Philos Stud (2011) 154:409–433
DOI 10.1007/s11098-010-9567-3

Misrepresenting consciousness

Josh Weisberg

Published online: 12 May 2010

Abstract An important objection to the "higher-order" theory of consciousness turns on the possibility of higher-order misrepresentation. I argue that the objection fails because it illicitly assumes a characterization of consciousness explicitly rejected by HO theory. This in turn raises the question of what justifies an initial characterization of the data a theory of consciousness must explain. I distinguish between *intrinsic* and *extrinsic* characterizations of consciousness, and I propose several desiderata a successful characterization of consciousness must meet. I then defend the particular extrinsic characterization of the HO theory, the "transitivity principle," against its intrinsic rivals, thereby showing that the misrepresentation objection conclusively falls short.

Keywords Consciousness · Higher-order thought · Misrepresentation · Empty higher-order thought

An essential first step in theorizing about consciousness is pinning down the data a theory must explain. In contemporary consciousness studies, two competing characterizations are widely used to fix the initial explanatory data. The first involves an *intrinsic concept* of consciousness, one tied to the experiential "feel" of a conscious state. The second involves an *extrinsic concept* of consciousness, one tied to the special connections a conscious state possesses or the special role it plays in a subject's mental life. These initial conceptual steps can have an enormous impact on the direction and success of a theory; mischaracterizing the data can lead to the illusion of intractable problems or to the false promise of easy reduction.

J. Weisberg (✉)
Department of Philosophy, University of Houston, 515 Agnes Arnold Hall, Houston, TX 77204, USA
e-mail: jweisberg@uh.edu

Using an intrinsic concept to characterize the data is the more popular theoretical starting point.[1] From this conceptual perspective, conscious states are defined by their experiential quality, by the *redness* of a conscious red experience or the *painfulness* of a conscious pain. These qualitative properties, known as "qualia," are conceived of as independent from causal, functional, or intentional features of the mind. Conscious states have qualia intrinsically, by virtue of their internal constitution and not as a result of any connection to other states, processes, or entities. By way of rough metaphor, conscious states "glow from within" with the light of qualia. An intrinsic conception takes seriously conceivability intuitions concerning the possibility of inverted or absent qualia. The apparent ease of imagining qualia inverts and zombies provides support for an intrinsic concept. Further, the intrinsic view gives pride of place to first-person reflection on the mind. If qualia seem intrinsic upon introspective reflection, then that's good reason to hold that they are indeed intrinsic, even in the face of contrary theoretical pressure. And while the view need not lead to dualism and mystery, it certainly "takes the phenomenal seriously" by regarding the challenge of explaining consciousness as a "hard" or "harder" problem, plagued with explanatory gaps and analytic and epistemological difficulties.[2] These problems may be surmountable, but embracing an intrinsic concept of consciousness at the outset of theorizing raises the explanatory bar high indeed.

An extrinsic concept of consciousness, on the other hand, characterizes conscious states in terms of their connections to the rest of the mind and to the world. Conscious states are not defined by their intrinsic qualities; rather, they are defined in terms of their functional role, broadly conceived. They may play a special role in cognition, broadcasting content to a variety of mental systems.[3] Or they may make the subject aware of certain aspects of the perceptual environment.[4] Or they may make the subject aware of her own mental states, providing a seemingly direct access to the mind.[5] Common to all these characterizations is the idea that consciousness is essentially a matter of connections. Something extrinsic to the state—the role it plays in the overall economy of the mind, the representational links it has to mind and world—makes it a conscious state. Accordingly, intuitions concerning inverted or absent qualia are downplayed: if the state has the right connections, it is conscious even if intuition urges otherwise. Further, an extrinsic concept eases the challenge of fitting consciousness into the scientific worldview. Extrinsic concepts of consciousness are generally amenable to functional characterization and so to functional reduction.[6] Likewise, representational characterizations of consciousness offer the prospect of explaining consciousness in terms of a

[1] Nagel (1974), McGinn (1989), Chalmers (1996), and Levine (2001).

[2] On the Hard Problem, see Chalmers (1996). On the Harder Problem, see Block (2002). See also Nagel (1986), McGinn (1999), and Levine (2001).

[3] Dennett (1991), Baars (1997), and Dehaene and Naccache (2001).

[4] Harman (1990), Dretske (1995), and Tye (1995).

[5] Armstrong (1980), Rosenthal (1986), and Lycan (1987).

[6] Lewis (1972); see also Jackson (1998).

naturalized theory of representation.[7] But proponents of an extrinsic view need not reject the importance of commonsense intuition and first-person reflection. They can point to a number of folk-psychological platitudes suggesting an extrinsic concept of consciousness. For example, conscious states are available for reasoning in ways nonconscious states are not, they are "transparently" about the world, or they are states we are suitably aware of ourselves as being in. These commonsense starting points allow proponents of the extrinsic conception to claim to have captured central elements of our pretheoretic notion of consciousness. This provides some defense against charges of subject changing and failing to "take the phenomenal seriously."

Still, it might be thought that an intrinsic concept more clearly captures what is unique about consciousness. Thomas Nagel's (1974) widely cited characterization of consciousness holds that when an organism is in a conscious state "there is something that it is like to *be* that organism—something it is like *for* the organism" (p. 519, emphasis in original). Nagel rejects any extrinsic concept of consciousness because, he contends, they are all fully compatible with the absence of consciousness.[8] His conception rests on the intuition that the true nature of consciousness is only accessible "from the inside" by similar kinds of creatures. Consciousness is in principle dissociable from the rest of the mind and world; knowing about connections will not reveal its nature. The prevalence of Nagelian "what it's like" language in the contemporary literature may suggest that an intrinsic concept best captures what folk and philosophers mean by 'consciousness.'

But, perhaps unsurprisingly, there is another way to interpret Nagel's phrase. There are two rival readings, corresponding to the two rival conceptions of consciousness. Supporters of an intrinsic concept focus on the "something" in the "something it is like for the organism." The "something" is held to be an intrinsic quality of experience, perhaps only knowable from the first-person perspective. We can label this interpretation the *qualitative reading* of Nagel's "something it's like" phrase.[9] Supporters of an extrinsic concept, on the other hand, focus on the "for the organism" in the "something it's like for the organism." This suggests a connection to the rest of the mind, a mode of access by a sentient subject. This connection might be cashed out in causal, functional, or representational terms, but it is the connection that matters, not the intrinsic nature of the states involved. We can call this interpretation the *subjective reading* of Nagel's "something it's like" phrase.[10,11]

[7] See, e.g. Dretske (1995).

[8] Nagel (1974, p. 519).

[9] Chalmers (1996), Block (1995), and Levine (2001).

[10] Lycan (1996) and Rosenthal (2002).

[11] An additional possibility is a "mixed" reading of Nagel's phrase. On this interpretation, the quality of conscious experience is intrinsic to the conscious state, but the state's being conscious as opposed to nonconscious is due to extrinsic factors. For the purposes of this paper, I will consider such a reading to fall into the intrinsic camp. What it's like for the subject in conscious experience is still determined by something intrinsic to the state—it's not a matter of connections. Further, it is unclear what the extrinsic connections contribute if they do not fix what it's like for the subject. What does it mean to say that extrinsic factors determine that there's something it's like for the subject, but not a determinate way that it's like? In any event, the key point is that there's a viable extrinsic reading of the "what it's like" phrase. I will address the plausibility of a mixed reading in more detail at the end of the paper.

Which concept of consciousness a theory employs is crucial to our evaluation of that theory. Failure to keep clear on this initial theoretical step can lead to confusion and question begging. This paper focuses on a particular example of just this sort of confusion. "Higher-order" (HO) theories of consciousness explicitly endorse an extrinsic concept of consciousness in fixing the data the theory aims to explain.[12] A well-known objection to the theory, *the problem of misrepresentation*,[13] loses sight of this crucial fact. I will argue that if we keep clear on just what the HO theory aims to explain, the objection fails. I'll argue that the prima facie pull of the objection is driven by (perhaps implicit) adoption of an intrinsic concept of consciousness. But since this is not the explanatory target of the theory, the objection fails. In response, one might take the pull of the objection as further reason to adopt an intrinsic concept. However, I will argue that the intrinsic concept is in need of additional support in this context and it is question-begging to assume it here. Further, given the great theoretical difficulties incurred by adopting an intrinsic concept, I'll contend that an extrinsic concept is better justified as a means of fixing the explanatory data. The misrepresentation objection to the HO theory does not show that an intrinsic concept is warranted; rather, it shows how well-entrenched this conceptual dead end can be.

The paper is organized as follows. I start by presenting the particular extrinsic concept embraced by the HO theory and by sketching a particular version of the view, David Rosenthal's HOT theory. Then I turn to several versions of the misrepresentation objection, developed by Alex Byrne, Karen Neander, Joseph Levine, and Uriah Kriegel. Next, I argue for a conditional claim: *if* the data is fixed by the HO theory's chosen extrinsic concept, there is no problem of misrepresentation. I close by defending the antecedent of the conditional, establishing that there is only a problem of misrepresentation if we misrepresent consciousness as being intrinsic to conscious states.

1 The transitivity principle

The HO theory fixes its explanatory target by drawing out a commonsense distinction between conscious and nonconscious mental states. The theory holds that ordinary folk allow that mental states sometimes occur nonconsciously, and it further contends that this distinction extends even to sensory states, like seeing red or being in pain. Commonsense intuition finds nothing incoherent or contradictory in the idea that we might subliminally see a red stimulus or fail at times to consciously feel pinching shoes or an all-day headache. HO theorists take this to indicate that folk psychology does not conceive of mental states as essentially involving consciousness. The HO theory then asks what the difference is, in commonsense terms, between a conscious and nonconscious mental state. The answer, it's claimed, is that a conscious state is a mental state a subject is aware of

[12] The main defenders of HO theory are Armstrong (1968, 1980), Rosenthal (1986, 1997, 2005), and Lycan (1987, 1996). See also Carruthers (2000).

[13] Byrne (1997), Neander (1998), Levine (2001), and Kriegel (2003). See also Van Gulick (2001, 2004).

herself as being in. The contrapositive of this claim, in particular, is taken to capture the commonsense distinction between conscious and nonconscious states: if a subject is in no way aware of herself as being in a mental state, that state is not a conscious state. This is the extrinsic concept of consciousness that fixes the data the HO theory aims to explain.

There is one minor emendation to this way of fixing the data. From a commonsense perspective, it is possible to be aware of oneself as being in a mental state without that state being conscious. If a subject consciously infers that she is in a particular mental state, that state won't intuitively be conscious. If I am told by another person, a reliable source whom I believe (my wife, say), that I am in some mental state, I may well be in some sense aware of myself as being in a particular mental state without that state being conscious. Likewise, if I observe a certain pattern in my own behavior, I might consciously infer on the basis of theory that I must be in some mental state. That, too, will not make a state conscious. Proponents of this extrinsic concept contend that for a mental state to be conscious, the subject must be aware of it without employing any conscious inference. The awareness must occur in a seemingly immediate way; it must not seem that any process or inference lies between me and my conscious state. Accordingly, a mental state is conscious when a subject is aware of herself as being in that state in a seemingly immediate way. All this is still meant to be pretheoretic.[14] It marks off an everyday sense of what it is to be in a conscious state.

David Rosenthal, who has developed and defended the "higher-order thought" (HOT) version of the HO theory, calls this extrinsic conception of consciousness the *transitivity principle* (TP).[15]

> TP: A subject's mental state is conscious when the subject is suitably aware of herself as being in that state.

This concept is meant to capture, pretheoretically, what consciousness is, what the term 'consciousness' picks out. The task of the HO theory is then to explain how this process is instantiated in us. The concept is clearly extrinsic: a conscious state is fully characterized in terms of its connection to the subject's awareness, appropriately specified. It makes no mention of the intrinsic qualities of conscious experience, nor does it make any claims about properties that can vary independently of any functional, causal, or intentional processes, events, or entities. The term 'suitably' marks off the need for the awareness to seem to the subject to be immediate.

Note that I am not here arguing for the correctness of this concept of consciousness. For present purposes, I only wish to establish that this concept fixes the data that the HO theory intends to explain, and that the conception is argued to be a pretheoretic, commonsense notion, one having folk-psychological standing independent of the HO theory. The HO theory identifies this phenomenon as its central explanandum and attempts to address any remaining explanatory issues in the context of the mechanisms posited to explain the TP.

[14] Or folk-theoretic—I will use these terms interchangeably.

[15] Rosenthal (2000). See also Rosenthal (1997, 2005) and Lycan (2001).

A further claim of the HO theory, straddling the line between pretheoretic platitude and philosophical analysis, is that the TP effectively captures what is meant by Nagel's "something it's like" phrase. Recall that Nagel himself stresses the 'for' in "something it is like *for* the organism." HO theorists contend that, in keeping with the TP, there would not be anything it's like *for* the organism if the organism was in no way aware of its states. Whatever qualities may or may not be present, if the subject is not aware of those qualities, there will be nothing it's like to be that subject, *for* that subject. This analysis perhaps is not simple folk-psychological exegesis—Nagel's formulation is meant as a piece of philosophical clarification. But it does seem to capture the subjective aspect of Nagel's discussion and so it stands as one reasonable reading of his claim. Thus, it is open for the HO theorist to contend that they've explained why there is something it is like for the organism when they've explained, by way of HO theory, how it is that a subject is aware of the states she is in. There is something it's like for her because she is suitably aware of herself as being in those states.[16]

The TP fixes the data the HO theory aims to explain. It is a commonsense extrinsic concept of consciousness, one characterizing the nature of conscious states in terms of their connections to other states, processes, or entities. And the TP offers a reasonable interpretation of Nagel's "something it's like" phrase, suggesting that it captures what is central to that popular philosophical way of pinning down the phenomenon to be explained by a theory of consciousness. Whether the HO theorists are correct that the TP appropriately fixes the explanatory data and captures what is central to Nagel's phrase has not yet been defended. For present purposes, note that this is what HO theorists *take* their theory to be explaining and this is how they interpret the "something it's like" phrase. Next, I will turn to a specific version of the HO theory, the "HOT" theory of Rosenthal, in order to clarify how the theory is supposed to explain the phenomenon picked out by TP. Then I will introduce the objection thought to undermine the theory.

2 The HOT theory

The HOT theory purports to explain what it is for a mental state to be conscious. By 'conscious' it means just what the TP says: what it is for a subject to be suitably aware of herself as being in a mental state. HOT theory, like its nearby rival higher-order perception (HOP) theory, explains consciousness by positing a distinct representational state about the state the subject is aware of being in. This separate representational state is about another mental state—hence the "higher order" appellation. This higher-order representation, in the case of HOT theory a species of conceptual intentional state (hence 'thought'), must occur in a seemingly spontaneous way; there can be no conscious inference or observation mediating the subject's awareness of her conscious state. This accounts for the "suitably aware" marked off in the TP. HOT theory holds, therefore, that a mental state is conscious, in the sense specified by the TP, when the subject represents that state with a HOT

[16] Cf. Rosenthal (2002).

formed in a seemingly spontaneous fashion. If that occurs, holds the view, the subject will be suitably aware of herself as being in a mental state; there will be something it is like to be that subject—*for* that subject.

For present purposes, it is important to note that the content of the HO state—what the HO state makes the subject aware of—fully determines what it is like for the subject. This makes sense, given the view: if consciousness amounts to an awareness of oneself as being in a state, and HO representation fully accounts for this process, then the HO state fully determines how one is aware of oneself. The HO state is wholly responsible for what Karen Neander terms "the phenomenal labor."[17] It does all the work of fixing what it's like for the subject in conscious experience. This fact will very much be at issue in what follows, so it is worth stressing. Consciousness is defined according to the TP, on the HO view. Conscious states are states we are aware of ourselves as being in. This process of awareness is explained by HO representation: by representing myself as being in such-and-such a state, that state becomes conscious. But since the only access I have to this state is by way of HO representation, the HO representation fully determines what it is like for me.

HOT theory offers a route to a reductive explanation of consciousness because it is widely held that representation is explainable in naturalistic terms. If consciousness can be explained by representation, and representation can be explained in naturalistic terms, then consciousness can be explained in naturalistic terms. Of the several naturalized theories of representation on offer, all seem compatible with HO theory. Causal-nomological covariation, asymmetric dependence, teleological proper function, or some variety of causal-role all offer ways to account for representation with naturalistically-acceptable ingredients. Whichever way the complex debates over naturalized representation play out, HO theory can remain neutral; any means of naturalizing representation ought to do, providing the route to a naturalistic explanation of consciousness.

However, this appeal to naturalized representation opens the door to the key objection providing the focus of this paper. Representation, by its nature, can occur even if the object of representation—the thing that the representation is about—does not exist. Representation is marked by the possibility of misrepresentation. In the course of developing his HOP view, David Armstrong notes that any naturalized representational process, including introspection, can go astray. It's always possible, for well-known Humean reasons, for one element in a causal process to occur without the other; causes are logically independent from their effects. If we allow that representation is ultimately some species of natural, causal process, then representation can go wrong.[18]

But given the HO theory's commitment to representation as the crucial mechanism of consciousness, what happens if a HO state misrepresents? What happens, that is, if a HO state misinforms the subject about her conscious state, or even represents her as being in a state she is not in? Several philosophers have noted this worry and developed it into an outright objection to the HO theory. Alex Byrne,

[17] Neander (1998).

[18] Armstrong (1968).

Karen Neander, Joseph Levine, and Uriah Kriegel have all offered versions of the objection, each making the case in somewhat different ways that the possibility of higher-order misrepresentation presents a serious, perhaps fatal, problem for HO theory. Note that the objection does not turn on any empirical result; rather, it is the mere *possibility* of misrepresentation that is enough, according to the objectors, to expose the perhaps fatal flaw in the HO theory. The issue turns on conceptual questions.

3 The misrepresentation objection

The misrepresentation objection is brought out by considering the following three scenarios, each possible on the HO theory. First is what we can call the *veridical case*. In this case, I have a first-order state and a higher-order state that accurately represents it. For instance, I might have a first-order state of seeing red and a higher-order state that makes me aware of myself as seeing red. The second case we can call *mild misrepresentation*. In this case, I might have a first-order state of seeing purple, but be aware of myself as seeing red. In the third case, which we can call *radical misrepresentation*, I might lack the relevant first-order state altogether, but nonetheless represent myself as seeing red. We can ask what it will be like for the subject in each of these cases. In the veridical case, it will be like seeing red; one will have a conscious visual experience of red. But what will it be like for the subject in the case of mild misrepresentation? According to the HO view, it will be the same. Rosenthal writes that such a case will be "subjectively indistinguishable" from the veridical case for the subject.[19] That is, it will seem to the subject that she is seeing red, that she is having a visual experience of red, even though she is in a first-order state of seeing purple. And the same goes even for the case of radical misrepresentation: it will seem to the subject that she is seeing red, that she is having a visual experience of red, even though the first-order state is not present at all. Because the subject has no other first-person access to her states except by way of HO representation, she will not be able to tell the difference between the veridical case, mild misrepresentation, or radical misrepresentation. What it will be like to be her—*for* her—will be subjectively indistinguishable in these three cases.

According to the objectors, this spells doom for the HO view once the implications of this fact are made clear. Byrne (1997) contends the misrepresentation problem shows that the HO theory fails as an explanation of consciousness. The theory claimed to explain consciousness, in the "phenomenal" what-it's-like sense, by positing a two-tiered representational structure. But radical misrepresentation makes clear that the bottom tier is explanatorily idle: it makes no contribution to the HO explanation of consciousness. But all that's left then is the HO state itself; it alone must explain why there's something it's like for the subject. But how could a single thought do that? If a single thought can explain consciousness, why not just posit a conscious first-order state? But that can be seen to fall short: we want to know *why* this thought, as opposed to others, is conscious. All that the HO theorist

[19] Rosenthal (1997, p. 744). See also Rosenthal (2004).

seems to have left is an appeal to the etiology of the state—that it comes about in a seemingly noninferential way. But this too seems to offer nothing in the way of an explanation of why there is something it's like in conscious experience. Adding the same requirement to a first-order state fails to explain consciousness, so why would it do so with a HO state?

Neander (1998) and Levine (2001) press a slightly different way of making the misrepresentation worry salient. Neander notes that radical misrepresentation makes the first-order states irrelevant to determining what it's like for the subject. The experience would be the same, radical misrepresentation shows, even if the first-order state was different or absent altogether. But first-order states are defined as sensations—they are the qualitative aspects of experience, states marked by qualia. But qualia are supposed to be *constitutive* of experience. If the HO theory allows that qualitative states do not constitute what it's like for the subject, the theory is no longer talking about qualia, and thus has changed the subject. It is not a theory of consciousness at all.

Levine (2001) stresses a similar worry to Byrne's—that the HO theory's explanatory structure collapses—but he also contends, like Neander, that the very phenomenon of explanatory interest seems to disappear on the HO view, once we grasp the implications of the misrepresentation problem. This shows, he claims, that conscious qualities and our awareness of them cannot be divorced, on pain of failing to explain the subject initially drawing our interest.

Finally, Kriegel (2003) contends that radical misrepresentation forces the HO theory into an extremely counterintuitive claim. According to HO theory, conscious states are ones we are conscious of ourselves as being in. But there is no first-order state present in radical misrepresentation, by hypothesis. Still, it will *seem* to the subject that she is in a conscious state. There will be something it is like for her to have this experience—it will be subjectively indistinguishable from the veridical case. But what state is conscious in the radical case? It can't be the HO state itself—we are in no way aware of that state. But the lower-order state doesn't even exist. Surely, a state must exist to be conscious. Thus, radical misrepresentation apparently exposes a theoretical incoherence in HO theory—we can seem to be in conscious states when we are not. But if it seems any way to us at all, mustn't we be in a conscious state?

However, this may not simply be a counterintuitive conclusion. The objection can be unpacked as a more formal argument, issuing in a destructive dilemma for the HO theory. Consider the following way of fleshing-out the misrepresentation problem:

1. If there's something it's like for the subject, the subject is in a conscious state.
2. In radical misrepresentation, there is something it's like for the subject.
3. So, in radical misrepresentation the subject is in a conscious state.
4. The conscious state is either the first-order state or the higher-order state.
5. The first-order state does not exist in radical misrepresentation.
6. The subject is not aware of the higher-order state in radical misrepresentation.
7. So, either the subject is in a conscious state that does not exist, or the subject is in a conscious state that she is not aware of being in.

8. A subject can't be in a conscious state that does not exist.
9. So, the subject is in a conscious state that she is not aware of being in.

This conclusion effectively eliminates or badly limits the explanatory target of HO theory. It appears that the HO theory is refuted simply by allowing the possibility of misrepresentation: if there are conscious states the subject is not aware of being in, then HO representation will be irrelevant in explaining how those states are conscious. And if it is irrelevant in some cases, why think it is ever relevant?

Prima facie, the argument is valid, and it contains either premises the HO theorist explicitly accepts or premises that seem obvious. Premise 1 is just the Nagelian definition of consciousness, accepted by all parties, given the right interpretation. Premise 2 follows from Rosenthal's claim of "subjective indistinguishability": a state subjectively indistinguishable from a conscious state is also a conscious state. Premises 4–6 are offered ex hypothesis: they simply recount what is going on in radical misrepresentation. Premise 8 seems just obvious: how can one be *in* a state if the state does not exist, and a fortoriori, how can *that* state be conscious? Thus, the conclusion seems forced upon the HO theorist. But to accept the conclusion is to give up the view, or at least to acknowledge that the view is of disappointingly limited scope and that there are other ways for states to be conscious.

4 The response: conscious states as *represented* states

Despite how things may appear, however, the misrepresentation objection does not undermine HO theory. The objection loses sight of the HO theory's commitment to a fully extrinsic characterization of consciousness, given by the TP. And this leads to a misreading of how the HO theory intends the phrase 'the subject is in a conscious state.' Once these errors are rectified, it becomes evident that the HO theory can avoid the objection. In this section, I present the HO response to misrepresentation, in order to establish the conditional claim that if the data is fixed as the HO theory intends, then there is no problem of misrepresentation. Then in the final section, I'll defend the HO theory's way of fixing the data.

Fixing the meaning of 'the subject is in a conscious state' requires a characterization of what it is to be a conscious state in the first place. According to the HO theory, this is provided by the TP. The TP tells us that a conscious state is one that the subject is aware of herself as being in. However, nothing in the TP rules out the possibility of one's awareness being in error, even in radical error. The TP employs the intentional construction "aware of oneself as…" And one can be aware of oneself as being in a state even if that state does not exist. For example, I can be aware of myself as being handsome when I am not in fact handsome. That is to say, it seems to me in a direct way that I'm handsome, even though I'm not. Or I can be aware of myself as possessing magical powers even though no such powers exist.[20]

[20] Note that the phrase 'aware of' may be factive. But this sort of factivity is satisfied by the TP: even in radical misrepresentation the self one is aware of is guaranteed to exist by cogito-style reasoning. However, the particular state the self is in need not exist. Being aware of oneself as being in a state is similar to being aware of an object as possessing a property—we can be aware of the object as possessing

Being aware of oneself as being in this or that state or possessing this or that property does not entail that the state or property must exist. And because the TP employs this notion in characterizing consciousness, it follows that the state one is aware of oneself as being in during conscious experience—the conscious state—need not exist.

It then follows that the notion of being in a conscious state allows for the possibility of radical misrepresentation. The notion of being in a conscious state is derivative upon the notion of being a conscious state. That is, to understand what it is for one to be in a conscious state, we first need to have a way of distinguishing conscious states from nonconscious states. On the HO theory, this is done by the TP. It follows that on the HO theory, all there is to being in a conscious state is to be in a state in the sense derived from the TP's characterization of a conscious state. And all that amounts to is being aware of oneself as being in a state, without any commitment to the existence of that state. Therefore, one can be in a conscious state even if that state does not exist, so long as one is suitably aware of oneself as being in that state.

With this analysis in hand, the HO theorist can reject premise 8 in the above argument. The subject can be in a conscious state that does not exist because all there is to being in a conscious state is being aware of oneself as being in that state. "Being aware of oneself as…" introduces an intensional context. Conscious states are simply objects of representation, and as such they need not exist. In the same way that I can be aware of the yard as containing a deer, even though no deer exists in the yard, and in the same way I can be aware of myself as being hungry, even when the biological state of hunger doesn't exist in me (I've just eaten a big meal, say), I can be aware of myself as being in a mental state that I'm not in fact in. And since that is all that's required for being in a conscious state according to the TP, I can be in a conscious state that does not exist, contrary to premise 8 above.

But surely, one might argue, it strains commonsense intuition to think that someone can be in a conscious state that does not exist! The worry gains support from noting that in general, when one attributes a mental state to someone, the state must exist if the attribution is true. For example, if I say of someone that they're in a depressed state and no such state exists, then what I've said is false. Generally, state attribution is factive—it does not introduce an intensional context. What's more, the HO theorist's intensional reading of the TP (and the derivative reading of 'being in a conscious state') seems to allow for the possibility of a creature that is in a range of conscious states, even though none of its conscious states exist. And it may be that none of the creature's conscious states have ever existed, despite there being something it's like for the creature. This seems highly counterintuitive, suggesting that the TP cannot be given an intensional reading.

But the strain with common sense is not as great as it appears. Common sense allows that we make errors about the conscious states we're in. What's more, if we

Footnote 20 continued
a property even if the property we're aware of the object as possessing fails to exist. The crucial construction here is 'being aware of X as Y,' rather than 'aware of.' The former introduces an intensional context even if the latter does not. On the factivity of 'aware of' see Huemer (1998, Sect. 1.1).

accept such errors, there seems to be no principled way to rule out the possibility of radical misrepresentation. First, there are lots of everyday situations where we can go wrong about the specifics of the conscious state we're in. For example, I might be aware of myself as feeling angry about an incident at work when in fact I'm angry about the poor play of my favorite football team. Or I might be aware of myself as hearing my name called out in a crowd, when in fact what I hear is a name with the same first syllable as mine. And I might even mistakenly take myself to be in pain. In so-called "dental fear," patients' fear combined with the whirr of the drill causes them to mistake the pressure of the dentist's hand in their mouth as pain, even though their teeth are fully numb.[21] These sorts of errors correspond to the "mild" misrepresentation scenario I described in Sect. 3 above. Such cases do not clash with our intuitions: even if they are somewhat uncommon, we don't find them intuitively incoherent or impossible. Yet even mild misrepresentation introduces a gap between the mental state I'm in and the mental state I'm aware of myself as being in. If I'm aware of my state of anger about the football team as anger about an incident at work, then I am not accurately aware of the state I'm in. Further, what it's like for me is as if I'm in a state of anger about the incident at work—the state I'm aware of myself as being in fixes how things seem in mild misrepresentation cases.

But if commonsense allows for a small gap between the state I'm in and the state I'm aware of myself as being in, why think it can rule out bigger and more radical mismatches? Is there any property of a state that cannot be misrepresented? And if any particular property can be misrepresented, why can't they all be? There does not seem to be a principled way of ruling out more radical mismatches. What's more, it's not clear how one in general distinguishes between misrepresenting an object's properties and representing another object that isn't present. I may, for example, misrepresent piles of laundry in my basement as pink elephants. But I could just as easily be described as representing nonexistent pink elephants. If the line between mild and radical misrepresentation can't be drawn in perceptual cases, it's not clear how it could be drawn in metarepresentational cases. Thus, there seems no principled way to rule out the possibility of radical misrepresentation, once mild misrepresentation is allowed.[22] If commonsense allows for mild misrepresentation, and there's no principled difference between mild (though extensive) misrepresentation and radical misrepresentation, then it seems that radical misrepresentation cannot be ruled out by commonsense. This supports an intensional reading of the TP and the derivative notion of being in a conscious state.

One might respond that so long as some state or other is there to be misrepresented, then the case is still mild misrepresentation. We will then still have an existing state to label as conscious and we will not be forced into saying that one can be in a conscious state that does not exist. But then it's not at all clear what holding onto the first-order state does for us beyond preventing an awkward way of speaking. Indeed, this is to surrender the idea that the first-order state engages in any

[21] See Rosenthal (2005, p. 209).

[22] Cf. Byrne (1997, p. 129, footnote 48).

of the "phenomenal labor," as Neander and Levine charge. The difference between such a view and the HO theory is minimal.

But it still may seem that something in the HO response to misrepresentation clashes with ordinary intuition. At the very least, saying that one can be in a conscious state that does not exist clearly strains normal discourse. At this point, however, the HO theorist can simply acknowledge the strain with common sense, but argue that it is an acceptable price to pay for a reductive explanation of consciousness. The HO theory starts with a commonsense way of fixing the data, the TP, and then posits HO representations as the best explanation of how the TP is realized in us. HO representation in turn is cashed out naturalistically, clearing the way for reductive explanation. And this of course opens up the possibility of radical misrepresentation. But it is often that case that a successful theory forces a revision of commonsense intuition, even those intuitions used to characterize the target data in the first place. Such a revision is justified by the explanatory virtues of the successful theory. So even if an intensional reading of the TP generates a strain with commonsense, that is not enough to deny the intensional reading, given the prospects of a reductive explanation of consciousness offered by the HO theory.[23] And because commonsense does allow for mild misrepresentation, the revision is not as great as it may have looked at first blush.[24]

But something still seems deeply amiss. The revision being demanded here is not just of a minor way of speaking. We're being asked to accept that we can be in a conscious state that does not exist, however, one wishes to speak about it. Independently of questions of commonsense language, this is problematic: surely, if we're in a conscious state, that state must exist. But at this point the HO theorist can hold the line and ask what underwrites this belief—what is it that makes us so sure that conscious states aren't just states we represent ourselves as being in? All that's left to support this belief, I contend, is a (perhaps implicit) commitment to an intrinsic conception of consciousness. From the perspective of an intrinsic conception, there is something substantial to consciousness, something beyond what can be accounted for by representation. How could mere representation account for the conscious experience of the rich reds and pinks of a sunset or the creamy smoothness of a freshly pulled pint of Guinness? And surely our contact with the reds and pinks of the sunset or the creamy taste of the Guinness is more intimate than mere representation can explain. Or to raise the worry in another way, there seem to be a range of representational systems with the capacity to represent

[23] This is not, of course, to claim that the HO theory is true—that is an empirical question. Rather, it is to point out that if it's true, we can expect some revision to our ordinary ways of thinking about consciousness. That's a common price to pay for good theory.

[24] It is worth noting that an additional source of the tension here is that even on a HO view, the subject can not *directly* recognize her error in radical (or even mild) misrepresentation. This is in contrast to how we normally recognize perceptual error. In perceptual error, we directly perceive additional information or we consciously cross-calibrate our perceptions in one modality with perceptions in another. We can then just *see* that we were wrong. In the case of inner misrepresentation, by contrast, there is no analogous line of independent but direct counter-evidence. Thus, the idea of internal error will seem counterintuitive, even if we have strong theoretical and empirical reasons to accept such error. If we never get direct counterevidence, why think we are ever wrong? The HO theory explains this intuition while rejecting the implausible claim that we never make internal errors.

their own states which nonetheless fail to be conscious. Perhaps my laptop is an example of such a system. Mere representation, even of oneself, is just not enough for consciousness.

But this is just to embrace an intrinsic conception of consciousness. It is to say that there must be something intrinsic to the medium of representation that accounts for the fact that we are conscious. This begs the question against the HO theory and its extrinsic conception of consciousness. What's more, things are made worse if one thinks that the only way to distinguish between conscious and nonconscious representational systems is by appeal to intrinsic nonrepresentational properties of experience. Such properties enter into the content of conscious experience—they constitute, in part, what it's like for us—but they are not represented properties at all.[25] If one is tempted by such a view, then something intrinsic to the underlying states involved in consciousness will seem necessary for the presence of conscious experience. Further, those features can't be represented properties; by definition, the properties at issue are nonrepresentational. And because these properties seem to be features associated with experienced objects—the experienced colors and textures of things, for example—it's plausible that the nonrepresentational properties are instantiated by first-order states representing objects in the world. That would entail that if one is conscious of a red object, then there must be a first-order representation of that red object instantiated in the right medium to account for the experience. First-order states, the states one is aware of oneself as being in on the TP, must exist to provide special nonrepresentational features of the conscious experience of objects. The states we are aware of ourselves as being in, that is, must exist if what it's like for us is as if we're seeing a red object. The TP fails to guarantee this fact, so it cannot be a proper characterization of consciousness.

It is obvious that this conception of consciousness is intrinsic and thus question-begging in the current context. What's more, this particular intrinsic conception claims that there are nonrepresentational properties in the content of conscious experience. That certainly begs the question against the TP, with its explicit invocation of awareness, plausibly a representational notion. If one is tempted by this kind of intrinsic conception, the misrepresentation will seem especially problematic. Evidence for a prior commitment to this sort of conception is found in the surprise some feel upon discovering that "that's all there is" to the HO theory. It's just an inner form of representation, and representation construed extrinsically, as a matter of connections. What seemed at first like a robust alternative to "thin" representational and functional views is "exposed" by the misrepresentation objection as a kind of subject-changing bait and switch (see Neander and Levine's comments above). But if one is looking for intrinsic qualia in the first place, one is already in the grip of an intrinsic conception of consciousness. And that is to misrepresent the data the HO theory aims to explain.

The allure of the misrepresentation objection, therefore, turns in part on the pull of the intrinsic conception. And to make matters worse, the extrinsic nature of the TP seems particularly easy to miss if one is influenced by an intrinsic conception. The TP holds that conscious states are states we're aware of ourselves as being in.

[25] See, for example, Block (1996).

This appears to retain an intrinsic element, namely the presence of qualitative sensory states as the objects of awareness. If those states' intrinsic properties play a constitutive role in what it's like for one, then the intuition that something intrinsic to conscious states is necessary for experience will be satisfied. But once it's made clear by the misrepresentation objection that the TP is fully extrinsic it feels to those tempted by an intrinsic conception that something essential has been lost. The intrinsic qualia constitutive of consciousness according to such a view have disappeared! But if one gives up on an intrinsic conception, the fact that the intrinsic properties of the represented state need not exist is not a problem. All there is to conscious experience is being aware of oneself in the right way. And that can be fully accounted for by HO representation.[26]

A final important point. Worries closely analogous to the misrepresentation problem arise for all other extrinsic concepts of consciousness. The extrinsic concept of *global accessibility*, for example, holds that a mental state is conscious when it is appropriately available to a broad range of mental systems.[27] However, we can imagine a scenario where the state so poised in the global workspace is excised while all the subsystems accessing the workspace are stimulated *as if* the state were present. What would it be like to be a subject in this scenario? If all the accessing systems are engaged in just the way they would be in the normal case, there would be no detectable difference in what it's like for the subject. Rejecting this conclusion is to hold that something intrinsic to the state in the workspace constitutes consciousness; it is to embrace an intrinsic conception. If one sticks with an extrinsic concept this sort of "false access" is a possibility. Again, one might take this as reason to go intrinsic; my point is just that extrinsic concepts lead to the possibility of misrepresentation, false positives, and absent or inverted *intrinsic* phenomenal character, and the like.

It is more obvious that the extrinsic concept employed by first-order representationalism (FOR) is open to a parallel objection. The FOR view holds that conscious states are states making us appropriately aware of the world. In conjunction with this

[26] There is another related worry nearby, one that might be thought to lend support to the misrepresentation worry. HO theory seems committed to the claim that a nonexistent state can have a property, the property of being conscious. But nonexistent things can't have properties—they don't exist! (See Mandik 2009). This, however, is a general problem for all theories of intentionality. And in this context, it seems reasonable to invoke Harman's plea (1990).

> Let me concede immediately that I do not have a well worked out theory of intentional objects…. Indeed, I am quite willing to believe that there are not really any nonexistent objects and that apparent talk of such objects should be analyzed away somehow. I do not see that it is my job to resolve this issue. However this issue is resolved, the theory had better end up agreeing that Ponce de Leon was looking for something when he was looking for the fountain of youth, even though there is no fountain of youth…. If a logical theory can account for searches for things that do not, as it happens, exist, it can presumably also allow for a sense of "see" in which Macbeth can see something that does not exist (pp. 37–38).

Nonexistent conscious states seem no worse off than nonexistent fountains of youth. If the fountain of youth can have the property of being looked for, a nonexistent state can have the property of being conscious, given that 'being conscious' just means 'being an object of awareness.' Whatever answer works for the one works for the other as well. Thanks to Pete Mandik for pressing this issue.

[27] Baars (1997), Dennett (1991), and Dehaene and Naccache (2001).

view FOR theorists argue that qualia are perceptible properties objects are *represented* as having.[28] Further, FOR views tend towards externalism about mental content in general and the content of qualitative representations in particular. Qualia, as they say, just ain't in the head. But the possibility of misrepresentation is inherent in representation. So what happens when I visually misrepresent the world as containing a bloody dagger, on this view? Where are the qualia? If they are not in the head, where could they be? But these are the very qualities of experience, the feel of what it's like for the subject. How could they be constitutive features of my experience if they do not exist? The answer, of course, is that they are represented as being in the world—they are the representational *content* of the FOR. And that is enough, on the FOR view, to make them the qualities of consciousness. Thus, a similar objection can be made against the extrinsic concept of consciousness employed by FOR.[29] It appears that all extrinsic concepts face some version of the misrepresentation objection. Why think, then, that the objection stems from anything more than an implicit or explicit allegiance to the rival intrinsic concept? If one thinks that there's something intrinsic to conscious states accounting for their consciousness, then all extrinsic concepts will eventually be "exposed" as failing to capture the special inner light of qualia. But that is question begging in the current context.

This concludes my defense of the conditional claim that *if* the explanatory data is fixed by the TP, then there is no problem of misrepresentation.[30] By this point, the reader may feel that while the conditional may be true, if has become obvious that the TP cannot be what fixes the data or cannot be what fixes the data alone. Perhaps the TP is only a necessary, rather than a necessary and sufficient, condition for a mental state's being conscious and what must be added to properly fix the data is some sort of prohibition on misrepresentation. Or perhaps the TP does not even set a necessary condition for being a conscious state. What, then, determines the proper way to fix the data?

5 Defending the transitivity principle

In this section, I will argue that the TP is better supported than its intrinsic rivals.[31] I will contend that both TP and its intrinsic rivals capture important aspects of our folk-psychological commonsense conception of consciousness. However, the TP

[28] Dretske (1995) and Tye (1995). See also Harman (1990).

[29] Cf. Mandik (2009).

[30] While it should be clear from the text how my response answers the objection as formulated by Neander, Levine, and Kriegel, an additional word is in order concerning Byrne. Byrne's objection focused on the HO theory's alleged explanatory failure: how could a single HO state *explain* phenomenal consciousness? First, HO theory is not trying to explain an intrinsic feature with an extrinsic mechanism; rather, the feature requiring explanation is itself extrinsic. Second, if the TP provides the explanandum, then the HO theory is clearly a better explanation than the FOR theory, because FO states do not makes us aware of our mental states. Finally, etiology matters for explaining the appearance of immediacy characterized by the TP.

[31] I will not argue here for the TP over its *extrinsic* rivals. For that, see Rosenthal (2005) and Lycan (1996), for a start. See also Carruthers (2000).

provides a better "mesh" with current and potential empirical research. All other things being equal, we should favor a concept that provides the best fit with empirical data, so the TP stands as the better concept. I will close by arguing that a restricted version of the TP explicitly designed to rule out the possibility of misrepresentation fails to provide an improved conception of consciousness.

There are two main ways to provide support for a concept of consciousness. First and foremost is showing that the concept best captures our folk-psychological intuitions about consciousness. This assures that we are indeed focused on the right phenomenon and that we have not illicitly changed the subject or operationalized away the problem. It is also important for fixing the pretheoretic appearances that must be accounted for by a theory. Even if a theory rejects some aspect of common sense, it still must explain why it *seemed* from the everyday perspective that things were that way pretheoretically. Failing to do so leaves us with nagging doubts and open questions. It won't be a satisfying explanation of consciousness. Thus, the better a concept of consciousness captures commonsense intuition, the better it will do at fixing the right phenomena to be explained.

Second, if a concept of consciousness satisfies the previous constraint, we can consider how well it "meshes" with empirical theorizing, in Ned Block's term.[32] If a concept of consciousness so defines its extension that no empirical explanation is even possible, that is a mark against it. All things being equal, a conception that allows for a fit with empirical theorizing is to be desired. That does not mean that we cannot discover after long empirical effort that consciousness defies explanation; it may be that some things are too difficult for the likes of us to explain. But so long as the first desideratum is met, a concept not ruling out the very possibility of empirical explanation has the advantage. If two concepts equivalently capture our intuitions about consciousness, the one with better mesh is be favored, at least without further argument to the contrary.[33]

I contend that the TP and its intrinsic rivals both capture important aspects of our folk-psychological conception of consciousness. The TP marks a commonsense distinction between conscious and nonconscious mental states. First, it's clear that in everyday discourse we accept that mental states can occur both consciously and nonconsciously. Cases of repressed desire and emotion, of hidden motives and passions, and even of subliminal perceptions and sensations are well understood and entrenched in everyday ways of speaking. Folk do not find such cases incoherent or paradoxical; indeed, they are central to many plots of popular books, television, and movies, not to mention our everyday gossip about colleagues and neighbors. We can then ask what the difference is between conscious and nonconscious states, so conceived. An intuitive answer is that we are aware of the conscious cases while we are not aware of the nonconscious cases, even though they still occur in us. Consider

[32] Block (2007).

[33] A third means of providing support for a concept of consciousness might be proper connection with the history of philosophy. If a concept is rooted in long-standing traditional debates, it may gain some measure of support. Here, both views can point to tradition, with the intrinsic view noting Locke's secondary qualities and the 20th century debates over sense data. The TP can trace its roots back to Brentano, Kant, Locke, and even Aristotle. So neither side gains a distinct advantage, if there is support to be gained in this way.

a case where I am angry, but my anger is not conscious. I may adamantly deny that I am angry, while my wife maintains that I am indeed angry. Later, I may come to realize she is right. What happens then? Plausibly, I become aware of my anger, in the appropriately unmediated manner of consciousness. I become aware of myself as being angry and my anger becomes conscious. And if I'm in no way aware of my anger, it's not intuitively conscious. This is what the TP captures. It is the folk-psychological distinction between conscious and nonconscious mental states.

Some critics have argued, however, that the TP is not well-supported by common sense. Charles Siewert, for example, contends that the appeal of the "conscious of" locution does not point to the TP; rather, it highlights the fact that we often use the phrase 'conscious of' to pick out the thing we are conscious of in the world. We are conscious of apples or trees or loud noises, but this does not entail that we are thereby conscious of our mental states themselves. Further, we sometimes use 'conscious of' to indicate knowledge of worldly facts; for example, I might be conscious of the trade gap or the situation in the Middle East. But that, too, does not indicate that I'm aware of anything mental. So, while we do indeed use the phrase 'conscious of' in everyday discourse, we shouldn't fall into the "'consciousness of' trap" and think this points to the TP.[34]

But Siewert is focused on the wrong phenomenon. HO theorists readily agree that we often use 'conscious of' to refer to our awareness of things in the world. Indeed, Rosenthal calls this "transitive consciousness" and makes much of its importance. What is at issue, rather, is a folk-psychological answer to the question of what accounts for the difference between conscious and nonconscious mental states. And here, according to proponents of the TP, folk will accept the necessity of being aware of one's state. It is true but irrelevant that folk also use 'conscious of' in other ways. What matters is that when faced with the question at hand, folk accept the TP. To reiterate, folk will not consider a state conscious if the subject is in no way aware of it. And this is equivalent to the TP, indicating its folk-psychological acceptability, even taking into account Siewert's observation of other uses of 'conscious of.'

Another possible worry about the folk-psychological standing of the TP comes from Alex Byrne. Byrne defends the FOR account of consciousness against the HO theory by arguing that the Nagelian "something it's like" phrase need not by read according to the TP. He contends that there being something it's like for the subject does not require an awareness of a mental state or an awareness of the subject. The syntax alone of the phrase does not entail this sort of reading. Further, Byrne also notes that even if the subject is aware of something in the Nagel case, it does not follow that the subject must be aware of a mental state or aware of herself. She might be aware of a book or a tree or a tomato. The HO reading of Nagel's phrase is not forced on us.[35]

Again, this is quite correct, but irrelevant to my point. While the TP reading of Nagel's phrase may not follow by syntactic argument, and while it may be the case that we are often aware of something other than ourselves or our mental states, when we focus on the folk-psychological difference between conscious and nonconscious

[34] Siewert (1998, pp. 194–197).

[35] Byrne (2004).

mental states, the TP provides a reasonable commonsense gloss of the distinction. And if that's the case, then the 'for' in 'something it's like *for* the subject' is plausibly read according to the TP. If the subject is in no way aware of her state, how can it be *for* her in the sense Nagel intends? This is not to say that the phrase can't be read in a number of different ways: syntax rarely restricts a phrase to one unambiguous interpretation. Rather, on the most plausible folk-psychological reading, the TP interpretation follows. And, again, it is true but irrelevant that we are often (indeed, usually) aware of things besides mental states. It's just that in considering the difference between conscious and nonconscious mental states, we note the intuitive necessity of an awareness of mental states. And that is enough to justify the TP in the current context—it captures an important aspect of our everyday understanding of consciousness.[36]

But what about the introduction of an intensional context by the TP? Haven't I already conceded that this is counterintuitive, undermining the claim that the TP captures a commonsense way of characterizing consciousness? Here it is important to recall two points argued for above. One, common sense does allow that we're sometimes wrong about the conscious states we're in. So even if the degree of possible error is surprising, the TP accurately captures the tolerance of error present in our everyday notion of consciousness. Common sense does not license infallibility; the TP reflects this fact. Two, the presence of an intensional context follows from accepting the HO theory. It is not forced on us by the TP. The TP is the commonsense starting point employed by the HO theory to pin down the explanatory data. The theory then posits the mechanism of HO representation to explain the TP and that creates the possibility of radical error. The TP is neutral about this possibility. It does not explicitly rule it out and its acceptance of mild misrepresentation suggests the possibility. But the issue of intensional context only arises when we consider a theoretical explanation of the data characterized by the TP; the TP itself remains in line with common sense.

An intrinsic conception, on the other hand, focuses on the qualitative "feel" of conscious experience: the redness of a red experience or the painfulness of conscious pain. Such a concept clearly captures something important about consciousness. It does intuitively seem that a central feature of consciousness is that it feels a certain way, that it is marked by special qualities. Defenders of the intrinsic conception further hold that the easy conceivability of inverted and absent qualia show that the qualitative feel of conscious experience is importantly independent from any other aspect of the mind. And these conceivability intuitions are said to be rooted in commonsense: normal folk will easily grasp the issue and find their intuitions drawn towards the intrinsic conception.

I do not contest that our everyday idea of consciousness is closely connected with the feel of conscious states, though I have argued that commonsense clearly accepts the presence of nonconscious mental states, even emotional or sensory states.

[36] See also Lormand (2004) and Hellie (2007) for a complex debate over Lormand's attempt to justify, by way of a complex linguistic analysis, a HO reading of Nagel's "what it's like" phrase. As I've indicated, I doubt that a linguistic analysis alone will be able to provide a definitive reason for favoring one interpretation. Instead, I believe that folk usage, *combined with broader theoretical considerations*, gives us reason to favor one reading over another.

Nor will I contest the claim that folk find inverted and absent qualia easily conceivable, though this is more controversial.[37] It seems to me that this is a fact that must be explained by a theory of consciousness, even if it such a theory ultimately rejects the idea that qualia *really* can come apart from all other mental processes. But in any event it seems clear that the intrinsic conception captures an important element of the folk conception. I contend that both TP and the intrinsic conception are in good standing in regards to folk psychology. Some ways of describing the mind or of describing various conceivable scenarios suggest one concept, while others favor the rival concept. There seems no good way to weight these scenarios so we can rule that one sort is more important than another; theorists will no doubt weight more heavily those scenarios leading to their favored concept. Instead, I purpose that the rival concepts both come off equally well in this context. Of course proponents of one concept must at least explain the appearances captured by the rival view, but so long as that is done, there seems little to decide between the two approaches here.

That leaves us with the second desideratum: how well a concept of consciousness meshes with empirical concerns. Here the TP has an advantage. The TP dictates that a theory of consciousness must explain how it is that we are aware of ourselves as being in mental states. This awareness is well-modeled by HO theories. Further, the sort of representation employed in HO theory is amenable to explanation in naturalistic terms, or so it is hoped. This provides a clear route to a naturalized theory of consciousness. This is in marked contrast to the data fixed by an intrinsic concept. And the TP has a range of applications in current empirical psychology and neuroscience. While the researchers themselves may not explicitly endorse the TP, a good case can be made that it is the guiding principle in a wide range of experimental paradigms focused on the differences between conscious and nonconscious processing, including research into implicit cognition, metacognition, blindsight, and change blindness, among others.[38] In addition, Block (2001) argues that the work of a number of empirical researchers, including Parvisi and Damasio (2001) and Jack and Shallice (2001), employ a notion of "reflexive consciousness," which amounts to the phenomenon picked out by the TP. Though Block goes onto argue that these researchers have missed the data fixed by an intrinsic concept, it's clear that these researchers themselves find the TP useful in their work. This is good evidence of empirical mesh.

The intrinsic concept, by contrast, is all but defined by its lack of mesh. Even Block, who attempts to establish an empirical connection, characterizes his intrinsic "phenomenal consciousness" as being independent of any causal, functional, or intentional notion. Indeed, he uses variations on the inversion and absent qualia thought experiments to introduce his concept. And David Chalmers contends that this sort of concept leads to what he calls the "hard problem of consciousness," the great difficulty of "locating" consciousness in a physical world.[39] Further, as Block acknowledges, even if we can justify an identity between our qualia and particular brain states, we would be at a loss to determine if aliens or human-like robots

[37] See Chalmers (1996, Chaps. 2, 3) and Kirk (2005), for example, on the conceivability of zombies.

[38] See, e.g., Merikle et al. (2001), Dienes and Perner (1996, 1999, 2004), and Weiskrantz (1997).

[39] Chalmers (1996).

possessed qualia, even if we had full knowledge of their internal physiology. He calls this the "harder problem of consciousness."[40] The hard and harder problems show that even if mesh is possible with an intrinsic concept, it is far less effective than the mesh offered by the TP. On this score, the TP has the advantage.

I conclude that the TP is well-justified as a concept of consciousness fixing the data a theory must explain. This is not to say, of course, that the conception couldn't be wrong. We might find, in the course of our empirical investigation that another concept, even an intrinsic one, is a better fit. This is an empirical question however, not one decided by a priori conceptual arguments. Finally, it may be that something completely unanticipated occurs in the course of empirical research, a conceptual revolution on the order of Einstein or Copernicus. In such a case, all bets are off—the TP may fall away as irrelevant or incorrect. But we will still need to explain the folk-psychological appeal of the TP—why, that is, it seemed pretheoretically that we are aware of our conscious states. In any event, for the purposes of this paper I have established both the conditional claim (*if* the TP fixes the data, then the misrepresentation objection fails) and its antecedent, with respect to the relevant intrinsic rivals. The misrepresentation objection therefore fails.

One lingering piece of business. It might be thought that we can embrace a revised version of the TP, one that does not let the misrepresentation objection get off the ground.[41] If such a characterization can be justified, we can avoid the entire worry. And this seems to be a desirable theoretical position, given the argumentative hackles raised by misrepresentation. If there's no need to even go near misrepresentation, why expend all the mental energy required for this defense? The revised version of the TP can be stated as follows:

> TP*: A subject's mental state is conscious when the subject is suitably aware of herself as being in that state and that state exists.

TP* seems to have all the advantages of the TP spelled out in this section with none of the alleged drawbacks: no worries about seeming to be in a conscious state when one is not, no fear of qualia "disappearing," no risk of the collapse of a two-tiered explanatory structure. But upon closer reflection, TP* does not offer a worry-free characterization of consciousness. It either collapses into the TP itself or it implicitly embraces an intrinsic concept of consciousness, with all the resulting explanatory difficulties.

TP* requires that the target of our awareness must exist if a mental state is conscious. But we can now ask, what happens if the awareness occurs in the absence of its target? That certainly seems like an open possibility, especially given the plausibility of mild misrepresentation. We can agree that the result will not be *called* consciousness, but that is clearly just a linguistic matter. We want to know if there will still be something it's like for the subject in this "pseudo-consciousness." There are two possibilities. First, it will seem to the subject that she is in the nonexistent state. That is, the mechanisms of inner awareness will make the subject aware *as if* the target state is present. If these mechanisms fully account for what it's

[40] Block (2002).

[41] See Kriegel (2003, 2006). See also Van Gulick (2001, 2004) and Gennaro (1996).

like for the subject, then from the subject's point of view it will seem she is in the nonexistent conscious state. The situation will be subjectively indistinguishable from the veridical case. But that leaves the proponent of TP* in *exactly* the same position as the proponent of the TP. We can ask all the same questions, raise all the same purported worries. Nothing has been gained by moving to the strengthened version of the TP.

The other possibility may therefore seem appealing. We can deny that it will seem to the subject that she is in the conscious state. But that raises the question of *why* this is so. And it is incumbent on the proponent of TP* to explain why this is so using only naturalistically acceptable means. The claim here is that the target state itself contributes something to conscious experience, that it is constitutively involved in the "phenomenal labor," to use Neander's term. But how could it do so unless there is something *intrinsic* to the target state itself accounting for its contribution? If there isn't something intrinsic to the state doing the work, it is still open for full-blown radical misrepresentation to occur. And claiming that it is something intrinsic to the target which *causes* the awareness mechanism to fire is insufficient. Given the independence of cause and effect, misfiring is still possible. So the intrinsic properties of the target state itself must play a *constitutive* role in consciousness. TP*, on this reading, is therefore an intrinsic conception of consciousness. It gives up the extrinsic game. A theory embracing TP* will then have to explain how it is that intrinsic properties of a mental state constitute consciousness; how they do so, that is, independently of any connections to other states or processes. Again, a theorist may feel that taking on these burdens is preferable to the alternative. But we have not been offered a nearby but theoretically improved version of the extrinsic TP conception. This is not a small change, a moderate emendation to the TP. It is a full-blown jettisoning of the basic idea of the TP.

It might seem there is room between the two extreme readings of TP* that I've sketched. Perhaps we can say that the target state only makes its contribution to consciousness when properly bound to the HO state.[42] But we are still lacking a satisfying story of *why* this occurs. If the process is truly extrinsic, we face the possibility of false positives: all extrinsic concepts face some version of this problem, as I argued in the previous section. If it is not, then we need an explanation of how the intrinsic properties contribute to consciousness and why they only do so when the proper binding occurs. Can such properties be detected empirically? If they are "stimulated" in the absence of any HO process, is there something it's like for the subject? Why not? Either the process can be fully characterized extrinsically, and so the possibility of misrepresentation remains, or it brings in an unexplicated intrinsic property. There is no room between these possibilities.[43]

Finally, we can ask why it is that one would be drawn to this sort of conception, beyond the desire to avoid worries of misrepresentation. Is it really evident from first-person reflection that we cannot misrepresent what state we are in? How would

[42] See Kriegel (2006). This may also capture what is intended by the "mixed" reading of Nagel's "what it's like" phrase, discussed in footnote 11 above.

[43] For a more complete defense of these claims, see Weisberg (2008).

this fact be manifest in experience? And if it is not manifest in experience, what intuitions are being saved by embracing this modified version of the TP? Do the folk think we can't ever be in error about what mental states we are in? Do they hold that there must really *be* something intrinsic to conscious experience, as opposed to there merely *seeming* to be something intrinsic? I contend that the folk do not have psychological intuitions this detailed and conclusive. Instead, our folk-psychological conception is committed (often implicitly) to a range of platitudes and those platitudes fail to provide anything like this clear a conclusion about the need to avoid misrepresentation. And if the misrepresentation worry is not a problem for the TP, we have no other good reason to attempt a strengthening of the TP. The TP captures an important aspect of our folk psychology and it provides a fruitful mesh with empirical research. Only if we misrepresent consciousness as being intrinsic do we have any worry about misrepresentation.

Acknowledgments Distant ancestors of this paper were presented at the SPP, Stanford, 1999, the conference on self-representational approaches to consciousness, University of Arizona, 2005, and the CUNY Cognitive Science Symposium and Discussion Group. My thanks to those audiences for helpful discussion. Thanks to Jared Blank, Ned Block, Alex Byrne, Gregg Caruso, Rocco Gennaro, Gil Harman, Uriah Kriegel, Pete Mandik, Liz Vlahos, and especially David Rosenthal. Thanks also to an anonymous reviewer for Philosophical Studies.

References

Armstrong, D. M. (1968). *A materialist theory of the mind*. New York: Humanities Press.

Armstrong, D. M. (1980). *The nature of mind and other essays*. Ithaca: Cornell University Press.

Baars, B. J. (1997). *In the theater of consciousness: The workspace of the mind*. Oxford: Oxford University Press.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2). (Reprinted from *The nature of consciousness: Philosophical debates*, pp. 375–416, by N. Block, O. Flanagan, & G. Güzeldere, Eds., 1997, Cambridge, MA: The MIT Press.)

Block, N. (1996). Mental paint and mental latex. *Philosophical Issues, 7*, 19–49.

Block, N. (2001). Paradox and cross purposes in recent work on consciousness. *Cognition, 79*(1–2), 197–219.

Block, N. (2002). The harder problem of consciousness. *The Journal of Philosophy, 11*(XCIX), 391–425.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30*(5–6), 481–499.

Byrne, A. (1997). Some like it HOT: Consciousness and higher-order thoughts. *Philosophical Studies, 86*(2), 103–129.

Byrne, A. (2004). What phenomenal consciousness is like. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 203–225). Amsterdam: John Benjamins Publishers.

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

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

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*, 1–37.

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

Dienes, Z., & Perner, J. (1996). Implicit knowledge in people and connectionist networks. In G. Underwood (Ed.), *Implicit cognition* (pp. 227–256). Oxford: Oxford University Press.

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences, 22*, 735–755.

Dienes, Zoltan., & Perner, Josef. (2004). Assumptions of a subjective measure of consciousness: Three mappings. In R. J. Gennaro (Ed.), *Higher order theories of consciousness* (pp. 173–199). Amsterdam: John Benjamins Publishers.

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

Gennaro, R. J. (1996). *Consciousness and self-consciousness: A defense of the higher-order thought theory of consciousness*. Amsterdam: John Benjamins Publishers.

Harman, G. (1990). The intrinsic quality of experience. In J. Tomberlin (Ed.), *Philosophical Perspectives: 4 action theory and the philosophy of mind*. Atascadero, CA: Ridgeview Publishing Co.

Hellie, B. (2007). 'There's something it's like' and the structure of consciousness. *Philosophical Review, 116*, 441–463.

Huemer, M. (1998). *A direct realist account of perceptual awareness*. Unpublished dissertation, Rutgers University, Graduate Program in Philosophy. Accessed from http://home.sprynet.com/∼owl1/dis.htm.

Jack, A. I., & Shallice, T. (2001). Introspective physicalism as an approach to the science of consciousness. *Cognition, 79*, 161–196.

Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Clarendon.

Kirk, R. (2005). *Zombies and consciousness*. New York: Oxford University Press.

Kriegel, U. (2003). Consciousness as intransitive self-consciousness: Two views and an argument. *Canadian Journal of Philosophy, 33*, 103–132.

Kriegel, U. (2006). The same-order monitoring theory of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness*. Cambridge, MA: MIT Press/Bradford Books.

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

Lewis, D. K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, L*(3), 249–258. (Reprinted from *Papers in metaphysics and epistemology*, pp. 248–261, by D. K. Lewis, Ed., Cambridge: Cambridge University Press.)

Lormand, E. (2004). The explanatory stopgap. *Philosophical Review, 113*, 303–357.

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

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

Lycan, W. G. (2001). A simple argument for the higher-order representation theory of consciousness. *Analysis, 61*(1), 3–4.

Mandik, P. J. (2009). Beware of the unicorn: Consciousness as being represented and other things that don't exist. *Journal of Consciousness Studies, 16*(1), 5–36.

McGinn, C. (1989). Can We Solve the Mind-Body Problem. *Mind, 98*, 891. (Reprinted from *The nature of consciousness: Philosophical debates*, pp. 529–542, by N. Block, O. Flanagan, & G. Güzeldere, Eds., 1997, Cambridge, MA: The MIT Press.)

McGinn, C. (1999). *The mysterious flame: Conscious minds in a material world*. New York: Basic Books.

Merikle, P. M., Smilek, D., & Eastwood, J. D. (2001). Perception without awareness: Perspectives from cognitive psychology. *Cognition, 79*, 115–134.

Nagel, T. (1974). What is it like to be a bat. *Philosophical Review, 83,* 435–445. (Reprinted from *The nature of consciousness: Philosophical debates*, pp. 519–527, by N. Block, O. Flanagan, & G. Güzeldere, Eds., 1997, Cambridge, MA: The MIT Press.)

Nagel, T. (1986). *The view from nowhere*. New York: Oxford University Press.

Neander, K. (1998). The division of phenomenal labor: A problem for representational theories of consciousness. In J. E. Tomberlin (Ed.), *Philosophical perspectives 12: Language mind and ontology* (pp. 411–434). Boston: Blackwell Publishers.

Parvisi, J., & Damasio, A. (2001). Consciousness and the brainstem. *Cognition, 79*(1–2), 135–160.

Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49*(3), 329–359.

Rosenthal, D. M. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: Philosophical debates* (pp. 729–753). Cambridge, MA: The MIT Press.

Rosenthal, D. M. (2000). Consciousness and metacognition. In D. Sperber (Ed.), *Metarepresentation: Proceedings of the tenth Vancouver cognitive science conference*. New York: Oxford University Press.

Rosenthal, D. M. (2002). How many kinds of consciousness. *Consciousness and Cognition, 11*(4), 653–665.

Rosenthal, D. M. (2004). Varieties of higher-order theory. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 19–44). Amsterdam: John Benjamins Publishers.

Rosenthal, D. M. (2005). *Consciousness and mind*. Oxford: Clarendon Press.
Siewert, C. (1998). *The significance of consciousness*. Princeton, NJ: Princeton University Press.
Tye, M. (1995). *Ten problems of consciousness*. Cambridge, MA: The MIT Press.
Van Gulick, R. (2001). Inward and upward: Reflection, introspection, and self-awareness. *Philosophical Topics, 28*, 275–305.
Van Gulick, R. (2004). Higher-order global states (HOGS): An alternative higher-order model of consciousness. In R. J. Gennaro (Ed.), *Higher-order theories of consciousness* (pp. 67–92). Amsterdam: John Benjamins Publishers.
Weisberg, J. (2008). Same old, same old: The same-order representation theory of consciousness and the division of phenomenal labor. *Synthese, 160*(2), 161–181.
Weiskrantz, L. (1997). *Consciousness lost and found: A neuropsychological exploration*. Oxford: Oxford University Press.