Commentary on Pete Mandik's Conscious-state anti-realism

My name is Daniel Kostic and I'm an associated researcher at The Berlin School of Mind and Brain.

I want to thank Richard Brown for organizing this conference. I think it's an excellent idea and obviously a tremendous success.

I will be commenting Pete Mandik's paper **Conscious-state anti-realism.**

This is a very good paper with a very clear structure, neatly laid out and in general a great pleasure to read. But, I wouldn't be doing a great job as a commentator if I didn't find something to be critical about.

I will first lay out the main points in Mandik's account and then I will raise some questions about it. I must say that I don't have a strong opinion on either of the original accounts in this debate, but I will point out some issues that arise from Mandik's account in this particular paper.

In this paper Mandik sets out to argue that the HOT theory can defend against the first-person operationalism (FPO) only on a relational reading according to which "consciousness consists in a relation between a HOT and an actually-existing mental state" (p. 1). On Mandik's view of a relational reading the HOT theory must involve three items: two relata and a representing relation between them. Presumably, nonrelational reading eliminates the representing relation. He goes on to argue that an ambiguity or a tension between these two readings in recent literature allows him to claim that HOT theory is actually committed to nonrelational reading but in that case it collapses into a version of FPO.

Let's go step by step here.

According to HOT theory a conscious mental state consists of one's having a

higher-order thought about that mental state. On the other hand the FPO is a thesis according to which we have to deny “the possibility in principle of consciousness of a stimulus in the absence of the subject’s belief in that consciousness” (Dennett, 1991)(p. 132). Given the general understating of the notions realism and anti-realism in the literature, the FPO seems like a clear-cut case of anti-realism.

In section 2 Mandik clarifies his key terminology and key theses regarding consciousness, (FPO, HOT, etc).

He begins with the general distinction between realism and anti-realism in philosophy. He claims that realism about dogs, for example, is a conjunction of an existence claim and independence claim, where the latter is cast in terms of "mind independence". The existence claim requires a very little precification for his purposes but the independence claim needs to be carefully formulated.

The simplest way to specify an independence claim would be to state it in terms of minds: "X exists independently of any mind existing." (Mandik, p. 2) But swapping a "mind" for "X" would render this formulation as incoherent.

A better way to specify this claim would be to cast it in terms of specific kinds of mental states, e.g.: "Minds exist independently of anyone thinking, believing, or judging that minds exist." (Ibid. pp. 2-3)

This is a kind of precification that he will be relying upon in the rest of his paper. The key question then will be whether one’s conscious experience exists independently of one’s thinking, believing, or judging it to exist". (ibid p.3)

He then goes on to give some historical examples of anti-realist positions based on the distinction between existence and independence claims. There are two varieties of anti-realists, nihilists who deny the existence claim and idealists who deny the independence claim. For example a Berkelyan idealist about dogs (funny pun indeed) doesn't claim that dogs don't exist, but that they don't exist independently of being perceived.

Mandik claims that Dennett's formulation of FPO admits only an idealist reading, because it only denies the independence claim.

A further notion that he wants to clarify in this section is "consciousness".

According to Rosenthal, the main proponent of the HOT theory, in thinking about consciousness we can distinguish two kinds of consciousness: "transitive consciousness" (being conscious of something) and "state consciousness" (a mental state's being conscious).

In this sense then one might say that Dennett's claim that FPO is anti-realist about "consciousness" actually means being anti-realist about "consciousness of a stimulus".

The anti-realist thesis about consciousness can now be formulated in the following way: "it denies the possibility in principle of a conscious experience of a stimulus in the absence of the subject's belief in that consciousness" (Ibid pp. 3-4). This formulation of the FPO is assumed in the remainder of his paper.

In the subsection 2.2 Mandik rehearses the Dennett's Orwellian/Stalinesque argument for the FPO.

He describes the phi phenomenon in which subjects exposed to stationary stimuli in a form of fast flashing circles that are different in color, report seeing a moving object that changes its color in the mid of trajectory. The appearance and disappearance of the circles on a screen are so fast so they appear to a subject as a single circle that moves across a trajectory and then suddenly disappears. In these experiments it appears that a single moving circle starts its trajectory in one color and then at the certain point on the trajectory it changes color.

Here Mandik invokes Dennett's distinction between two competing explanation of this phenomenon, i.e. Stalinesque and Orwellian. Orwellian explanation appeals to false memory whereas Stalinesque explanation claims that it's not a false memory of experience that creates this illusion but the experience itself which is false. In Dennett's words, Orwellian explanation revises history (as in famous 1984) whereas Stalinesque stages show trials.

On Stalinesque account there is a temporal order of mental events that is responsible for creating the phi illusion: "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." (Ibid p 5)

The main difference between these two explanations is that: "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)."

So the question arises, how do we know which one of these two explanations is the right one? On Dennett's view no amount of empirical data would help us to decide it.

Now, Mandik explains what role the HOT theory plays in this case and claims that Rosenthal introduces versions of both Orwellian and Stalinesque scenarios that can be distinguished by the HOT theory. Mandik rightly notes that the fact that there are some Orwellian and Stalinesque scenarios that can be distinguished by HOT does not suffice to refute the FPO. A more compelling job at refuting the FPO the HOT theory can do by distinguishing some Orwellian and Stalinesque scenarios that are initially indistinguishable.

And this is the main point of this section: Mandik wants to show that HOT theory can distinguish between Orwellian and Stalinesque scenarios only on a relational reading.

According to HOT theory a state is conscious if there is a HOT that is about that state. On a relational reading there are three items, a HOT and a state it is about, and a relation between them which is "a representing relation". On this reading state is conscious when for example a visual experience is represented by a HOT, and a state is not conscious if there is no such representing relation accompanying it. This is a bit odd formulation and shall be discussed in a moment.

Now let's get back to the phi phenomenon having this relational reading in mind.

In order for a phi phenomenon to appear as moving circle that changes its color one has to have a HOT whose content is visual experience of a moving, color changing circle.

Mandik wants to argue that there is no such representing relation and furthermore if the HOT theory is true no such relation figures in it. I happen to

agree with this, but for different reasons, which I will explain in more detail later on. If there is no such relation then the HOT can't distinguish between the HOT Orwellian and HOT Stalinesque explanations of the phi phenomenon. In any case this means that the content of one's consciousness is just a content of a HOT, not a combination of a HOT and a representing relation.

In the fourth section Mandik introduces his unicorn argument which is a response to the inconsistent triad of intentionality.

The triad of representationalism goes 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. (p. 11)

Obviously each of the propositions taken individually is plausible, but one cannot hold them to be jointly true.

The unicorn argument denies the first proposition while keeping the other two. In other words denying the first premise amounts to saying that there is no representing relation. As Mandik puts it: "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" (Ibid p. 11). I think this is a controversial claim.

Mandik wants to argue that there is no such representational relation and furthermore if the HOT theory is true no such relation figures in it. If there is no such relation then the HOT theory can't distinguish between the HOT Orwellian and HOT Stalinesque explanations of the phi phenomenon. And this further means that the content of one's consciousness is just a content of a HOT. But I see no problem with that, this is what Rosenthal (2011) seems to claim in his reply to Block.

So if I understand Mandik correctly, on nonrelational reading the representing is not an epistemically and metaphysically independent relation between what is represented and a representation. In other words, on a nonrelational reading the representing relation is not something over and above the representation, it is

actually the HOT itself. Somehow it seems that without a representing relation as epistemically independent from the HOT we end up with what I am inclined to call "representational danglers", i.e. the HOT can't accurately pick out an actual first order state that is causally responsible for its formation and in that way collapses into the FPO. That is to say, without a representing relation the HOT can't help us distinguish which first order state is responsible for its forming in both Orwellian and Stalinesque scenarios.

I hope this is a correct understanding of Mandik's distinction.

It wasn't quite clear to me what motivates Mandik's relational reading of the HOT theory in the first place. Perhaps he would want to expand on that. For example, it's not at all obvious to me how a quote from Rosenthal on page 7 invites the relational reading.

It's not clear from Mandik's quote exegesis that the literature can support a tension between these readings. The other issue is whether nonrelational reading actually makes the HOT theory collapse into an FPO. I'll explain that in a moment.

Let's go step by step here as well.

It's not quite clear to me what Mandik means by representing relation. I agree with him when he says that "Representation may involve relations, but it is not constituted by a relation to that which is represented" (p. 11). And he continues by claiming that: "It follows from there being no representation relation that there is no such relational property as the property of being represented" (Ibid.). This seems very controversial. If correct this would allow for representational danglers, or HOTs that have no content, but if they have no content they can't be HOTs in the first place, i.e. if nothing is represented then we have no representation.

I am not sure that HOT theory is ever committed nor that it has to be committed to a relational reading because there seem to be some relations between a first order state and a HOT, e.g. a bottom-up causal relation from a first order state to a HOT, or a top-down intentional relation from a HOT to a

first order state, but that doesn't mean that there is a representing relation over and above these categories. It would sound quite odd to say that there is a special, additional causal relation between a cause and an effect. The same applies to this case.

On the other hand I think there are some resources in the latest accounts of the HOT theory, such as a combination of the quality space theory and Higher Order Awareness, that can help us deal with cases of unconscious perception. I don't see that the issue of whether it collapses into the FPO strongly depends on the distinction between relational and nonrelational reading.

Furthermore the quote from Rosenthal (1995) only mentions the content of representations, although he mentions the words "relational" and "non-intrinsic", "external to a state", I can't see in this quote that he actually claims that a representing relation is something over and above the representation or its content. So I don't really see how this invites a relational reading. Perhaps Mandik would want to elaborate on this as well.

I agree with Mandik in a certain sense, inasmuch as I think that there is no reason to think that HOT theory in any case requires a relational reading.

Furthermore, the fact that there is a bottom-up causal relation (the first order state causes a HOT) and there is an "intentional relation" which branches top-down either to the actual first order state or to an intentional content of a HOT, doesn't mean that there is a special independent representing relation. But if Mandik is right about this then what we are left with on a nonrelational reading? I think it leaves us with representational danglers, which is something that wouldn't make much sense, for it would allow for HOTs that have no content and if they have no content they are not HOTs. It would be wrong to conclude from this that HOTs without content undermine the HOT theory. It rather means that HOTs without content are not actually HOTs. So the very idea of representational danglers doesn't seem to make much sense. In this case he would have to argue that either there is no bottom up causal relation between a first order state and a HOT or that there is no top-down intentional relation between a hot and a first

state. However, the case of empty thoughts requires that we have an intentional relation without a bottom-up causal relation. But intentional relation can be between a HOT and its intentional content regardless of the first order state. The only difference is that in the normal case without empty thoughts the intentional content matches the first order state. So on both relational and nonrelational readings the HOT is representing something, it's not strictly speaking false. The only difference is that in the case of unconscious perception such as masked priming, blindsight or the phi phenomenon the first order state does not correspond to whatever is causally responsible for it, but it only seems to correspond to whatever mechanism is responsible for forming the false perception, like in the phi phenomenon where we perceive a moving circle instead of many stationary circles. Or like in the case of empty thoughts where the HOT is representing only its intentional content.

I'm not quite convinced from the quotes that there is a tension between relational and nonrelational reading. Maybe there is, but from the Mandik's account it's not quite clear. Perhaps he would like to address this issue in his response.

Rosenthal argues in "*multiple drafts*" (1995) that being conscious is not an intrinsic property of mental states, because the same mental state can become unconscious from conscious and vice versa. He argues that consciousness is rather a property of being represented by a HOT. It's not quite clear to me then what Mandik takes it to be the difference between a relational and nonrelational reading of HOT theory in this context. This formulation doesn't seem to invite relational reading.

A state is conscious just in case we're able to report it non-inferentially. So for Rosnethal there is a fact of the matter to tell whether a state is conscious or not.

In fact I happen to agree with this statement: "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. " (p. 10)

Another issue to distinguish here is about misrepresentations and erroneous introspection. But that's topic for another discussion.

Furthermore, it's not quite clear how a nonrelational reading of HOT theory commits it to a version of FPO from the paper. That part of the paper ends a bit abruptly and perhaps more needs to be said about this crucial step in his account. Perhaps Mandik would want to elaborate on this as well.

I think there are some resources in the body of Rosenthal's work that can perhaps help him withstand this challenge. Rosenthal claims in *exaggerated reports* (2011) and in *how to think about mental qualities* (2011) that actually there are cases in which a HOT can represent states that occur unconsciously, for example in subliminal perception or masked priming. In these cases the HOT is about whatever we believe it's about, i.e. it's about whatever it seems to us is a qualitative state regardless of the actual unconscious states that is in fact causing the HOT. In this case Rosenthal proposes using the quality space theory to capture the actual first order state (which is unconscious).

Rosenthal claims in Multiple drafts and the facts of the matter (1995), and it's implicitly or explicitly stated in many of his writing on HOT theories, that a first order state doesn't have to be conscious for a HOT to represent it. He seems to be claiming that the representing relation is always between the HOT and its intentional content, that's what makes a state conscious. In some cases the intentional content matches the first order state, but in some other cases it doesn't. The other issue is how to determine which actual state is causally responsible for that HOT or a first order state. But in any event it seems to me that the quality space theory is designed to answer that worry. It should be noted that this case is a bit different than the empty thought case, inasmuch as we can use a QS to map unconscious perception.

So the question whether HOT that represents merely its intentional content is a representational dangler remains open in the case of empty thoughts, but not

in the case of the color phi. But if it is a representational dangler then it can't be a proper HOT. So both objections don't seem to do much harm to the HOT theory.

Mandik seems to claim that the color phi phenomenon is an example of an inexistent first order state. He claims that we can't accurately determine which first order state the HOT is about and since it actually doesn't exist we end up having a false HOT. I don't think that the color phi is a good example of an inexistent first order state. It's only an example of an unconscious perception that is causally responsible for the HOT. Just because there is no actual motion to be experienced, there is still a causal relation between whatever unconscious perception or mechanism is responsible for the first order state and consequently for the HOT. On the other hand there is certainly an intentional top-down relation from a HOT to whatever intentional content of that HOT is. What creates a confusion here is that these two relations in the case of color phi don't match. But Rosenthal has already proposed a solution for unconscious perceptions such as masked priming or blindsight (and I can't see why color phi can't fall in this category as well), namely the quality space theory, that can be used to correctly match the intentional and causal relations in color phi.

A more interesting question regarding empty HOTs is not whether they can represent non-existent objects but rather whether they can represent an unconscious mental state. Recently Rosenthal has proposed to use a quality space theory in combination with Higher order awareness (HOA) to explain and integrate unconscious perception, that would normally be considered an empty HOT (if we are not aware of it can't have a what-it-is like quality).

Let me summarize the key points of my criticism.

1. I am not quite sure that the HOT theory actually allows a relational reading. It seems that this reading would imply an additional representing relation that is over and above the causal, or intentional relations, which if denied would in effect make a HOT a representational dangler. But that wouldn't make much sense because in that case we wouldn't have a proper HOT but something else.

2. I don't think that color phi is a good example of inexistent first order state.

3. I don't think that even on a nonrelational reading the HOT theory collapses into the FPO. There are moves on Rosenthal's part that don't seem to allow this, for example the quality space theory.

Let me say again that this is a very good paper, with plenty of great ideas, it provides a resource for much better understanding of the HOT theories and it provokes a further discussion on this topic.

Daniel Kostic