Conscious-state anti-realism

1. Introduction

One of the central theses that Daniel Dennett defends in his work on consciousness is a kind of anti-realism about consciousness, what Dennett calls "first-person operationalism," a thesis that "brusquely denies the possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness" (Dennett, 1991)(p. 132). One of Dennett's most famous arguments toward this conclusion appeals to the alleged empirical underdetermination of theory-choice between "Stalinesque" and "Orwellian" explanations of certain temporal anomalies of conscious experience (pp. 115-126). The explanations conflict over whether the anomalies are due to misrepresentations in memories of experiences (Orwellian) or misrepresentations in the experiences themselves (Stalinesque).

David Rosenthal has offered that his Higher-order Thought theory of consciousness (hereafter, "HOT theory") can serve as a basis for distinguishing between Orwellian and Stalinesque hypotheses and thus as a basis for resisting first-person operationalism (hereafter, "FPO")(Rosenthal, 1995, 2005a, 2005b). The gist of HOT theory is that one's having a conscious mental state consists in one having a higher-order thought (a HOT) about that mental state. (Such a HOT must also not be apparently arrived at via a conscious inference, but this further constriction on the HOTs that matter for consciousness is of little importance to the present paper.) I'll argue that HOT theory can defend against FPO only on a "relational reading" of HOT theory whereby consciousness consists in a relation between a HOT and an actually-

existing mental state. I'll argue further that this relational reading leaves HOT theory vulnerable to objections such as the Unicorn Argument (AUTHOR, 2009). To avoid defend against such objections, HOT theory must instead admit of a "nonrelational reading" whereby a HOT alone suffices for a conscious state. Indeed, HOT theorists have been increasingly explicit in emphasizing this nonrelational reading(Rosenthal, 2011)(Weisberg, 2011)(Weisberg, 2010). However, I'll argue, on this reading HOT theory collapses into a version of FPO.

The remainder of the paper will go like this: In section 2 I'll say some more about anti-realism, FPO, and the Orwellian/Stalinesque argument. In section 3 I'll lay out a HOT-theoretic version of the Orwellian/Stalinesque distinction that depends on a relational reading of HOT theory. In section 4 I'll spell out the case for a nonrelational reading of HOT theory and how HOT theory is thereby led to a kind of FPO.

2. Anti-realism, consciousness, and FPO

2.1. Clarifying consciousness anti-realism

In this subsection I want to rapidly clarify key terminology ("anti-realism" etc.) and key theses relating it to consciousness (FPO etc). My aim in the present section is not to argue that one set of construals is better than another, but instead to lay out a series of stipulations to facilitate the rest of the discussion.

Consciousness aside for a moment, let's think about the general structure of realism/anti-realism theses and debates between them. A realist position, say realism about dogs, is a conjunction of an existence claim and an independence claim, where the independence in question is often glossed as "mind independence". An imprecise statement of dog realism is

"dogs exist and exist mind-independently." Each conjunct admits of multiple precisifications. I'll have little to say in the present paper about precisifications of the existence claim. Let it suffice that I intend existence claims to be tenseless and actual-world directed. So, items in the past and future exist, though no item in a nonactual possible (or impossible) world does. The extinction of dogs won't, then, falsify dog realism.

Precisifications of the independence claim require more care, especially if we want to formulate coherent claims of mind-independence about things that are themselves mental. One precisification of independence that will not serve present purposes is one stated simply in terms of minds: X exists independently of any mind existing. Clearly, plugging "minds" in for "X" generates an incoherence. Precisifications that avoid such incoherence appeal instead to specific kinds of mental state, say specific kinds of thought, belief, or judgment. "Minds exist independently of anyone thinking, believing, or judging that minds exist" contains no obvious incoherence. Precisifications of the independence claim along this line will be the key ones I have in mind for the rest of the paper. Of interest will be the question of whether one's conscious experience exists independently of one's thinking, believing, or judging it to exist.

Given that realist theses are each a conjunction of an existence claim and an independence claim, opponents of realism come in two varieties: Nihilists, who deny the existence claim, and idealists, who deny the independence claim. A Berkeleyan idealist about dogs (a "bark"-leyan?) does not deny that dogs exist, but instead denies that dogs exist independently of being perceived.

I will simply set nihilism aside in this paper, and reserve "anti-realism" for the idealist variety. While Dennett's critics sometimes accuse him of denying that consciousness exists, it should be clear that Dennett's statement of FPO doesn't support such a reading. In denying "the

possibility in principle of consciousness of a stimulus in the absence of the subject's belief in that consciousness," Dennett is clearly not denying an existence claim, but instead an independence claim. The kind of anti-Dennettian that I am interested in can be briefly described as holding that we can roughly sort mental states into two varieties, roughly, experiences and thoughts, and that states (and facts about them) of the first variety obtain independently of states of the second variety.

One further set of issues I want to address before leaving this subsection concerns which facts about consciousness are at issue. What we get directly from the Dennett quote is that FPO is anti-realist about "consciousness of a stimulus". Some consciousness theorists, especially HOT theorists, will detect an ambiguity in this phrase. Many, if not all, follow Rosenthal in distinguishing "transitive consciousness" (being conscious of something) from "state consciousness" (a mental state's being conscious). If there is such a distinction, then the possibility opens of having a state in virtue of which one is conscious of something, for instance, a perceptual state in which one is conscious of a red rose, without that state itself being a conscious state. Other theorists do not urge such a distinction, and perhaps (though I'm unsure) Dennett counts among them. However, regardless of where one stands on this issue, there is an interesting anti-realist thesis to be stated explicitly in terms of state consciousness. Modifying the Dennett quote accordingly yields a thesis that "brusquely denies the possibility in principle of *a conscious experience of a stimulus* in the absence of the subject's belief in that consciousness" (altered text italicized). For the remainder of the paper, I shall be interpreting FPO as including this thesis.

2.2. The Orwellian/Stalinesque argument for FPO

The phi phenomenon is a species of illusory motion, as when one views the flashing stationary lights on a marquee. Color phi is a species of the phi phenomenon in which the stationary stimuli differ in color and the apparently moving object changes color mid-trajectory. Subjects in a color phi experiment look at a computer screen upon which a green circle appears then disappears. A small time later in a position a small distance away from where the green circle was a red circle of the same size appears and then disappears. The time elapsed between the disappearance of the green and the appearance of the red is very short. It's so short that, as a subject in this experiment, it would appear to you as if a single circle appears, moves across the screen, and then disappears. Further, the single moving circle would appear to start off green and change to red midway in its trajectory. This is color phi and it is weird.

Color phi is not just weird because we don't know how the brain creates illusory motion from nonmoving stimuli. Here's the really puzzling thing about color phi: How does the brain know to change the moving green circle to red before the red circle appears? Clairvoyance aside, clearly it cannot. So the experience of the red-to-green change needs to have happened after the brain receives information of the appearance of the red circle. We want further details in explaining this, and here we feel pulled toward two competing explanations, explanations that Dennett famously dubs "Orwellian" and "Stalinesque".

My mnemonic for Dennett's labels is that "Stalinesque " shares an "s" and a "t" with "show trial", and "Orwellian" has an "r" in common with "revisionist history". Both explanations have key roles for the notions of consciousness and of falsehood, but differ with respect to which states are conscious and which ones are false representations. Let's start by looking at the revisionist history, that is, the false memory, posited by the Orwellian explanation. On this explanation, the key mental events and their temporal order are as follows: First there's a

conscious experience of a green circle, next there is a conscious experience of a red circle, and finally there is a false memory of a single circle having moved and changed from green to red. On the Orwellian explanation there is neither a conscious experience of motion nor one of color change, but instead a false memory that movement and color change were experienced.

Let us turn now to the Stalinesque explanation, which posits a show trial. On this explanation, the false mental state posited is not a memory but an experience. On the Stalinesque explanation, the key mental events and their temporal order are as follows: First there is an unconscious receipt of information concerning the green circle, next there is an unconscious receipt of information concerning the red circle, and finally, based on these raw materials, a conscious experience is assembled—a false experience of a green circle moving and changing to red mid-trajectory.

On the face of it, these seem to be distinct competing explanations of the empirical data. The Orwellian explanation posits two accurate conscious experiences of two stationary, differently colored circles followed by a false memory of having experienced a single moving circle that changes color. The Stalinesque explanation posits a false conscious experience of motion and mid-trajectory color-change and an accurate memory of that experience. To highlight their differences: Orwellian posits a false memory and accurate conscious experience, whereas Stalinesque posits a false conscious experience and an accurate memory (of what the experience was).

If these are indeed distinct explanations, then which one is the correct one? Dennett argues persuasively that no amount of evidence, either first-personal or third-personal, will determine theory choice here. I'm persuaded. I find it easy to be so persuaded.

To attempt to persuade yourself of Dennett's conclusion, first imagine being a subject in a color phi experiment. What you introspect is that there's been a visual presentation of a moving, color-changing circle. Your introspective judgment is that you've experienced such an episode. But to resolve the Stalinesque v. Orwellian debate on introspective grounds, your introspective judgment would need to wear on its sleeve whether its immediate causal antecedent was a false memory (Orwellian) or a false experience (Stalinesque). But clearly, no such marker is borne by the introspective judgment. So much for the first-person evidence!

So now, imagine being a scientist studying a subject in a color phi experiment. Imagine availing yourself of all of the possible third-personal evidence. Suppose you avail yourself to evidence gleaned via futuristic high-resolution (both spatially and temporally) brain scanners. Such evidence, let us suppose, will allow you to determine not only which brain events occur and when, but also which brain events carry what information, and which brain events are false representations. This is, of course, to presume solutions to very vexing issues about information, representation, and falsehood (solutions that might beg the question against a Dennettian anti-realism about representation and perhaps, thereby, against Dennettian anti-realism about consciousness, but I won't pursue this line of thought). However, we will here suppose that such solutions can be arrived at independently of resolving issues about consciousness. Clearly, then, the evidence that you have will, by itself, tell you nothing about which states are conscious. So much for the third-person evidence!

To surmount this hurtle for strictly third-person approaches, you may feel tempted to either ask the subject what their conscious experiences are like, or allow yourself to be a subject in this experiment. However, either way you will only gain access to an introspective judgment with a content that we have already seen as underdetermining the choice between Orwellian and

Stalinesque.

Given that there's no real difference between the Orwellian and Stalinesque scenarios, what matters for consciousness is what they have in common, namely the content of the belief or thought that one underwent a conscious experience of a color-changing moving circle. There's nothing independent of this belief content that serves to make it true, so having a belief with such-and-such content is all there is to being in so-and-so conscious state.

3. HOT Orwellian and HOT Stalinesque scenarios

Dennett's Orwellian/Stalinesque argument turns on a kind of underdetermination of theory by evidence. Of course, what evidence underdetermines, additional theory can sometimes settle. Rosenthal constructs HOT-theoretic versions of the Orwellian and Stalinesque scenarios that are distinguishable given the resources of HOT theory (Rosenthal, 1995) (p. 362). However, that there are *some* Orwellian and Stalinesque scenarios that are distinguishable from each other doesn't suffice to refute FPO. Dennett himself admits that some Stalinesque scenarios are distinguishable from some Orwellian scenarios (especially at macroscopic time-frames) (Dennett, 1991)(p. 117). What matters instead is that there are some Orwellian and Stalinesque scenarios that are not distinguishable from each other. I aim in the present section to show that there are Orwellian and Stalinesque scenarios that HOT theory serves to distinguish only on a relational reading of HOT theory.

One way to convey the gist of HOT theory is by saying that a state is conscious when a HOT is about the state. Reading this relationally, we have two relata and a relation between them. The relata are the HOT and the state that it is about. The relation the HOT bears to its target is an

"aboutness" relation, or as I'll prefer to say, a "representing relation". When a visual experience of a red circle is accompanied by a HOT that bears the representation relation to it, then the visual experience is a conscious one. If, instead, the visual experience is unaccompanied by any such HOT, the experience is an unconscious one. Sometimes HOT theorists themselves put HOT theory in ways that invite the relational reading. For example, Rosenthal (Rosenthal, 2005b) writes that his is "a theory according to which a mental state is conscious just in case it is accompanied by a higher-order thought (HOT) to the effect that one is in that state" (p.322). Perhaps, in the final analysis, Rosenthal's commitment to the relational reading may be merely a superficial appearance. I'll return to this issue in section 4. For the present section, I will keep the relational reading at the forefront.

With this relational reading of HOT theory in mind, let us think through how color phi can be explained. In color phi, it seems to one that one has an experience of a moving circle that changes color. In order for it to seem to one that one is having an experience of a moving, color-changing circle, there needs to be a HOT that has content along the lines of that one is visually experiencing a moving, color-changing circle. We might wonder further about what the causal antecedents are of this HOT, especially as concerns links in the causal chain after the information from the stationary flashing circles has hit the eye of the beholder.

One possibility is that none of the causal antecedents of the HOT is a visual experience of motion and color change. Instead, the causal antecedents are visual experiences of the stationary red and green circles. Further, it is a consistent elaboration on this possible scenario that no causal consequence of the HOT is a visual experience as of motion and color change. Since nothing antecedent or consequent to the HOT answers to the description that constitutes the HOTs content, the HOT is false. Since the HOT is not itself an experience (it is instead a

thought) and has occurred after the experiences that triggered its occasion, we can regard it as a memory (albeit, a false one). Given the possibility we've just consistently described, this reading of the HOT theory casts it as close to Orwellian. However, to be fully Orwellian, there needs to be posited, in addition to a false memory, an accurate conscious experience. Can we complete an Orwellian explanation sketch that's consistent with HOT theory? I think that we can, but some care needs to be taken.

The way to introduce an accurate conscious experience into the above sketch in a way that is consistent with HOT is to go looking for one or more states that the HOT is about. If this sketch is to be Orwellian, some choices for what the HOT is about will be better than others. On a highly natural reading of what the HOT is about, it is about an inexistent state, namely a visual experience of motion and color change. The inexistence of such a state is what makes the HOT false. One problem with this reading is that the Orwellian is supposed to be positing the *existence* of a conscious state, and it is highly strained to posit the existence of something that is admitted in the same breath to not exist. I hope I will be forgiven in dismissing the Meinongian perspective required to view existing inexistents as welcome company. Anyway, there's another problem: It is difficult to regard the inexistent state as accurate. The inexistent state is a representation of movement and color change upon the computer screen, and, in actuality, no such motion or color change exists. And since Meinongianism is here not taken seriously, there's no serious way of taking the suggestion that the inexistent state is an accurate representation, albeit one that accurately represents an inexistent state of affairs.

There is another possibility for interpreting what the HOT is about, namely that it is about the two separate experiences of the differently colored circles. In being about those accurate experiences, they are thereby rendered conscious: On the occasion of the HOT about them, the

experiences become conscious. This may have a slight air of strangeness, but there's no obvious problem in a representation of something representing it falsely. Indeed, the scenario described here is a possibility that Rosenthal explicitly endorses (Rosenthal, 2005a)(pp. 240-241). (That is, he endorses it as a possibility. He does not assert that it is an actuality.)

Thus completes my sketch of a HOT Orwellian explanation of color phi. Let's try to fit a Stalinesque explanation into the HOT mold as follows:

Recall that a Stalinesque explanation posits a false conscious experience of motion and mid-trajectory color-change that has as causal antecedents the unconscious receipt of information concerning the stationary presentations of the green circle and the red circle. To fit such an explanation into the HOT mold, the HOT theorist needs to posit a HOT that is about an experience that is itself (the experience) a false representation of motion and color change. Otherwise, without such a HOT, the false experience won't be conscious. But in order to introduce this HOT, a means must be devised of determining that the HOT is about the false representation and not about the accurate representations. Otherwise, the accurate representations will be the conscious ones and this won't be Stalinesque. Supposing that this can be determined, we therefore have a Stalinesque reading of a HOT-theoretical explanation of color phi.

It looks, at least prima facie, that HOT theory is consistent with Orwellian and Stalinesque explanations. However, once these explanations are fit into the HOT mold, are opportunities thereby made available for adjudicating between them?

Note the key similarities in the Orwellian and Stalinesque stories. See fig 1. On both stories there is a HOT, the content of which is that there's an experience of motion and color change. Also, on both stories there are accurate experiences of the stationary red and green circles. The

key differences are that, on the Orwellian story, the HOT bears the representation relation to the accurate experiences and not to the (inexistent) inaccurate experience of motion and color change. On the Stalinesque story, the HOT bears the representation relation to the inaccurate experience of motion and color change and not to the accurate experiences of the stationary red and green circles. If we assume that the HOT theory is true, then in order to discover whether color phi is Orwellian or Stalinesque we would need to discover whether the HOT bore a representing relation to the accurate experiences or not.

Orwellian HOT story:

(acc green exp)&(acc red exp) (inexist. motion exp) [false HOT]

!.....rep rel…….……...…!

Stalinesque HOT story:

(acc green exp)&(acc red exp) (motion exp) [acc HOT]

!…..rep rel....!

fig 1. On the Orwellian story, a false HOT bears a consciousness-conferring representation relation ("rep rel") to the accurate experiences of the stationary red and green circles. On the Stalinesque story an accurate HOT bears a consciousness-conferring representation relation to the inaccurate experience of motion and color change. Both stories contain accurate experiences of the stationary circles. And both stories contain HOTs the contents of which are that one is experiencing motion and color change. The stories differ in which relata the HOT bears the representation relation to.

To give a preview of the worry that I ultimately want to press against HOT theory, there are good reasons to think that there is no such thing as a representation relation and so, if the HOT theory is true, no such relation figures in it. But without recourse to such a relation, there's no relevant difference between the HOT Orwellian and the HOT Stalinesque explanations: On either case, the content of one's consciousness just is the content of the HOT, and that content is the same on either story.

4. Non-relational HOT theory and FPO

In (AUTHOR, 2009) I press an argument, "the Unicorn Argument" or just "the Unicorn," against HOT theories. At the heart of the argument is a view about how best to think of representation in the face of the representation of inexistents such as unicorns. This view can be seen as emerging as a response to the famous inconsistent triad of intentionality. One way of presenting the triad is like this:

1. Representing is a relation borne to that which is represented.

2. There are representations of inexistents.

3. There are no relations borne to inexistents.

While all three propositions of the triad are independently plausible, they cannot be jointly true. The heart of the Unicorn involves a denial of the first item in the triad while retaining the last two. The resulting view might be summed up as holding that there is no such thing as a representing relation: Representation may involve relations, but it is not constituted by a relation to that which is represented. It follows from there being no representation relation that there is no such relational property as the property of being represented.

This line of thought is pressed against the HOT theory by reading hot theory as committed to the existence of such relations and relational properties. On what I'll call the "relational reading" of HOT theory, a state is conscious only if a HOT bears the representation relation to the state. On this reading of HOT theory, the property of being conscious just is the property of being represented by a HOT. But if there are no such relations and relational properties, and there is

such a property as a state's being conscious, then this cannot be what a state's being conscious consists in.

HOT theorists often present the view in a way that seems to invite the relational reading. However, in responding to the Unicorn and closely related objections turning on "empty" higher thoughts (e.g. (Byrne, 1997)(Neander, 1998)(Block, 2011)), HOT theorists have urged a reading of their view that I'll call the "non-relational reading".

Weisberg (2010), in responding to the Unicorn, cites approvingly a remark of Harman's (1997), part of which includes the statement "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." Rosenthal (Rosenthal, 2011) writes, in response to Block's (2011) attack based on empty HOTs:

> Block describes me as having retreated from an 'aboriginal' theory, on which the targets of HOTs always exist, to a 'new version' on which they need not …. This is not so; in my earliest publication about consciousness I noted the possibility of absent first-order states …. For ease of exposition, I often introduce the theory by saying that a state is conscious when it's accompanied by a HOT, noting that this characterization is not strictly accurate. And there's no harm in putting things in those relational terms when the existence of HOTs' targets is not under consideration. All that matters for a state's being conscious is its seeming subjectively to one that one is in that state. On the HOT theory, that's determined by a HOT's intentional content….(p. 436).

With this nonrelational reading of HOT theory in mind, it becomes overwhelmingly difficult to see how HOT theory isn't just a version of FPO. In publications attacking FPO, Rosenthal

describes FPO as, among other things, a view whereby "facts about…when states become conscious are exhausted by how things appear to consciousness" (Rosenthal, 2005b, p.323). Note how similar such a description of FPO is to Rosenthal's own description of HOT theory in publications highlighting its invulnerability to empty-HOT based attacks: "A state's being conscious is a matter of mental appearance - of how one's mental life appears to one."(Rosenthal, 2011, p. 431).

I find it hard to shake the impression that there's a tension within HOT theory itself between a relational reading and a nonrelational reading. Further it seems that the nonrelational reading is highlighted when defending against empty-HOT and Unicorn types of objections and that the relational reading is highlighted when defending against FPO. In a publication targeting FPO Rosenthal (1995) seems himself to be promoting a relational reading of HOT theory:

> Because many mental states aren't conscious at all, it's implausible that the property of being conscious is an intrinsic property. All mental states have some sort of content properties—intentional content in the case of intentional states and sensory content in the case of bodily and perceptual sensations and most emotions. Such content properties are arguably intrinsic to mental states. By contrast, mental states can be conscious at one moment and not at another; so we have no reason to regard the property of being conscious as being intrinsic to such states. Accordingly, a state's being conscious requires the occurrence of something extrinsic to it. And it may well be, therefore, that no mental state is conscious when it first occurs. But this doesn't mean there are no facts of the matter about consciousness; states are conscious when, and only when, the relevant events occur. (p. 364)

If there's a way to resolve the apparent tension between relational and nonrelational readings of HOT theory, I do not know what it is. I do hope, though, that the present paper aids in

progress toward its resolution. In the meantime, it seems to me that the balance is tipped toward a nonrelational reading and thus, if my arguments are correct, a reading of HOT committing it to FPO.

References

Block, N. (2011). The higher order approach to consciousness is defunct. *Analysis*, *71*(3), 419–431. doi:10.1093/analys/anr037

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

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown and Company.

Harman, G. (1997). The Intrinsic Quality of Experience. In N. Block, O. J. Flanagan, & G. Guzeldere (Eds.), *The Nature of Consciousness* (pp. 663–675). Cambridge, MA: MIT Press.

AUTHOR (2009).

Neander, K. (1998). The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness. *Nous*, Nous, *32*(S12), 411–434.

Rosenthal, D. (1995). Multiple drafts and facts of the matter. In T. Metzinger (Ed.), *Conscious Experience* (pp. 359–372). Exeter: Imprint Academic.

Rosenthal, D. (2005a). First-person operationalism and mental taxonomy. In *Consciousness and Mind* (pp. 229–256). Oxford: Oxford University Press.

Rosenthal, D. (2005b). Content, Interpretation, and Consciousness. In *Consciousness and Mind* (pp. 321–335). Oxford: Oxford University Press.

Rosenthal, D. (2011). Exaggerated reports: reply to Block. *Analysis*, *71*(3), 431–437. doi:10.1093/analys/anr039

Weisberg, J. (2010). Misrepresenting consciousness. *Philosophical studies*, *154*(3), 409–433. doi:10.1007/s11098-010-9567-3

Weisberg, J. (2011). Abusing the notion of what-it's-like-ness: A response to Block. *Analysis*, *71*(3), 438–443. doi:10.1093/analys/anr040