On Dispositional HOT Theories of Consciousness

Higher Order Thought (HOT) theories of consciousness contend that consciousness can be explicated in terms of a relation between mental states of different 'orders', in which the higher order state – generally a thought – is *about* the lower order mental state. One form of such a theory holds that a conscious mental state is one which causes a thought with the content that the subject is in that mental state. We can outline various forms of HOT theory more precisely if we invent a small piece of notation: where α is a mental state, let T[α] be the thought that one is in the state α (we can naturally extend this notation to yet higher order thoughts in the obvious way, so that, for example, T[T[α]] is the thought that one is having the thought that one is in α). Let's call any thought of the form T[α] a *nested* mental state. The generic form of the HOT Theory can then be given thus:

For any subject, S, and mental state, α, α is a conscious state if and only if

(1) S is in α,

(2) α bears an appropriate relation, R, to T[α].

As we shall see the relation R can be of various sorts. As noted, the most typical is simply that α *cause* T[α] to occur. This entails that for α to be a conscious state of S, S must also have the thought T[α] (where S actually *has* the thought T[α] we call T[α] an *occurrent* thought). Additional niceties can intrude; for example, there are good reasons to demand that the causal relation from α to T[α] should not be mediated by inference. Distinctly articulated versions of such a theory have been developed by David Rosenthal (19??) and David Armstrong (19??). While a great deal could be said about HOT theory in general (for a critical overview see my

1999, chapter 3), I want to focus here on a new version of the HOT theory devised by Peter Carruthers (19??). His account contains some novel features provide an interesting analysis of the problem of qualia and a nice way to deal with certain issues of 'cognitive overload' that arise in other forms of HOT theory. However, I will not expound the virtues of Carruthers's theory, for I believe that its novel features lead, in the end, to insuperable difficulties.

We can approach these difficulties by asking a simple question: can HOT theories require that the higher order thought, T[α], which 'makes' α conscious be itself a conscious mental state? The answer is apparently an obvious 'no' since such a requirement would generate a vicious infinite regress of nested conscious states. Not only is it the case that it is phenomenologically plain that when I am conscious of some mental state, α, I am not also *conscious* of each of an infinite hierarchy of states T[α], T[T[α]], ..., T[...T[α]...], etc. but there must also be neurologically founded limitations on the number and complexity of thoughts that any of us can actually entertain at one time. But on the other hand, HOT theory cannot rule out the possibility of a higher order thought's being conscious since, generally speaking, it is certainly possible to become aware that one is having a lower order thought. The generic theory outlined above, with R as the relation of causation, is designed to meet this difficulty. One *can* become conscious of the higher order thought, T[α], that makes α a conscious state if T[α] should bring about the still higher order thought T[T[α]], but there is no requirement that this thought should occur to one in order for α to be a conscious state. Since it seems evident that we are often conscious without being conscious of being conscious but that sometimes we do enjoy such higher order consciousness, this would appear to be an advantage of the standard form of HOT theory given above.

Most interestingly, Carruthers disagrees. There are – according to Carruthers – two reasons for requiring that the consciousness conferring higher order thought itself be a conscious thought. The first is that – under rather special circumstances – the defining conditions of the HOT theory given above can be fulfilled without α being a conscious state. The example that Carruthers uses is this:

> Suppose that I am disposed to make judgements on utilitarian grounds whenever practical considerations of morality arise. I therefore believe that the morally correct thing to do is whatever will cause the greatest happiness to the greatest number. But I am not aware of having this belief. Indeed, if challenged, I may be inclined to deny it .... Yet in the course of a discussion of the merits and demerits of judging actions in terms of what will cause the greatest happiness to the greatest number, I may find myself speaking of the people who maintain such a view as 'we', and becoming angry when their views are criticised, thus manifesting the higher order belief, that I believe myself to believe utilitarianism to be true. (1996, 173)

The problem here, of course, is that since the higher order belief – which I possess all along – is not itself conscious it does not spur me into a reevaluation of my (lower order) moral beliefs. It does not flush the lower order beliefs into the light of consciousness where they can be properly examined and, in this case, consciously accepted.

This consideration is not entirely decisive. It is not clear to me that such cases could not largely be accounted for in terms of lower order beliefs. One gets angry at criticisms of utilitarianism simply because one favours utilitarianism, not necessarily because one believes that one favours utilitarianism. The use of 'we' might indicate the coming to believe that one favours

utilitarianism rather than a sign that one already believed it. Carruthers notes that in the imagined scenario this defence of utilitarianism 'may strike me with the force of self-discovery' (1996, 173) but it could equally be surprise at suddenly consciously seeing the force of the utilitarian position which goes with the adoption of the higher order belief.

Carruthers's second reason for espousing the view that consciousness conferring higher order thoughts must themselves be conscious thoughts is simply that '...as a matter of fact it does seem to be the case that whenever I have a conscious experience or thought it is always *available* to *conscious* thought' and '... the only cases where higher order thoughts seem to make a difference in behaviour ... are where they are conscious ones' (1996, 173). Carruthers admits that this is indecisive but thinks such phenomenological evidence is sufficient to invoke a methodological principle of 'minimizing accidents'. All things considered it would be preferable to have a theory that did not leave it an accident that effective higher order thoughts are always (or usually) conscious. But this idea seems to ignore the greatest source of potency that higher order states possess according to HOT theory – namely the 'ability' to confer consciousness upon lower order states. These lower order states being conscious results in major behavioural effects and so indirectly the higher order states possess wide ranging causal powers whether or not they are conscious.

Though it is important to see that there are no *compelling* reasons to accept the requirement that the consciousness conferring higher order thoughts be themselves conscious, it is not my main business to complain about the reasons why Carruthers favours his modified HOT theory. In any case, haven't we shown above that any HOT theory that requires these higher order thoughts to be conscious is committed to an impossible infinite proliferation of conscious

nested mental states? In fact, Carruthers's account dodges this bullet, for his version of HOT theory balances the extremely strong requirement that the higher order states be conscious with the novel, and much weaker, condition that lower order states need only be *disposed* to produce conscious higher order thoughts in order to be conscious. This disposition is causally grounded in the cognitive architecture of the subject. Carruthers says 'what makes an occurrent thought ... to be conscious ... is that it is made available to further thought through the operation of a regular feed-back loop whose function is to make such thoughts available to yet further thoughts' (1996, 195). Such thoughts are *conscious* since they are available as the contents of higher order thoughts, but the higher order thoughts may not actually occur, and their failure to occur does *not* by itself prevent the lower order thought from being conscious.

Here Carruthers has introduced a fundamental distinction among the kinds of HOT theories available to us. Let's call versions of HOT theory, such as the typical one outlined above, that require that the higher order thought actually occur in order for the lower order state to be conscious an *occurrent* HOT theory. A HOT theory like Carruthers that requires only that the lower order state be *apt* to cause the appropriate higher order thought can be called a *dispositional* HOT theory. Carruthers goes further, of course, in demanding that the higher order thought be conscious (if it occurs) so we might add a new form of HOT theory naturally labelled dispositional *conscious* HOT theory. We can then diagram the set of theoretical options thus:

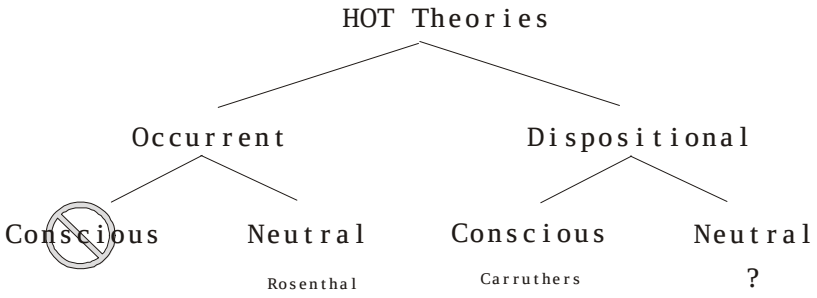

Figure 1

To its credit, the dispositional version of HOT theory avoids two 'cognitive overload' problems. In addition to avoiding positing a literally infinite number of ever higher order thoughts (an advantage it shares with the occurrent neutral HOT theory), it also limits the number of occurrent thoughts at any one level. Carruthers notes that any conscious perceptual experience is vastly rich in content and any occurrent higher order thought theory of consciousness requires either one higher order thought for every distinctive element of that content, or a single 'super-thought' that duplicates all the content in the perceptual experience. Carruthers regards it as highly implausible that whenever we are, for example, perceptually conscious, we have such a vast number of higher order states (or one with exceptionally rich content) that appear to do little but duplicate the content of their lower order counterparts.

It seems to me that this latter point is a genuine advantage of the dispositional HOT theory. But there are serious disadvantages that create such difficulties for both dispositional conscious and dispositional neutral HOT theories (and especially the former) that this advantage becomes insignificant. To these I now turn.

In occurrent HOT theory, if we can prevent the higher order thoughts from occurring we can prevent the lower order states from being conscious. Thus, for example, if we had some kind of machine, call it a 'neural meddler', that interfered with the cognitive mechanisms which normally permit the lower order state to cause the higher order thought which makes the former conscious then we would have a 'consciousness inhibitor'. In the case of the dispositional HOT theories, things are somewhat more complex. Consider a match. It has a disposition to light if struck. Does it have this disposition in a vacuum, where it cannot light whether or not it is struck? I'm not sure if there is a definite answer to this question, but I feel sure that if we meddled with the match itself – as opposed to varying the external circumstances – we could destroy the disposition. If, for example, we covered the match with a coating of wax that prevented it from lighting we would have 'disabled' the disposition. Similarly, if we meddle with someone's brain so that the lower order states are made incapable of causing the appropriate higher order states we have, as the phrase 'incapable of causing' suggests, eliminated the disposition to cause higher order states. And thus our meddler has eliminated consciousness, in just the same way that consciousness would be eliminated by the prevention of the *occurrence* of higher order states under the dictates of the occurrent HOT theories. Another way to put this point, in terms that Carruthers favours, is that the neural meddler prevents the lower order states from being *available* to consciousness, and without this availability there can be no consciousness of those states.

But if this is so, dispositional HOT theories face a serious objection. Consider a modified neural meddler which blocks the disposition to cause higher order states only for those states and for those time periods when the lower order states would not, *in fact*, cause a higher order

thought. (To continue the analogy with the match, we can imagine a device that somehow only coats with wax matches that are not actually going to be struck.) Such a meddler would be extremely difficult to produce in practice (not that the original meddler is exactly 'off the shelf' machinery just yet!) since it requires an ability to know under what conditions a lower order thought will occur but will not actually cause a higher order thought. If we suppose that it is in principle possible to predict the operation of a brain at the neural level, then the information from such predictions could be fed into the meddler so that it would be active only for those lower order states, and only at those times when these lower order states do not actually bring about a higher order thought. Of course, the practical difficulty of developing such a meddler is irrelevant to the point of principle at issue here.

Now, a curious consequence follows. Let us take two people, one with a modified neural meddler attached to his or her brain and one without, but who otherwise begin in identical neurological states (and in identical environments). Both of these people will have an identical history of higher order thoughts, since the meddler will never prevent a lower order state that actually was going to produce a higher order thought from causing that thought. They will also have identical histories of lower order mental states, for the meddler has no effect on these. Yet they will be markedly different in their states of consciousness, for the unfortunate person with the meddler will lack an entire set of conscious states enjoyed by the other – namely those that as a matter of fact do not produce any higher order thoughts (but – in the unmeddled brain – could). This is a necessary consequence of the dispositional HOT theory, since it is explicitly designed to allow that states are conscious simply if they are able to produce higher order thoughts, not if they actually do produce those thoughts.

This consequence of dispositional HOT theory is not only curious, it is disturbing and implausible. There is absolutely no difference in the behaviour of our two individuals and no difference in their history of mental states. There is nothing to mark the difference between the two of them except an *entirely inert* meddler. The meddler never has to actually function to produce this drastic alteration in consciousness. That is, two brains identical in their neural states and their dynamics will differ in consciousness solely because one has an *inert* piece of machinery within it! No such implausibility follows from *occurrent* HOT theory.

Perhaps the implausibility can be underlined if we imagine that the modified meddler is oscillating between being 'off' – incapable of functioning – and 'on' – capable of functioning, even though it never will. There will be a corresponding oscillation in consciousness (more conscious states when the meddler is disabled, fewer when it is enabled) which would presumably be very striking but in fact would be seemingly be completely unreportable by the subject despite being a huge difference in phenomenological experience.

There is a kind of 'inverted' version of this objection. Consider a device that increases or *boosts* the aptitude of a lower order mental state to produce higher order thoughts (call it a boosting meddler)[1]. Under occurrent HOT theory, such a device would increase the number of conscious states inasmuch as it would increase the number of higher order thoughts brought about by lower order states. This effect would be apparent under dispositional HOT theory as well. But, as before, dispositional HOT theory permits a more subtle tampering with consciousness. For we can imagine implanting a modified boosting meddler, which only boosts the aptitude to cause

[1] In terms of our match analogy, a boosting meddler might be something that modifies the match so that it will ignite at a lower temperature, thus increasing its aptitude to light when struck.

higher order states in lower order states which (1) would not otherwise have the power to bring about higher order states and which (2) even with the boosting meddler in place will not quite have enough power to *actually* cause a higher order thought. (Again, this would require remarkable knowledge (and fore-knowledge) of how a particular neural system is going to work, but we can grant this knowledge in principle[2].) So, even though there is absolutely *no* increase in the number of higher order thoughts, there is a striking increase in consciousness. Perhaps this result is less implausible than the previous one, insofar as in this case the meddler actually does some work, but it remains very implausible.

Such consequences of the dispositional HOT theory are difficult to swallow. I do not, however, think they are the most serious objection that can be made, although the more serious objection holds against only the dispositional *conscious* HOT theory. Recall that this theory requires that the higher order thought which confers consciousness upon the lower order state to be itself a conscious thought. This does not produce the viciously infinite hierarchy of ever higher order thoughts because a state need merely be disposed to cause a higher order (conscious) thought to be itself conscious.

Let us consider whether there is a limit, imposed by the finiteness and particularity of our cognitive architecture, on the complexity of nested thoughts we can entertain. It seems as certain that there are thoughts of the form 'I am aware that I am aware that I am aware .... that P' which are sufficiently deeply nested as to be in fact entirely incomprehensible, given normal human cognitive capacities, as that there are numbers which are too big for my calculator to multiply. I

[2] For the match example, a modified boosting meddler only modifies those matches that are in fact not going to be struck. So the modified boosting meddler never makes any difference to what matches actually light or do not light.

suspect that this limit is in fact quite low – at least for me – as I have a good deal of difficulty being aware of just a few levels of awareness.

Suppose, then and without loss of generality, that a level of nesting at which nested thoughts become unentertainable is n. Then it is easy to show that no thought of level n-1 could be conscious. For if it is impossible to entertain, because of inherent cognitive limitations, a nested thought of level n, it cannot be the case that any thought is apt to cause a nested thought of level n. (Any more than a match could be disposed to light if struck if, because of inherent chemical conditions, no striking could raise the temperature sufficient to ignite it.) Obviously, if a level n-1 thought cannot cause a level n thought it cannot cause a level n *conscious* thought, and so, according to the dispositional *conscious* HOT theory, the level n-1 thought cannot be conscious. But if no level n-1 thought could be conscious then no thought of level n-2 could be disposed to produce a *conscious* level n-1 thought (even if it might be disposed to produce a level n-1 thought). Hence, no level n-2 thought could be a conscious thought. This argument by 'vicious descent' can clearly be generalized as far as necessary, with the disastrous result that *no* mental state can be conscious according to the dispositional conscious HOT theory.

Biting (instead of dodging) the bullet, it might perhaps be replied that there is no level of nested thought which is impossible to entertain. It is true that the concept of thought, and the entertaining of thoughts, admits of no intrinsic, purely abstract, limitation in complexity. But the point here is that there is a natural limitation, imposed by the cognitive architecture implemented by the finite brain, to the complexity of entertainable thoughts[3]. This limitation is based upon

[3] The distinction here is of course reminiscent of Chomsky's between performance and competence. While competence – the abstract structural possibilities of language – is not limited by natural constraints, it would be ludicrous to claim that there was an actual human disposition

natural law which reveal to us the range of possible dispositions which our neurological machinery can instantiate. The dispositional conscious HOT theory cannot avail itself of the mere abstract structural possibilities of thought, since it depends upon the actual dispositions inherent in the thoughts we *actually* possess.

I think the only conclusion that can be drawn is that the dispositional *conscious* HOT theory cannot be correct (at least for finite, real-world cognitive systesm). Within the field of the HOT theories, this seems to leave both of the two neutral theories, either dispositional or occurrent as possible contenders. For the neutral version of dispositional HOT theory is immune to the vicious descent argument which yields only the 'theorem' that there is a level of nested thought of which we cannot be conscious (which is one level below the level at which we can no longer entertain thoughts of that complexity at all). But the objections given earlier above seem to me to tell quite clearly in favour of the occurrent version as against the dispositional form. It is of course also possible, and I think likely to be true, that no HOT theory will provide an acceptable account of consciousness.

William Seager
University of Toronto at Scarborough

---

either to produce or to understand sentences with, say, 10^{37} nested relative clauses.