Comments on Pete Mandik, "Conscious-state Anti-realism" / Consciousness Online 4

Alex Kiefer

I think Pete has raised a serious challenge to HOT theory in this paper, not unconnected of course to recent challenges it has faced from Ned Block (2011) and from Pete himself (2009). Whatever the outcome of this discussion, I believe that Pete's paper will help to clarify HOT theory's commitments, by encouraging its proponents to spell out how, exactly, it differs from Dennett's First-Person Operationalism (if it does differ) once a relational understanding of HOTs' representing is abandoned.

As I see it, Pete raises two interesting but, importantly, distinct questions about what he calls the "non-relational reading" of HOT theory. The first is whether it constitutes a form of anti-realism about conscious states, and the second is whether it collapses into First-Person Operationalism (FPO). Of course, if you take conscious-state anti-realism to be definitive of FPO, then the distinction between these two questions is of little interest. But I think the core claim of FPO is stronger than the anti-realism thesis Pete discusses. Thus, it is possible (and will shortly be actual) for me to argue that non-relational HOT theory is likely a form of anti-realism in Pete's sense, and yet is distinct from FPO. First I'll discuss the anti-realism issue, and concede that non-relational HOT theory constitutes an anti-realist position in the sense in question. Then, I'll argue that this theory can nonetheless distinguish in principle between the Orwellian and the Stalinesque explanations of the color phi phenomenon. Finally, I'll conclude with some reflections on the differences between FPO and HOT theory as I understand them.

Conscious-state anti-realism

According to Pete's stipulation, a conscious state is real only if it (a) exists, (b) independently of anyone's judging or thinking that it does. The type of conscious-state anti-realist in which Pete is interested denies (b), the independence claim. I think it's clear that Dennett is an anti-realist in this sense. But I think it's also fairly clear that Rosenthal and other proponents of HOT theory are anti-realists in this sense, whether or not HOT theory is construed as involving a representation-relation. This is because what *makes* a state a conscious state, according to HOT theory, is a person's assertoric thought (i.e. judgment, belief) that he or she is in that state. Assuming that judging oneself to be in a particular state involves the judgment that there is such a state, then, HOT theory is committed to the claim that there are no conscious states without judgments or thoughts to the effect that they exist (albeit not normally judgments that they exist *qua* conscious), so HOT theory is a version of anti-realism even on the relational reading.

Still, I think the point Pete is after in his paper has more bite than this. A HOT theorist who wants to resist the anti-realist label might argue that even if no conscious states exist without our judging/thinking that they do, this doesn't mean that the existence of a conscious state is *just* a matter of such judgments, since there are facts about conscious states besides those that stem from our judgments about them (for example, facts involving properties that those types of states have whether or not they occur consciously.) If this is the case, it might be more accurate to describe HOT theory as involving the claim that a state's property of being conscious is judgment-dependent than the claim that conscious states are not real. But this is so only if we *can* identify some properties that conscious states have whether or not they are

conscious, and this we arguably can't do if we adopt a non-relational reading of HOT theory. Once we grant that HOTs do not bear representation-relations to their target states, the argument goes, it becomes arbitrary to identify any actually occurring brain state with the notional state (the state we think of ourselves as being in). Thus all there ever is to being in a conscious state is having a HOT with certain contents, which sounds like a full-fledged idealism (and thus anti-realism) about conscious states. It also sounds a lot like Dennett, as Pete notes, but I'll come back to that shortly.

The end of the last paragraph is a much-compressed version of an argument that Pete gives in his Unicorn paper (p14-15). I think that this argument can be resisted to some extent: even if we forsake the idea of a representation-relation, it needn't be entirely arbitrary to identify some actual brain state as the state that a given HOT is about, and this is all that's needed to accommodate our loose talk of states' being represented. But I won't argue that point here: if you are a non-relational HOT theorist, then you accept that facts about consciousness depend entirely on the contents of our higher-order thoughts, and the existence (or not) of the states we believe ourselves to occupy is theoretically inert. So the interesting thing about conscious states, at any rate (i.e. consciousness) is entirely judgment-dependent.

The Orwellian/Stalinesque distinction

I have granted that HOT theory is committed to anti-realism in Pete's sense, and so long as it's kept in mind what "anti-realism" means here, I do not see the harm in this. But realism aside, the more interesting issue raised in Pete's paper seems to me to be the question of whether non-relational HOT theory collapses into a version of FPO, on the basis of an inability to

distinguish between Orwellian and Stalinesque scenarios of the kind considered by Dennett. Pete argues that if HOT theory could distinguish the Orwellian from the Stalinesque explanation of the color phi phenomenon, it would have to be on the basis of which states the relevant HOTs bear representation-relations to, and since non-relational HOT theory can't appeal to such relations, it can't make the distinction.

This argument depends on the way Pete reconstructs the two competing explanations of color phi in HOT-theoretic terms. On Pete's account, there is one relevant HOT in each scenario: in the Orwellian case a false memory that inaccurately represents two accurate experiences of stationary dots, and in the Stalinesque case an accurate HOT that represents an experience of motion and color change which is itself a misrepresentation of the stimuli. Since the HOTs in the two scenarios are indistinguishable in terms of their (if you like, "non-relational") representational contents, and there is, on the non-relational reading of HOT theory, nothing *to* consciousness beyond the contents of HOTs, Pete concludes that non-relational HOT theory must treat the Orwellian and the Stalinesque scenarios as identical.

Pete's reconstruction focuses, naturally enough, on the conscious state that is reported by the subject in the color phi experiment: that of seeing a green dot travel across the screen and change color to red along the way. But there may be more HOTs here than meet the eye (or, rather, ear). The Orwellian scenario posits two conscious experiences followed, rapidly, by a false memory of those experiences. Although the details are tricky, I believe this provides enough wiggle room to keep HOT theory afloat and distinct from FPO, given certain assumptions to be discussed below. I will first introduce an alternative way of reconstructing Dennett's two scenarios within the framework of HOT theory that does allow for discoverable

differences between the Stalinesque and the Orwellian, and then argue that this alternative is at least as satisfactory a way of making the distinction as Pete's.

I follow Pete in imagining that we have access to the resources of future neuroscience, including spatially and temporally high-resolution brain scanners. I also assume, as Pete does, that we can tell what information each brain state so identified carries, and whether each state is a false representation. Pete proposes to put aside potential difficulties and possible begged questions about consciousness involved in clarifying the notions of representation, etc., and I'll do so for now as well. Now, if we can tell what information each brain state carries and whether it is a misrepresentation, presumably we can tell (to some degree of determinacy, anyway) *what* each state represents. One might think, then, that one could use HOT theory to distinguish Dennett's two scenarios in the following way. On the Orwellian scenario, one would expect to find, very shortly after the presentation of the green dot: (a) a perceptual state carrying information about the green dot stimulus, and (b) a higher-order thought about that first-order state. One would then expect to find, very shortly after the presentation of the red dot: (c) a perceptual state carrying information about the red dot, and (d) a higher-order thought representing this first-order state. Subsequently, one would expect to find (e) a memory ("higher-order memory"?) to the effect that one had experienced a change from a green to a red dot.

Now, still assuming that we can solve the problems about mental representation and so forth that have been shelved for the moment, this seems clearly to be a distinct scenario from the Stalinesque one, by HOT theory's lights. According to HOT theory, in the Stalinesque scenario one should expect to see: first, (a) some brain activity responsive to the green dot

stimulus (no commitment need be made here about whether this brain activity is robust enough to qualify as an "unconscious perception" of the dot), but *no* simultaneous higher-order representation (about one's seeing a green dot); secondly, (b) some brain activity responsive to the red dot stimulus, but again, no relevant simultaneous higher-order representation; and finally, (c) a higher-order thought or series of higher-order thoughts representing the green dot moving, and then the red dot moving.

Some potential fuzziness is created here by the way Rosenthal states his requirement on appropriate HOTs: among other things (these other things being irrelevant for present purposes), a HOT must be *roughly simultaneous* with the state it is about (so that if one saw a dog last week and *now* thinks "I am seeing a dog", then whatever happens, one's seeing of the dog last week doesn't thereby become conscious.) Given "roughly", there may be a question about whether, in the Stalinesque scenario, the HOT representing oneself as experiencing a dot changing from green to red is best construed as a misrepresentation of the recently lapsed brain states bearing information about the dots, or as an accurate representation of a distinct perceptual state (perhaps a non-existent one). There may be no obviously best way of answering this question, but whichever way it is answered, the overall description of the scenario differs from that of the Orwellian one in two ways: (1) in this second scenario, no higher-order representations about one's seeing of the dots are to be expected prior to the one representing oneself as experiencing motion and color change, and (2) the state occurring at the end of the series involving the experienced motion and color change is in this scenario a thought rather than a memory.

I doubt that we can get significant theoretical mileage out of this second distinction. Once differences in contents are accounted for (i.e. "am seeing" vs "just saw"), and the vagueness implicit in the rough-simultaneity requirement for HOTs is acknowledged, it's not clear that either HOT theory or folk psychology generally provide resources for distinguishing higher-order thoughts from higher-order memories in such a case. But it's clear enough that, on my understanding at least, HOT theory predicts the presence of HOTs with specific contents prior to the existence of the representation of experienced motion and color change if the scenario is Orwellian, and a lack of those very HOTs if the scenario is Stalinesque.

I resist Pete's alternative translation of the Orwellian scenario into HOT-theoretic terms for several reasons. The first is just that if it's translated instead in my way, then HOT theory has the resources to distinguish the scenarios even on the non-relational reading, so unless we have good reason not to pursue a translation along my lines, we should do so in the interests of charity at least. But, further, I believe that the reconstruction I've proposed is a slightly more natural way of construing Dennett's Orwellian scenario in terms of HOT theory. The Orwellian scenario really involves three distinct "moments" of consciousness—first, an accurate experience of a green dot, then an accurate experience of a red dot, then a false memory of motion and color change. The most straightforward way to understand this in HOT-theoretic terms seems to me to be to posit a rapid succession of three HOTs, the first two of which accurately represent the subject's states and thus "render them conscious", and the third of which misrepresents what has just occurred in consciousness and thus renders those original conscious experiences inaccessible to the subject. On Pete's reconstruction, what we'd have instead is a false memory of motion and color change that doubles as a HOT conferring

consciousness on the accurate experiences of stationary green and red dots. But given that "what it's like" for the subject is determined by the contents of his or her HOTs, the subject would not by hypothesis have any conscious experiences as of stationary dots on this reconstruction (not even fleeting ones immediately obliterated by a false memory). In short, the HOT that makes one conscious of oneself as perceiving motion and color change in the Orwellian scenario can't also serve to make one conscious of oneself as perceiving stationary red and green dots, by HOT theory's lights.

Less needs to be said about the Stalinesque scenario, but I would like to point out that on my view, and certainly once we abandon a relational reading of HOT theory, the Stalinesque scenario doesn't require both a false experience of motion and color change, distinct from whatever information the brain carries about the stationary dots, and a HOT representing this false experience. It seems sufficient to qualify as Stalinesque that we have some low-level processing of the stationary dots, and a subsequent HOT that represents oneself as experiencing motion and color change. Even if we take this HOT to represent accurate (unconscious) perceptual states, we have an inaccurate conscious experience as of motion and color change where none really occurred, which is all that the Stalinesque scenario requires. Pete suggests that if the accurate representations are the conscious ones, the scenario won't be Stalinesque. But the only representing that matters once one abandons a relational understanding of HOT theory is the way the conscious person represents his/her mental life to him/herself.

Let me conclude this section by responding to a potential objection, and addressing some of the complications I put aside earlier, concerning the assumptions we've made about

what sorts of information about brain states will be available to future neuroscientists. It might be thought that, at the time-scales involved in the color phi phenomenon, it's simply not possible for all the HOTs posited by my version of the Orwellian explanation to occur in succession, so that my reconstruction of that explanation is flawed. I confess that such a swift succession of HOTs with subtly varying contents seems to me to be an ambitious posit, but nothing at issue turns on the point. Whether such HOTs occur is of course an empirical question, and if we can observe or deduce from what we know about the brain that the first two HOTs do not or could not occur, this would provide grounds for accepting the Stalinesque interpretation as the *correct* account of the color phi phenomenon and rejecting the Orwellian one.

A more powerful version of the objection goes as follows: I'm putting too much faith in the idea that we will eventually be able to distinguish, on the basis of observations of the brain, between a quick succession of three HOTs with closely related contents (Orwellian) and a single HOT (Stalinesque). After all, "brain states" are abstractions from a continuous process of evolving neural activity, and the distinction between the Orwellian and the Stalinesque that I am proposing depends on our being able to carve these processes up very finely into a temporal succession of states with distinct representational contents, perhaps implausibly finely. I grant that this is a concern. For the two scenarios as I've described them to be actually distinguishable by HOT theory, we must eventually have the means to tell (a) when a HOT is present, and (b) at least roughly what its content is, *by means other than correlating brain*

states with verbal reports.[1] This is because the first two HOTs posited in my Orwellian scenario are too fleeting to be reported on, and are immediately "wiped out" by the false memory.

I admit that the prospects for such fine-grained identification of HOTs without knowledge of their ties to actual reports are uncertain, and if the apparent possibility of such identification turns out not to be genuine, I will be the first to concede to Pete that HOT theory cannot distinguish between Orwell and Stalin. For now, it's enough to say in response that whether we will someday have the theoretical and technical resources to do what is required is a thoroughly open empirical question. If there are arguments from what we now know about the brain to the effect that this is impossible, I'd love to hear them. In any case, it seems to me that the prospects for a HOT theory importantly distinct from FPO turn on such questions.

First-Person Operationalism

Finally, I return to the question of how I interpret FPO's defining thesis, and how FPO differs from HOT theory. If FPO is committted to the indistinguishability of the Orwellian and the Stalinesque, then it must involve something more than conscious-state anti-realism, since we've seen that non-relational HOT theory, an anti-realist view in the sense in question, is able to distinguish the scenarios. I think the issue is this: rather than being committed only to the

[1] Dennett discusses one such possibility: Subjects could be prompted to respond with a button-press whenever they experience a red spot, etc. He argues that this wouldn't settle whether the Orwellian or Stalinesque explanation is correct, because the subject could respond with the button-press to an unconscious red dot stimulus. While this is true, it does seem to me that in the absence of evidence to the contrary, or some reason to believe the subject is not being cooperative, we are entitled to take a button-press initiated as a response to that prompt as indicating consciousness. How the subject is taking this instruction might be tested by pairing it with a condition in which the subject is asked to press the button "whenever you become aware of a red dot to any degree, whether fully consciously or not".

claim that there are no conscious states independently of our judgments/thoughts about them, FPO is committed to the claim that there are no conscious states independently of our *reports* about them. To put it another way: FPO is committed not only to the view that conscious states depend on our thoughts or judgments about them, but to a narrow view of what counts in principle as evidence for those thoughts or judgments. On Dennett's view, the *actual heterophenomenological record* obtained from the subject is authoritative about that subject's consciousness; on Rosenthal's view, such a record might in principle leave out a conscious state that the subject didn't report but that was nonetheless "reportable" in the sense that a HOT with a certain content was fleetingly present.

There is thus a sense in which FPO is more anti-realist about conscious states than is (non-relational) HOT theory. FPO takes facts about consciousness to be nothing more than facts about the narratives we tell ourselves about our "phenomenological gardens", which can be read off at a glance from the heterophenomenological record. HOT theory, on the other hand, takes facts about consciousness to be facts about the occurrence of HOTs--theoretical posits with a non-constitutive link to verbal reports. Given our current state of knowledge/ignorance about the mind, I think it counts in favor of HOT theory that it can in principle make the Orwellian/Stalinesque distinction, one which people find intuitively plausible. But it is too early to tell whether this apparent boon to the theory will turn out to be a liability, by committing it to a distinction where there is none to be found.

References

Block, Ned (2011). "The higher order approach to consciousness is defunct", *Analysis* vol.71 issue 3, July 2011.

Mandik, Pete (2009). "Beware of the Unicorn", *Journal of Consciousness Studies*, 16, No. 1.