or by the way in which working scientists actually employ the theory. Ad hoc additions don't help. Second, although the addition might well be conceptually consistent with our cases of radical confabulation, mere consistency doesn't aid explanation. Third, the two-component view has the virtue of falsifiability; the one-component view, in its present form, does not. Obviously we have treated just lightly some difficult issues in the philosophy of science, but our intent is simply to show that the disjunctive strategy is suspect. To remove the shadow of suspicion from this strategy, at least if Rosenthal wishes to continue to promote the HOT theory as an explanatory theory for empirical science, he must accept the burden of showing that the one-component version is not ad hoc, that it provides explanatory advantage, or that it is falsifiable.

4.

We conclude that radical confabulation presents a serious problem for Rosenthal's HOT theory. There seems to be an important body of anomalous cases for which the HOT theory—either the two-component version or the disjunctive strategy—fails to provide any explanatory advantage. We wish to emphasize that our intent is not destructive. We share an admiration for the effort that is expended in rising to the challenge of consciousness and seeking to devise an empirical theory that aims to handle even those aspects that are often treated as being beyond the reach of science. Our task, as we approach it here, is to examine the explanatory adequacy of the proposed theory. Our hope is that the discovery of anomalies leads to fruitful revision, the design of a more adequate theory of broader scope, enhanced consistency, and greater explanatory power.[40]

Notes

The order of authorship was determined arbitrarily; this manuscript is completely collaborative.

[1] In this paper we focus on Rosenthal's version of HOT theory. For other versions of HOT theories of consciousness, see Rocco Gennaro 2004.

[2] That we are not conscious of any mediating factors should not be taken to imply that there are no mediating factors.

[3] Natural language is nonessential to the formation of HOTs; hence, prelinguistic infants, feral children, and nonhuman animals are in principle capable of conscious experience. See Rosenthal 1991, 472–73, and 1997, 741–42.

[4] Rosenthal does though make an important distinction here: the claim is that some mental states, those that exhibit "thin phenomenality," have qualitative properties, even though there is nothing that it is like for one to have said properties. Mental states that exhibit "thick phenomenality," by contrast, are states that have qualitative properties along with a what-it-is-like for one to have the experience (e.g., 2005, 190–92). Subliminal perception and the cocktail party effect are offered as illustrations of how, empirically, qualitative states can be independent of what-it-is-like. Below, whenever we use p-consciousness, we intend that it be understood in Rosenthal's sense of "thick phenomenality."

[5] Rosenthal does not endorse this characterization of his position; we further explain our position below.

[6] We don't intend to belittle the significance of conceptual issues, but they trouble *all* scientific theories, and those theories—quantum mechanics, evolution by natural selection, and so on—continue to generate fruitful inquiry in the domains for which they were designed.

[7] Prominent among such investigators have been Rolls (2001, 244–65), Weiskrantz (1997, 71–76), and Dienes and Perner (e.g., 1999). We will return to consideration of their work below.

[8] Of states that possess the property of phenomenal consciousness, it is often said that it is *like something* to have them (Farrell 1950; Sprigge 1971, 167–68; Nagel 1974). Notational variants of what-it-is-like-to-be include raw feels, qualia, experience, and subjective feel. Many hold that this property can only be defined ostensively, such that it can only be pointed to in experience, as when you explain "sting" by saying how your hands feel when you hit a fastball off the handle or the end of the bat on a cold day. As a more formal alternative to ostension, Carruthers defines phenomenal consciousness as "events that we can recognize in ourselves, non-inferentially, or 'straight off,' in virtue of the ways in which they feel to us, or the ways in which they present themselves to us subjectively" (2000, 14).

[9] To take just one example among these, pain, a property that is typically regarded as being the most highly accessible to a subject, it is striking just how many investigators of consciousness, representing varied theoretical perspectives and areas of specialization, countenance talk of unconscious, unnoticed, unexperienced, unfelt, or sub-clinical pains (e.g., Carruthers 2005, 185–86; Chalmers 1996, 17; Dartnall 2001; Guttenplan 2000, 28; Searle 1992, 164–67; Tye 1995, 115 and 2000, 182; Vertosick 2000, 152 and 175; Wilkes 1993, 186). And though it is often said that these are incoherent or counterintuitive ways of

thinking about pain, Lycan observes that "ordinary people quite frequently speak of pains that go unfelt, without any sense of contradiction" (2003, 9 and 13). To convince skeptics, he has even begun to compile examples of unfelt pain from the popular press.

[10] This example is frequently cited by Dennett (e.g., 1991, 61; 1996, 13 and 95; and 1998, 351).

[11] According to Rosenthal's "homomorphism theory," "mental qualities are properties of states that figure in the perceptual functioning of a particular sensory modality, and whose similarities and differences are homomorphic to those which hold among the properties perceptible by that modality." Rosenthal claims that the homomorphism theory provides the theoretical grounds for distin-guishing among the conscious, the nonconscious but mental, and the merely physical.

[12] Dennett's (1991, 30–31, 337) observations concerning palate training for wine tasters, aural training for musicians, and the experiences of an apprentice piano tuner mirror Rosenthal's views. Rosenthal's and Dennett's theories converge in many respects. Concerning their apparent differences see Rosenthal 2005, 321–35 and Dennett 2000.

[13] In both personal communication and in recent publications (e.g., 2005, 209–13), Rosenthal has explicitly pressed this point.

[14] Rosenthal has emphasized this point in personal communication.

[15] Those of us who live in earthquake-prone regions are perhaps more easily persuaded by this claim than those who do not. We have grown accustomed to not infrequent vibrations accompanied by fear, and the two combined are indeed painless. To those who don't live in earthquake-prone regions, perhaps a ride on a roller coaster would suffice to persuade you.

[16] The same is true for well-studied cases of the confabulation of intention (Wegner 2002, 171–86); after the completion of actions, people commonly revise what they think they intended, and there is no necessary reference to p-consciousness.

[17] One of the explanatory challenges that HOT seems, as yet, unable to respond to is the need to distinguish between distinct explananda: beliefs that merely make reference to what-it-is-like and beliefs that have an actual what-it-is-like aspect. More will be said about this below.

[18] Cf. also Rosenthal 2005, 211, where he makes the same point concerning dental fear.

[19] In personal communication, Rosenthal has conceded that matters might not be so plain here. Again, we take this matter to be significant because, among other things, it seems to suggest that the HOT theory carries more explanatory weight when targets exist, i.e., when there are two components. And we believe that radical cases of confabulation show that the two-component view fails.

[20] Also known as the Anton-Babinski Syndrome.

[21] Our discussion of hallucinations and pseudohallucinations in this and in the next paragraph is based on Davis 2004.

[22] We don't concede this point; we simply choose not to press it here.

[23] The patient in question did recognize some of his hallucinations *as* hallucinations. That a hallucination is recognized as such by the subject does not imply that it is inconsistent with the three criteria for

hallucinations indicated in the text.

[24] At one point, Rosenthal entertained the possibility that targetless HOTs would be both rare and pathological. He has since retreated from that position (2005, 29). Below we try to show that he was correct to revise the claim that such cases are rare.

[25] Yet more difficult cases could easily be cited here, e.g., the case of color-blind color vision among synesthetes who see numbers as tinged with hues (Ramachandran and Hubbard 2003, 57). But each case involves distinctive complexities, and our only intention at this point is to suggest that there is a family of cases that seriously challenge the explanatory efficacy of HOT theory.

[26] If we are describing this phenomenon accurately, this case too seems to lend support to a two-component view, albeit a version wherein HOTs are passively compliant to the demands of first-order sensory states.

[27] Rosenthal identifies this transition of his ideas in several places, e.g., 2005, 56n25.

[28] Rosenthal does propose an evolutionary account of how it came to be that HOTs tend to accurately reflect actual first-order states (2005, 15–16, 218–19, 303–5). He argues that HOTs evolved as a means of attributing mental states to others, attributions that are based on behavioral observations and other stimuli; this talent for the attribution of mental states was then turned inward. But this argument doesn't help his case here because it provides no reason to expect HOTs that would be involved in enabling the deaf to hear or the blind to see. Evolved capacities can fail to be adaptive, because the environment or informational inputs change (Buller 2005, 57). But on the HOT theory we have no reason to believe that either has changed in such a way as to cause radical confabulation.

[29] It is worth mentioning that there is a significant tension between responding to the problem of the rock by claiming that HOTs target mental states while simultaneously claiming that HOTs can target nonexistents. How could targeting a nonexistent give rise to what-it-is-like when targeting a rock cannot?

[30] This is a point that he has emphasized in personal communication.

[31] Our discussion here is drawn from Churchland 2002, 122–23.

[32] In other words, we are saying that Rosenthal has not yet succeeded in putting to rest the "problem of the rock."

[33] For a defense of one version of the dispositional view, see Carruthers 2000 and 2005.

[34] Rosenthal concedes that when mental states are conscious, "it seems phenomenologically that there is no HOT present" (2005, 111). But he emphasizes that the appearance is irrelevant, "because HOTs are posited as the best explanation of what it is for a mental state to be a conscious state." And this is yet another reaffirmation that the case for—or against—HOT needs to be made in terms of explanatory adequacy.

[35] Personal communication.

[36] It should be noted though that Dennett's resistance to talk of mental levels might reflect more a difference in terminology than in substance (e.g., Rockwell 1996).

[37] Here we are quoting Carruthers (2000, 206).

[38] The point is not to deny the occurrence of improbable events;

rather, it is merely to say that when offering an explanation, some reason for expecting the improbable event should be stated, perhaps, for example, by describing relevant causal mechanisms (e.g., Railton 1993).

[39] We don't intend to deny Duhemian and other concerns pertaining to falsification. We appeal to it here merely as an (admittedly imperfect) constraint on the development of explanatory theories.

[40] We hereby express our heartfelt gratitude to David Rosenthal for his careful reading of and detailed comments on previous drafts of this paper. We also wish to express our gratitude to anonymous referees who motivated us to make apprpriate revisions. Moreover, we are grateful to participants in the 2007 LMPS Taipei Conference sponsored by the National Taiwan University (April 27–28) and to participants in the Soochow University Conference on Analytic Philosophy (June 20–23, 2007) for their many helpful comments. Especially worthy of note are several constructive suggestions offered by Richard Fumerton, Eric Peng, Wen-fang Wang, and Huei-Ying (Tony) Cheng. We have not agreed with or been persuaded by every comment or suggestion, but our understanding of the subject matter has been greatly enriched through these stimulating exchanges.

References

Atkinson, Joanna R. 2006. The perceptual characteristics of voice-hallucinations in deaf people: Insights into the nature of subvocal thought and sensory feedback loops. *Schizophrenia Bulletin* 32 (4): 701–8.

Buller, David J. 2005. *Adapting minds: Evolutionary psychology and the persistent quest for human nature.* Cambridge, MA: MIT Press.

Byrne, Alex. 1997. Some like it hot: Consciousness and higher-order thoughts. *Philosophical Studies* 86:103–29.

Carruthers, Peter. 2000. *Phenomenal consciousness: A naturalistic theory.* New York: Cambridge University Press.

———. 2005. *Consciousness: Essays from a higher-order perspective.* New York: Oxford University Press.

Chalmers, David J. 1996. *The conscious mind: In search of a fundamental theory.* New York: Oxford University Press.

Churchland, Patricia Smith. 2002. *Brain-wise: Studies in neurophilosophy*. Cambridge, MA: MIT Press.

Dartnall, Terry. 2001. The pain problem. *Philosophical Psychology* 14 (1): 95–102.

Davis, Derek Russell. 2004. Hallucination. In *The Oxford companion to the mind*, ed. Richard L. Gregory. 2nd ed. Oxford: Oxford University Press.

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

———. 1996. *Kinds of minds: Toward an understanding of consciousness.* New York: Basic Books.

———. 1998. *Brainchildren: Essays on designing minds.* Cambridge, MA: MIT Press.

———. 2000. With a little help from my friends. In *Dennett's philosophy: A comprehensive assessment*, ed. Don Ross, Andrew Brook, and David Thompson. Cambridge, MA: MIT Press.

Dennett, Daniel. 2005. *Sweet dreams: Philosophical obstacles to a science of consciousness.* Cambridge, MA: MIT Press.
Dienes, Zoltran, and Josef Perner. 1999. A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences* 22:735–55.
———. 2001. When knowledge is unconscious because of conscious knowledge and vice versa. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*, ed. J. D. Moore and K. Stenning. Mahwah, NJ: Lawrence Erlbaum Associates.
———. 2002. Developmental aspects of consciousness: How much theory of mind do you need to be consciously aware? *Consciousness and Cognition* 30.
Dretske, Fred. 1995. *Naturalizing the mind.* Cambridge, MA: MIT Press.
Farrell, B. A. 1950. Experience. *Mind* 59:170–98.
Frith, Chris D., and Uta Frith. 1999. Interacting minds—A biological basis. *Science* 286:1692–95.
Gennaro, Rocco, ed. 2004. *Higher-order theories of consciousness.* Philadelphia: John Benjamins Publishing Company.
———. 2005. The HOT theory of consciousness: Between a rock and a hard place? *Journal of Consciousness Studies* 12 (2): 3–21.
Goldman, Alvin. 1993. Consciousness, folk psychology, and cognitive science. *Consciousness and Cognition* 2:364–82.
Guttenplan, Samuel. 2000. *Mind's landscape.* Oxford: Blackwell.
Heilman, Kenneth M. 1991. Anosognosia: Possible neuropsychological mechanisms. In *Awareness of deficit after brain injury: Clinical and theoretical issues*, ed. George Prigatano, 53–62. New York: Oxford University Press.
Hempel, Carl G. 1965. *Aspects of scientific explanation.* New York: Free Press.
Hirstein, William. 2005. *Brain fiction: Self-deception and the riddle of confabulation.* Cambridge, MA: MIT Press.
Hobson, J. Allen. 2007. Normal and abnormal states of consciousness. In *The Blackwell companion to consciousness*, ed. Max Velmans and Susan Schneider. Malden, MA: Blackwell.
Janzen, Greg. Forthcoming. Phenomenal character as implicit self-awareness. *Journal of Consciousness Studies.*
Kriegel, Uriah. 2003. Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy* 33:103–32.
———. 2007. Philosophical theories of consciousness: Contemporary Western perspectives. In *Cambridge handbook of consciousness*, ed. M. Mosocovitch, E. Thompson, and P. O. Zelazo. New York: Cambridge University Press.
Lane, Tim. 2006. T-Belief. Paper presented at the Conference on Naturalized Epistemology and Philosophy of Science at Soochow University, May 31–June 3.
Levine, Joseph. 2001. *Purple haze: The puzzle of consciousness.* New York: Oxford Press.
Lycan, William G. 2003. Dretske's ways of introspecting. In *Privileged access: Philosophical accounts of self-knowledge*, ed. Brie Gertler. New York: Ashgate.
Nagel, Thomas. 1974. What is it like to be a bat? *Mortal Questions.* Cambridge University Press.
Neander, K. 1998. The division of phenomenal labor: A problem for

representational theories of consciousness. In *Language, mind, and ontology*, ed. James Tomberlin, 411–34. Oxford: Blackwell.
Nisbett, R. E., and T. D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231–59.
Railton, Peter. 1993. Probability, explanation, and information. In *Explanation*, ed. David-Hillel Ruben. New York: Oxford University Press.
Ramachandran, V. S. 1998. Consciousness and body image: Lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Philosophical Transactions of the Royal Society of London* 353:1851–59.
Ramachandran, V. S., and Edward M. Hubbard. 2003. Hearing colors, tasting shapes. *Scientific American* 5:53–59.
Ramachandran, V. S., and D. Rogers-Ramachandran. 1996. Synesthesia in phantom limbs induced with mirrors. *Proceedings of the Royal Society of London* 263:377–86.
Rockwell, Teed. 1996. Awareness, mental phenomena, and consciousness. *Journal of Consciousness Studies* 3:463–76.
Rolls, Edmund. 2001. *The brain and emotion*. New York: Oxford University Press.
Rosenthal, David M. 1991. Two concepts of consciousness. In *The nature of mind*, ed. David Rosenthal. New York: Oxford University Press.
———. 1997. A theory of consciousness. In *The nature of consciousness*, ed. Ned Block, Owen Flanagan, and Guven Guzeldere. Cambridge, MA: MIT Press.
———. 2000. Metacognition and higher-order thoughts. *Consciousness and cognition* 9:231–42.
———. 2002. Explaining consciousness. In *Philosophy of mind: Classical and contemporary readings*, ed. David J. Chalmers, 406–21. New York: Oxford University Press.
———. 2004. Varieties of higher-order theory. In *Higher-order theories of consciousness*, ed. Rocco J. Gennaro, 17–44. Philadelphia: John Benjamins Publishing Company.
———. 2005. *Consciousness and mind*. Oxford: Oxford University Press.
Rowlands, Mark. 2001. Consciousness and higher-order thoughts. *Mind and Language* 16 (3): 290–310.
Seager, William. 1999. *Theories of consciousness: An introduction and assessment*. New York: Routledge.
Searle, John R. 1992. *The rediscovery of mind*. Cambridge, MA: MIT Press.
Sprigge, Timothy. 1971. Final causes. *Proceedings of the Aristotelian Society* 45:149–70.
Swartz, B. E., and J. C. Brust. 1984. Anton's Syndrome accompanying withdrawal hallucinosis in a blind alcoholic. *Neurology* 34:969–73.
Tye, Michael. 1995. *Ten problems of consciousness: A representational theory of phenomenal mind*. Cambridge, MA: MIT Press.
Tye, Michael. 2000. *Consciousness, color, and content*. Cambridge, MA: MIT Press.
Vertosick, Frank T. 2000. *Why we hurt: The natural history of pain*. New York: Harcourt.
Wegner, Daniel M. 2002. *The illusion of conscious will*. Cambridge, MA: MIT Press.
Weiskrantz, Larry. 1997. *Consciousness lost and found: A neuro-*

physiological exploration. Oxford University Press.
Wilkes, Kathleen V. 1993. *Real people: Personal identity without thought experiments*. New York: Oxford University Press.