This is an expanded version of the paper I presented at the APA Pacific Division meeting in San Francisco, CA (2005)

From HOTs to Self-Representing States

Paul Raymont (paulraymont@trentu.ca, paulraymont@hotmail.com)

Trent University, Department of Philosophy

Jan.30, 2005

According to David Rosenthal, a mental state is conscious just in case its subject suitably represents herself as being in that state, where this entails that the mental state "is accompanied by a noninferential, nondispositional, assertoric thought to the effect that one is in that very state" (2002a, p. 410; see also Rosenthal, 1997, p. 742). This assertoric thought, since it is about another mental state, is a higher-order thought (HOT). Let us use 'HOT*m*' for a HOT that is about mental state *m* (where *m* is, say, visually representing something as being blue). The claim is that *m*'s being conscious consists in its being suitably represented by HOT*m*.

Rosenthal requires that HOT*m* be a different state from *m*, so that *m* does not represent itself. This is because, he says, "Absent some reason to the contrary, distinct mental functions call for distinct states" (2004, p. 30; see also Rosenthal, 1986, 344-5; 1997, p. 738). So the represented mental state, *m*, the one *of which* the subject is conscious, is not the representing state by means of which she is thus conscious. When a state is conscious, then, there are really two states: the consciousness-conferring state, HOT*m*, and the state on which consciousness is conferred, *m*.

I will argue that there is a good "reason to the contrary" for saying that both the first-order and the higher-order contents are carried by the same state, a self-representing state.[1] This reason arises from the possibility that HOT*m* misrepresents *m* (a source of

recent objections to HOT-theory).[2] For example, suppose that *m*, seeming to see something blue, is misrepresented by HOT*m* as seeming to see a red thing. Will one then be in a conscious state of seeming to see red or seeming to see blue?

As Rosenthal interprets the objection, the worry is that "there is no good answer to how things would be phenomenologically in these kinds of case" (2004, p. 32). But that is not the best way to think of the problem. Indeed, it is all too clear "how things would be phenomenologically" according to Rosenthal's model – they would be the way the HOT represents them to be. For, on his view, how things seem phenomenally to the subject, 'what it is like' to be thus conscious, is determined by the HOT. As Rosenthal says, "Since HOTs make one conscious of oneself as being in a particular state, what it's like for one to be in a state is a function of how one's HOT represents that state" (2002c, p. 658).[3] But then it looks like HOT*m* alone determines what things are like for the subject, and *m* turns out to be an idle wheel.

It is hard to see how the HOT-theorist could avoid this implication, for it derives from a part of HOT-theory that cannot be excised without sacrificing the theory. HOT-theory begins with the idea that a conscious state is one *of which* I am conscious. Since I am conscious of it I must have a representation of it, and this representation, since its object is itself a mental state, is a higher-order representation. Moreover, it is generally the case that how a represented item seems to me will be determined by how I represent it. So, for example, how a first-order mental state seems to me, or what it is like for me, will depend on how I represent it, and thus is determined by my higher-order representation. That is why Rosenthal concludes that it is HOT*m* that determines what things are like for the subject of *m*.

The problem is that on Rosenthal's theory, state consciousness is supposed to be constituted in a *relational* complex in which the state that is conscious (*m*) is a state that *is represented*. But cases of misrepresentation make it clear that it is really the *representing* state (HOT*m*), all by itself, that determines the nature of one's consciousness; and the represented state, which is supposed to be *the conscious state*, makes no contribution in determining what things are like for the subject.[4] Its being thus idle undercuts its claim to being the state that is conscious.

HOT*m* can misrepresent either because *m* lacks the qualities it is represented as having, or because *m* does not even exist. Let us develop the objection by focusing on the latter kind of misrepresentation. Call the HOT, the object of which does not exist, an 'empty HOT'. Regarding such HOTs, Rosenthal says, "A HOT's accompanying its target is subjectively indistinguishable from a HOT's occurring in the absence of that target" (2004, pp. 40-41).[5]

This remark conflicts with a key tenet of his HOT-theory, specifically, with the idea that a conscious state is a represented state. If having an empty HOT of being in *m* (a state of visually representing something as blue) really is *subjectively indistinguishable* from having a veridical HOT of being in that state, then it must seem phenomenally to the subject of the empty HOT as if she sees blue. What it is like for her at the time is just as it would be if she really were representing blue – otherwise the empty HOT would be subjectively *distinguishable* from the veridical one. Thus, *she really is in a conscious state* (and does not merely seem to be in one) – assuming that any state, the having of which is subjectively indistinguishable from having a conscious state, is itself a conscious state.[6] What, then, is the conscious state that she is in? Not *m*, for it does not exist (since

by hypothesis HOT*m* is empty), and the conscious state that she is in must at least be a real state. However, assuming the tenets of Rosenthal's account, *m*, the state that is represented, is the only candidate for being the state that is conscious.

In short, on Rosenthal's model one's conscious state, what it's like for one, remains the same – the empty and veridical alternatives being "subjectively indistinguishable" – regardless of whether the mental state that *is represented* exists, and therefore, on the terms of Rosenthal's model, regardless of whether the state that *is conscious* exists. That is a reductio of HOT-theory.

Rosenthal does not appreciate the force of this form of the objection from the possibility of misrepresentation by a HOT. This is evident from his remarks about the two basic ways in which a HOT can misrepresent its target (viz., by representing a target that does not exist or by representing a real target as having features it lacks). He says, "The distinction between an absent target and a misrepresented target is in an important way arbitrary" (2004, p. 32). In fact, though, it is only the former kind of misrepresentation, where the HOT is empty, that is compatible with there being at the time just one mental state on hand, namely, the misrepresenting HOT. We need acknowledge only this mental state in order to account for the misrepresentation, since any represented mental state is by hypothesis a mere fiction. And since this misrepresenting HOT is not itself conscious (given that it is not itself represented by any further HOT), the kind of misrepresentation that occurs when a HOT is empty is compatible with the subject's having *no* conscious state at the time. So Rosenthal is wrong when he says that the two kinds of misrepresentation "occasion the same kinds of phenomenological perplexities, if any" (2004, p. 32). For only the kind of

misrepresentation that occurs in the case of an empty HOT is compatible with one's having no conscious states.

For example, if HOT*m* is empty, then when one seems to see a blue ball, or is in a state that is 'subjectively indistinguishable' from seeing it, then not only might there in fact be no such ball, and not only might one's first-order mental state in fact be something other than a visual representation of a blue ball, but one might in fact have no conscious state at all at the time. Rosenthal's account allows this as a genuine possibility.

To take another illustration, consider the starting point of the *cogito*. I think? Maybe not; maybe I have confabulated that thinking. It is, after all, my current *conscious* thinking of whose existence I am supposed to be assured.[7] But if it is conscious thinking, then it is thinking *of which* I am conscious. In Rosenthal's model, then, it is *represented* thinking, thinking that I represent myself as carrying out. That is, the thinking in question is a mental state that I attribute to myself by means of a HOT that represents me as thus thinking. But that HOT might be empty, in which case there is no conscious thinking, no conscious state, at all.

What is in jeopardy is my certainty that I really am in a conscious state when I seem to be in one. For this certainty is rooted in the fact that I simply cannot go wrong in judging that I am in a conscious state whenever I seem subjectively to be, that is, whenever my condition is subjectively indistinguishable from some state of consciousness. But, on Rosenthal's account, I *can* go wrong in making this judgment (and do if the relevant HOT is empty). So the basis for my certainty is gone.

A HOT-theorist might reply that I am here at least assured of the reality of the HOT in the light of which I seem thus to be thinking. Granted; however, this does not

give us the result we want. For what I am in fact assured of is that there is at least some *conscious* thinking, and the HOT that sits at the top of the hierarchy, and that is supposedly responsible for the consciousness of my conscious state, is not itself, according to Rosenthal's theory, a conscious thought.[8] So being assured of the HOT's existence is not the same as being assured that I am in a conscious state.

Also, if the conscious state does not even exist, why posit any other state (e.g., a HOT) to account for its consciousness? There simply is no state that is conscious, and so no reason to posit a HOT to account for its being conscious. A HOT-theorist might reply that I need to posit the HOT in order to account for my at least *seeming* to be in a conscious state. Still, this is warranted only if the 'seeming to be in a conscious state' is a *conscious* seeming,[9] one *of which* I am aware. But, again, if this mental state (of things so seeming) really is conscious, then in Rosenthal's account it is one that I represent myself, by means of a HOT, as having. And that HOT could be empty, the mental state that it attributes to me being mere confabulation.

The upshot is that if HOT*m* is empty, then all it supplies is transitive consciousness, consciousness of (or as of) *m*, and there exists no state of the subject's that has intransitive state consciousness (that is, no state that is itself conscious). This undercuts a central part of HOT-theory, namely, the view that a special kind of transitive consciousness (namely, being conscious of *m* by means of a HOT) gives rise to intransitive state consciousness (*m*'s being conscious).

Rosenthal suggests that empty HOTs pose no difficulty because we can identify the conscious states in such cases with "merely intentional items" (2003, p. 345, n. 31).[10] He says that "the conscious state a HOT is about may be a merely notional state and may

not actually exist" (2000b, p. 232). For Rosenthal, this is because a conscious state is a state that the HOT makes one conscious of oneself *as* being in, and the state that is alluded to after the 'as' can be a merely intentional item, one that does not exist.[11]

This approach is still vulnerable to the above '*cogito*' worry, for my conscious thinking is a real occurrence. It is a dated, particular thinking, not a mere intentional object, and this is what I know with certainty. Also, if it were a mere intentional object, it would support no belief that I myself am anything more than an intentional object. The *cogito* establishes for me that I am more than a mere notional item because one of my states, a mental state *of which* I am conscious, is. What the *cogito* exploits is that my current conscious state is a real state *whereby things seem* thus-and-so, and not just a state that I *seem* (via an unconscious HOT) to be in.

A second difficulty with the idea that the represented mental state is a mere intentional object has to do with sensory states. When discussing how HOTs determine what it's like to be in a sensory state, Rosenthal emphasizes the incapacity of HOTs to generate new sensory qualities (2004, p. 40; 2002a, p. 413). He says that new concepts, and the HOTs that embed them, enrich sensory consciousness not by generating the sensory qualities of which one is newly conscious, but instead by enabling one to discriminate more finely among sensory qualities that were already there, so that one can be conscious of more finely individuated sensory states.[12] However, if a sensory state is unreal, a mere intentional object (as real as a unicorn), then my thought about that state does not *reveal* its nature – there is no real state already there and waiting to be revealed. So if the represented sensory state is a mere intentional object, then the HOT must be

credited with more creative power with respect to sensory qualities than it plausibly enjoys.

A third difficulty has to do with the causal relations of conscious states. For Rosenthal, if I am conscious of a headache, I will as a result do some things distinctive of having a headache. I will take aspirin, avoid loud music, and so on. Here, the cause is the headache, a real mental state that produces effects even if I am not aware of it. Surprisingly, Rosenthal denigrates the causal contribution of the HOT in virtue of which the headache is conscious. He says, "A state's being conscious *may to some extent* matter to its causal role" (2002a, p. 416; emphasis added; see also Rosenthal, 2004, p. 35). Still, most of the effects are, he thinks, accounted for by the headache, not the awareness of it in the HOT. By contrast, when the HOT is empty and the represented headache is a mere intentional object, the headache cannot produce any effects, for mere intentional objects have no causal clout. Nevertheless, if in such a case my mental condition is *subjectively indistinguishable* from having a real headache, this surely *will* produce some results that are distinctive of headaches (e.g., taking aspirin). What accounts for these effects if there is no real headache on hand to explain them? Rosenthal deals with such cases by having the HOT pick up the slack. He says, "If a conscious state is wholly confabulatory, any causal impact will be due to the HOT that represents us as being in the confabulatory state" (2000a, p. 212).[13] This seems *ad hoc*. It is odd that some at least of the causal relations typical of headaches are had now by the sensory states that are represented by HOTs, but at other times by the HOTs themselves. Moreover, to the extent that the HOT not only makes things seem subjectively indistinguishable from having a real headache but also now takes on some of the distinctive causal relations of real headaches, to that

extent it seems to become *a real headache*. What reason could there be for denying that it is? But if it is a headache, then the headache was not confabulated after all; it is a real, efficacious mental state, and not a mere intentional object.

Equally strong objections confront an alternative response to the problem of empty HOTs that Rosenthal considers, according to which we are to "construe the state that's conscious in these cases as being some relevant occurrent state that we're conscious of, but in an inaccurate way" (Rosenthal, 2003, p. 345, n. 31). In fact, though, Rosenthal's theory is incompatible with this alternative.[14] This is because it sits ill with other aspects of his account for Rosenthal to require that there be a real target state on hand whenever a HOT misrepresents. Recall that in his model, *how things seem* to the subject, 'what it is like' to be thus conscious, is wholly determined by the HOT.[15] Why, then, should it matter whether there is a real, but drastically mischaracterized, state on hand as opposed to there being no such state?

Also, it is unclear whether this alternative response would preserve the central insight of HOT theory, namely, the claim that a mental state is not conscious unless I am *conscious of it*. As Karen Neander has pointed out (1998, p. 423), the point of this claim is lost if it is taken to require only that I be conscious of the mental state in any old way, even if the nature of the represented mental state diverges utterly from the features that I represent it as having. On a more plausible reading, the HOT-theorist's claim should instead be that my mental state is not conscious unless I am conscious *of its nature* (e.g., of its sensory qualities). On this approach, a ticklish feeling cannot really be said to be conscious if I drastically misrepresent it as, say, a headache. In such a case, what it's like for me will be just as it would be if I really did have a headache, given that 'what it's

like' is determined by the HOT, by how it represents, or misrepresents, a mental state. But since what it's like for me is indistinguishable subjectively from having a headache, and not from a ticklish feeling, it is false that I am conscious in the appropriate way of the ticklish feeling. I am not conscious of it in the way that I would need to be in order to make it conscious, for I am not conscious of its nature. Therefore, since the ticklish feeling is not itself conscious, we cannot identify it with my conscious state. In short, on the most plausible reading of HOT-theory's central claim, we should not adopt this alternative reply, on which we identify the conscious state with "some relevant occurrent state that we're conscious of, but in an inaccurate way" (Rosenthal, 2003, p. 345, n. 31).

Even if a HOT-theorist could overcome these obstacles and pursue the above alternative line of response to the problem of empty HOTs, there remains an important difficulty. When a HOT misrepresents (say) a ticklish feeling as a headache, some of the typical effects of real headaches will again be present, and these are hardly attributable to the misrepresented ticklish feeling. So the HOT must again pick up the slack by taking on some of the distinctive causal relations of real headaches, in stark contrast to the causal relations of the represented state. As Rosenthal says, "If a HOT represents an actually occurring state as having different mental properties from those it actually has, the HOT and target state will correspondingly diverge in causal influence" (2000a, p. 212). But, again, if the HOT has the characteristic headache effects, then it is difficult to see why the HOT is not now a real headache; in which case, this is not a case of misrepresentation after all, and the HOT itself is a conscious state (a headache), contrary to the requirements of Rosenthal's model.

In conclusion, the problem of empty HOTs shows that when I am (transitively) conscious of one of my (intransitively) conscious states, both the first-order and the higher-order contents are carried by the same state, a self-representing state. As Victor Caston notes (2002, p. 781), such a state is not only what is represented but also what does the representing, and its having this latter function does require that it at least exist. That is, in self-representing states we have an exception to the rule that representing states can be 'empty' or exist without the items that they represent; for, since they are one and the same state, the existence of the represented mental state is required by the existence of the state that represents it. The upshot is that in order to avert the difficulties that beset Rosenthal's HOT-theory, we should endorse self-representing states.

NOTES

[1] For an excellent recent treatment of self-representing states, see Kriegel (2003).

[2] Difficulties connected with inaccurate HOTs have been noted by Byrne (1997), Neander (1998), Seager (1999), Levine (2001), and Balog (2000).

[3] Elsewhere, he says, "If I am conscious of myself [via a HOT] as being in a P state, it's phenomenologically as though I'm in such a state whether or not I am" (Rosenthal, 2004, p. 35). See also Rosenthal, 2004, p. 41.

[4] As Joseph Levine puts it, "It looks as if the first-order state plays no genuine role in determining the qualitative character of experience" (2001, p. 108). This worry is borne out by Rosenthal's remark that, "A higher-order awareness of a P state without any P state would be subjectively the same whether or not a Q state [one's actual first-order state] occurs. *The first-order state can contribute nothing to phenomenology* apart from the way we're conscious of it [via the HOT]" (Rosenthal, 2004, p. 32; emphasis added).

[5] Elsewhere, Rosenthal says, "A case in which one has a HOT along with the mental state it is about might well be subjectively indistinguishable from a case in which the HOT occurs but not the mental state" (Rosenthal, 1997, p. 744). He says also that the "nonveridical appearance of pain is indistinguishable, subjectively, from the real thing" (Rosenthal, 2002a, p. 415). See also Rosenthal, 2000a, p. 213; and his 2000b, p. 237.

[6] If that is not what Rosenthal means by 'conscious state', then he is stipulating a new use of the phrase, and we can set aside his theory as not being an account of what many others mean by it.

[7] I will take as my focus only this aspect of the *cogito*, that I cannot go wrong in believing that I am in a conscious state, and not the (equally certain) claim that I exist.

[8] For Rosenthal, a HOT can itself be conscious if it is the object of an even higher-order HOT. This is not the standard case, but happens, according to Rosenthal, in introspection. The problem of empty HOTs is recapitulated in such cases, for the represented states may yet be unreal, mere confabulations, and the highest-order HOT that sits at the top of the hierarchy represents but is not itself represented, and is therefore not conscious (on Rosenthal's theory).

[9] 'Conscious' in the relevant sense, namely, intransitive state consciousness.

[10] Rosenthal there attributes the objection from empty HOTs to Elizabeth Vlahos.

[11] As Rosenthal puts it, "Being in a conscious state is not being in that state and being conscious of being in it, but simply being conscious of oneself as being in the state" (2004, p. 41). Note that if we substitute an ellipsis for the negative claim in this sentence and isolate its positive component, we get, "Being in a conscious state is . . . *simply* being conscious of oneself as being in the state" (Ibid., emphasis added). This equates being in

a conscious state with suitably representing (or being conscious of) oneself as being in it; so that if one is conscious of oneself (via a HOT) as being in *m*, then one really is in *m*; it really is one's conscious state. This commits Rosenthal to a doctrine of infallibility about not only the existence but also the *nature* of one's current conscious state, something that he is elsewhere at pains to reject (Rosenthal, 1991, p. 32 n. 17; 2002a, p. 415).

[12] "How," he asks, "could merely having new concepts give rise to our sensory states' having new properties?" (Rosenthal, 2002a, p. 413) See also Rosenthal, 1997, p. 742; 1993, p. 362; 1991, p. 34.

[13] Elsewhere he says of such cases, "The causal role played by the HOT will matter even more" (2002a, p. 416).

[14] Indeed, when Rosenthal considers this alternative, he refers us to an earlier paper in which he does reject it (2000b, p. 232).

[15] See n. 3, above, for references.

REFERENCES

Byrne, A. (1997): 'Some Like It HOT: Consciousness and Higher-Order Thoughts', *Philosophical Studies* 86, 103-129.

Balog, K. (2000): 'Comments on Rosenthal's "Consciousness, Content, and Metacognitive Judgments"', *Consciousness and Cognition* 9, 215-219.

Caston, V. (2002): 'Aristotle on Consciousness', *Mind* 111, 751-815.

Kriegel, U. (2003): 'Consciousness as Intransitive Self-Consciousness: Two Views and an Argument', *Canadian Journal of Philosophy* 33, 103-132.

Levine, J. (2001): *Purple Haze*, chap. 4, Oxford: Oxford University Press.

Neander, K. (1998): 'The Division of Phenomenal Labor: a Problem for Representational Theories of Consciousness', in *Philosophical Perspectives* 12, 411-434.

Rosenthal, D. (1986): 'Two Concepts of Consciousness', *Philosophical Studies* 49, 329-359.

(1991): 'The Independence of Consciousness and Sensory Quality', *Philosophical Issues* 1, 15-36.

(1993): 'State Consciousness and Transitive Consciousness', *Consciousness and Cognition* 2, 355-363.

(1997): 'A Theory of Consciousness', in Block, Flanagan, and Guzeldere (eds.), *The Nature of Consciousness*, (pp. 729-753), Cambridge, MA: The MIT Press.

(2000a): 'Consciousness, Content, and Metacognitive Judgments', *Consciousness and Cognition* 9, 203-214.

(2000b): 'Metacognition and Higher-Order Thoughts', *Consciousness and Cognition* 9, 231-242.

(2002a): 'Explaining Consciousness', in Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, (pp. 406-421), Oxford: Oxford University Press.

(2002b): 'Consciousness and Higher-order Thought', in Nadel (ed.), *Encyclopedia of Cognitive Science*, (pp. 717-726), London: Macmillan, Nature Publishing Group.

(2002c): 'How Many Kinds of Consciousness?', *Consciousness and Cognition* 11, 653-665.

(2003): 'Unity of Consciousness and the Self', *Proceedings of the Aristotelian Society* 103, 325-352.

(2004): 'Varieties of Higher-Order Theory', in Gennaro (ed.), *Higher-Order Theories of Consciousness*, (pp. 17-44), Amsterdam: John Benjamins Publishing Company.

Seager, W. (1999): *Theories of Consciousness*, chap. 3. London: Routledge.