**Rocco J. Gennaro, “Defending HOT Theory and The Wide Intrinsicality View: A Reply to Weisberg, Van Gulick, and Seager,” Journal of Consciousness Studies book symposium on *The Consciousness Paradox* (MIT Press, 2012)**

*[Final copy in JCS vol 20 (11-12), 2013, pp. 82-100]*

I first want to thank all three commentators: Josh Weisberg, Robert Van Gulick, and William Seager. I appreciate the time they put into reading my book and working on their thought provoking commentaries. This symposium grew out of an author-meets-critics session at the Central APA conference in 2013. I also wish to thank the session chair (William Robinson) and the Central APA committee chair (Anne Jacobson) for arranging the session in the first place. I cannot of course respond to every single point made by my commentators, but I’ll focus here on what I take to be the most important and interesting criticisms.

**Misrepresentation** (Weisberg and Van Gulick)

Josh Weisberg focuses largely on the central problem of misrepresentation, including so-called “empty” or “targetless” HOT cases, which is probably the most serious problem for HOT theory. Robert Van Gulick also raises questions about my treatment of this problem. The main example of misrepresentation used both in my book and in the commentary comes from the following hypothetical case from Levine (2001) which I quote in my book as follows:

Suppose I am looking at my red diskette case, and therefore my visual system is in state R. According to HO, this is not sufficient for my having a conscious experience of red. It’s also necessary that I occupy a higher-order state, say HR, which represents my being in state R, and thus constitutes my being aware of having the reddish visual experience. . . . Suppose because of some neural misfiring (or whatever), I go into higher-order state HG, rather than HR. HG is the state whose representation content is that I’m having a greenish experience, what I normally have when in state G. The question is, what is the nature of my conscious experience in this case? My visual system is in state R, the normal response to red, but my higher-order state is HG, the normal response to being in state G, itself the normal response to green. Is my consciousness of the reddish or greenish variety? (Levine 2001, p. 108)

So the question is: In such a hypothetical case, what “is it like for the subject”? The higher-order (HO) content is green whereas the first-order (FO) content is red. Levine initially points out that we should reject the following two possible answers:

Option 1: The resulting conscious experience is of a *greenish* sort.

Option 2: The resulting conscious experience is of a *reddish* sort.

I agree with Levine on these two options. The main problem is that one wonders what the point of having *both* a LO and HO state, on HOT theory, if only one of them determines the (color of the) conscious experience. Levine briefly considers two further options which I find preferable, but dismisses them far too quickly:

Option 3: “When this sort of case occurs, there is no consciousness at all” (Levine 2001, p. 108).

Option 4: “A better option is to ensure correct representation by pinning the content of the higher-order state directly to the first-order state” (Levine 2001, p. 108).

Weisberg first notes that everything WIV invokes to deal with misrepresentation could be adopted by standard HOT theory. This *may* be true and would be welcomed by me, but neither he nor David Rosenthal has even tried option 3 or 4. Quite the opposite; they criticize it. If they did endorse option 3 or 4, however, it seems to me that they still would move somewhat closer to the WIV since there would then have to be some account of the required more intimate relationship between LO and HO mental states. In any case, their favored option 1 is still much more problematic to me. Weisberg rightly explains that Rosenthal has opted for 1, arguing that what really matters is how things seem to the subject—if we can explain that, we’ve explained all that we need to. Somehow, then, the HO representation *alone* is what matters. But Weisberg correctly points out that I reject this move. After all, doesn’t it defeat the purpose of HOT theory which is supposed to explain state consciousness in terms of the *relation* between *two states*? Moreover, I further explain in my book that “HOT theory is supposed to be a theory of (intransitive) state consciousness; that is, the *lower-order state* is supposed to be the conscious one” (2012, p. 60). (I will also return to this overall problem below with respect to targetless HOTs where a similar point applies.)

Weisberg insists that it is not enough to *stipulate* that misrepresentation does not result in a conscious state. We need to understand within a reductive framework how this possibility can be ruled out. However, I do not merely “stipulate” anything. As he does acknowledge to some extent, I *argue* for it and greatly elaborate on this idea especially in chapters four and nine, for example, with reference to various brain mechanisms, feedback loops, and Kantian considerations such as the necessary cooperation of the sensory (“sensibility”) and cognitive (“understanding”) aspects of mind. And, of course, I do locate my view within a reductive framework. For example, I quote the following passage in my book: “High level areas must send feedback to lower-level areas . . . so that neural activity returns in full circle” (Baars and Gage 2010, p. 173). And perhaps the most crucial point is that part of the reason for this may simply be that “higher areas need to check the signals in early areas and confirm if they are getting the right message” (p. 173). The subject’s HO concepts must *recognize* the incoming representations *as* something or other. I reference many others along these lines such as Feinberg, Edelman, Tononi, Weiskopf, and Cleeremans. I take this also to be an initial answer to Van Gulick’s questions: “Why is a match required for a consciousness? One can simply define a CMS (conscious mental state) in such a way that a match is required, but why should we accept such a definition?” (I should also note that I explicitly reply to Weisberg’s 2008 and 2010 papers in section 4.5.6. As far as I can see, he offers no further counter-reply to what I say there so I won’t elaborate here.)

Weisberg goes on to say thatat times I seem to accept that it is the HOT (or “MET” for metapsychological thought) which fully accounts for what it’s like for the subject, so long as there is a match with the LO state. He mentions that one of my main motivations in chapter six is to support the idea that HO conceptual representation can account for the detail of conscious experience. However, as Weisberg rightly notes, I also hold that if something LO is not conceptualized in *any* way, there will be nothing it’s like for the subject (since there wouldn’t be a HOT at all). But his characterization of what I say in chapter six is not quite right or, at least, very oversimplified.

The entire *initial* premise in question is as follows:

“Whenever a subject S has a HOT directed at e, the content c of S’s HOT determines the way that S experiences e (provided that there is a match with the lower-order state).”

Nonetheless, after significant discussion in chapter six, the *final premise*, labeled (2″) says:

“Whenever a subject S has a HOT directed at e, the content c of S’s HOT determines the way that S experiences e (provided that there is a *full or partial* conceptual match with the lower-order state, *or* when the HO state contains more specific or fine-grained concepts than the LO state has, *or* when the LO state contains more specific or fine-grained concepts than the HO state has, *or* when the HO concepts can combine to match the LO concept).”

The reasons for the above further qualifications above are many and discussed at length in chapter six, but Weisberg does not really engage at all with these crucial parts of my book. I basically try to explain what happens in some abnormal cases (e.g. visual agnosia) and in some other atypical contexts (e.g. perceiving ambiguous figures such as the vase-two faces) where mismatches might occur between the HOT and LO state. This also arises in chapter six with respect to well-known fineness-of-grain color cases.

For example, in the much discussed “the two paint chips case,” we might acquire conceptual content at the lower-order level (such as RED26 and RED27), which is more detailed and fine grained than those concepts present at the HOT level, at least initially. Any given subject, however, may have more coarse-grained concepts such as DARK RED and LIGHTER in the HOTs and will thus have only the more coarse-grained color experience when presented with the two paint chips separately. The crucial point here is that even though there isn’t an exact match between all the concepts at each level, the HOT’s concepts, which are the concepts that S already has, are consistent with, but more general than, the LO concepts. Suppose that red27 is lighter than red26. A subject S may not be able to distinguish red27 from red26 when presented with each separately. However, S may still be able to distinguish them when presented together because S may have, say, the concepts RED27 and DARKER or simply DARK RED and LIGHTER. Much the same is true for other cases I discuss in chapter six, such as recognizing multi-sided figures, differentiating very similar sounds, and recognizing similar animals.

Associative visual agnosia has been described as a case where a subject has a conscious experience of an object without any conceptualization of the incoming visual information. There appears to be a first-order perception of an object without the accompanying concept of that object (either first- or second-order, for that matter). But it seems to me that it is simply an unusual case where the HOT does not *fully* match up with the first-order visual input. That is, we might view associative agnosia as a case where the “normal,” or most general, object concept in the HOT does not accompany the input received through the visual modality. I use the example of seeing a whistle but not being able to identify it. However, the agnosic can still apply, say, the concepts ROUND or SILVER to that object. So there is a *partial match* instead, a case where a HOT might partially recognize the LO state concepts. So associative agnosia would be a case where the LO state could still register a percept of an object O (because the subject still does have the concept), but the HO state is limited to some features of O. Bare visual perception remains intact in the LO state but is confused and ambiguous, and thus results in a different phenomenological experience. It may seem that this way of handling associative agnosia is at odds with my treatment of Levine’s case. Notice, however, that in his case there is a mismatch in the sense of there being entirely *incompatible* properties represented in the LO and HO states. In Levine’s red and green diskette case example, the case cannot be both entirely red and green. On the other hand, in an ambiguous figure case, ambiguous LO content can be “recognized” by a HOT in two incompatible ways (as a vase or as two faces), resulting in two very different perceptual experiences. A similar point applies to the agnosia case: ambiguous and thus more general LO content is accompanied by more specific HO content. So there is an important difference between Levine’s case and scenarios where the LO content is ambiguous or more general than the HO content. HOTs can thus also serve the purpose of narrowing down the conscious perceptual content in these cases. The HOT is recognizing only part of what is present at the LO level. According to the WIV, however, if there were *no* LO state at all, then then there would be no conscious state at all. And if there are incompatible concepts in LO and HO, then there would be no conscious experience involving those concepts.

My responses above (and in chapter six) also serve as detailed answers to Van Gulick’s questions: What if the HOT (MET) is *mostly right* about MS but not completely so? Would that prevent integration? As far as I can tell, Van Gulick also does not engage much with chapter six when asking about “slight mismatches” or “partial mismatches.” In the above premise 2’’, I do use the disjunctive phrase “full or partial” mismatch. I also explicitly mention “partial matches” in discussing the above cases. Although it is true that I could have sometimes been clearer that, in the diskette case, I meant that there would at least be no *color* experience accompanying any visual perception, I did for example say at one point: “…if specific brain lesions are involved, perhaps the subject would at least experience a loss of color vision (achromatopsia) with a diskette case perception and...at the very least, if a misrepresentation occurs between some of the relevant concepts in M and MET, then *that aspect* of the conscious state would not exist” (2012, p. 62). And I take it that Levine also took “no consciousness” to be referring only to the color aspect of the diskette experience.

To elaborate further: Suppose we have concepts A, B, and C as part of a first-order mental state M and yet a HOT containing A, B, and D. If so, then the resulting conscious experience would only include A and B, especially if C and D are incompatible properties like the different colors in Levine’s case. Similarly, if one has concepts A, B, and C as part of M and a HOT containing only A and B, then the resulting conscious experience would only include A and B. But I fail to see why this would strike anyone as odd. We consciously experience the world through the prism of our concepts and if we do not recognize some aspect of the incoming information (or even a first-order perception), then we would not experience that perception *as* having that aspect. Moreover, we are again presumably talking about some rather abnormal phenomena, such as achromatopsia (loss of color vision). Anyone immersed in consciousness research realizes just how many bizarre abnormal phenomena there are, including motion blindness, neglect, prosopagnosia, and many others. In these cases, one’s conscious experience is indeed typically missing just one or maybe two central aspects of a normal experience, such as experiencing motion or recognizing a face. Yet there is still of course some kind of conscious experience in each case. So if these cases can *actually* happen, then I don’t see why my way of potentially allowing for admittedly odd conscious experiences with partial matches between HO and LO concepts is so hard to accept. Thus, I also take much of the above (and below, for that matter) to blunt the force of Van Gulick’s demand for more “details” and “further explanation” about why and how the integration works on the WIV. There is no doubt that more could always be said, but I think there is significant conceptual and neurophysiological detail in my book already, including in chapters six and nine.

Weisberg insists, however, that the real worry is about the *theoretical* possibility of mismatch. Could God, or neuroscientists of the distant future, make it so that a MET is sustained by something other than the matched LO state? Given the nature of representation, he sees no way to deny this remote possibility. What if the HO state was somehow triggered anyway?

The answer to this last quesstion is nothing, that is, there would be no resulting color experience in the conscious state. But I also point out (p. 64) how the following two claims are really two sides of the same coin:

(1) There is no resulting conscious state when a misrepresentation does occur, and

(2) Misrepresentations cannot occur.

I am honestly unsure which view is preferable but either one seems plausible to me. My sense is that (2) strikes others as more problematic because it is ruling something out in a stark way. But, as a practical matter, it doesn’t really matter very much. As I say in the book, statement (2) should really be understood as:

(2′) Misrepresentations cannot occur between M and MET and still result in a conscious state.

Perhaps I should clarify further here:

(2’’) Misrepresentations cannot occur between M and MET *and* result in a conscious experience reflecting mismatched and incompatible concepts.

Weisberg goes on to point out that Chalmers (1995) appeals to a sui generis acquaintance relation to explain our awareness of our conscious states. Levine also thinks something like that must be the case and argues that its presence is just another facet of the explanatory gap. Weisberg says and I agree that it is not at all clear how such a proposal can provide a reductive explanation without remainder of conscious experience. So I think there is real danger for the reductionist in going down this road.

So I agree with Weisberg (and Kriegel) here in opposition to any “acquaintance” alternative. I am not sure what to make of this “sui generis” alternative (2012, pp. 107-108). It seems at best to trade one difficult problem for an even deeper puzzle, namely, just how to understand the allegedly intimate but *non*representational “awareness of” relation between HOT and M. I am also inclined to treat the relation in question as representational for reasons given throughout the early chapters of my book. But, contrary to Weisberg’s argument, I do think that one can both adopt a WIV-like view and maintain a reductionist programme.

So with respect to Weisberg’s final “dilemma argument,” I would still reject his premise 2 (if [misrepresentations] can occur, Levine’s options 1 & 2 are the only ones available and HOT and WIV collapse into one another). And even if we accept premise 3 (“if misrepresentations cannot occur, we must posit a strong enough link to block misrepresentation”), I would reject premise 5, namely, that “if the link involves a causal connection, HOT and WIV collapse into one another.” I think that Weisberg oversimplifies the integrative relation between HO and LO parts on my view. Finally, once again, it is not so much that misrepresentations cannot occur *as such*, but rather, as I said above, that “misrepresentations cannot occur between M and MET *and* still result in a conscious experience reflecting mismatched and incompatible concepts. It is the *combination* of these claims that is most important.

**Sums and Complexes** (Weisberg and Van Gulick)

Somewhat related to the misrepresentation question, both Weisberg and Van Gulick challenge the superiority of the WIV over standard extrinsic HOT theory and even that the WIV is importantly different than EHOT theory. For example, at one point, Weisberg still wonders if there is no difference between HOT theory and WIV beyond the WIV’s insistence that we count the various representational bits as all part of one state. This has also led Kriegel (2009) to suppose that the WIV is merely a “sum” view, rather than a real “complex” view.” Van Gulick also asks: “What other integrative relations might one appeal to justify the claim a CMS is a complex and not a mere sum?” The basic difference is that a mere “sum” says that what makes two states part of a single mental state is merely our decision to treat them as such. This is a purely verbal or stipulative difference. In contrast, a “complex” is a sum whose parts are essentially connected in a certain way.

But of course I do respond to this charge explicitly and at length in section 4.5.3 of my book. Moreover, Kriegel’s view over the years has even seemed to move closer to the WIV, especially when he settles on so-called “indirect” self-representation in Kriegel 2009. He explicitly says that only an indirect self-representation is applicable to conscious states such that one part is directed at another part (Kriegel 2009, 215–226). In any case, all of chapter five in my book is devoted to a critique of “self-representationalism,” including Kriegel’s view. In addition, on a more conceptual level, I develop an account of the structure of conscious states in some detail via the notion of mereology, which is the theory of parthood relation (section 4.5.4 “More on Parts and Wholes”). I develop a way of understanding the WIV as “complex” states via such notions as “overlap” and “underlap.” Here is one quote: “we might say that M (x) *underlaps* MET (y), since there is a CMS (z) such that M is part of CMS and MET is part of CMS. However, there can also still be some *overlap* between M and MET insofar as a psychologically real relation holds between M and MET. On the neural level, much the same seems reasonable, since, for example, there are overlapping parts of feedforward and feedback loops that extend from M to MET and vice versa. If we construe the vehicles of M and MET in such a manner due to their essential integration, then M and MET can overlap in addition to the underlap” (2012, p. 95). As far as I can see, neither Weisberg nor Van Gulick delves into my analysis here, but I certainly agree that the details do matter.

**Targetless HOTs** (Weisberg)

Weisberg then returns to the misrepresentation problem and, more specifically, “targetless” or “empty” HOT cases where there is no target at all for a HOT. Rosenthal frequently refers to confabulation and dental fear as examples of targetless or “hallucinatory” HOTs. Confabulation typically involves falsely thinking that one is in an intentional state or making erroneous claims with regard to the causes of one’s intentional states. Dental fear occurs when a dental patient seems to experience pain even when nerve damage or local anesthetic makes it impossible for such a pain to occur. Perhaps the patient’s fear has been mistaken for pain, but it may also be that the patient has a HOT about being in pain when in fact no pain is present. Weisberg characterizes these phenomena as plausible cases of real-world HO misrepresentation which I merely dismiss as cases of introspective error or as cases of “top-down activation” of LO states. Indeed, I do argue that it is often best to think of these kinds of cases as introspectively generated or caused LO states. Thus, I fail to see how they are even relevant to any alleged misrepresentation between an *unconscious* HOT or MET and a lower-order mental state.

Weisberg explains that in a case of introspective error, I am consciously thinking that I am in pain or I am angry but I can misrepresent the content of my current conscious state. So there must be a conscious thought present in such cases. But, he thinks, that does not seem to accurately capture the phenomenology of cases where one’s expectations lead to strange experiences when those expectations are not met. For example, if I expect milk in my glass and get orange juice instead, I have a gross experience of something pulpy and sour and nasty (even though he likes OJ). He claims that the problem for the WIV is that this explanation can never be the case.

I don’t see why. In response, first, if I drink some OJ and do apply the right concept to it, then I will experience the conscious tasting of OJ (even if it is unexpected). In this case, there might *also* be another mistaken introspective “expectation” of milk but that will not be matched by the LO input. So the OJ conscious state wins out even though it was unexpected. After all, it is precisely the fact that he actually has the OJ experience that causes him to react in a surprised and disgusted way.

Second, let’s return to Weisberg and Rosenthal’s acceptance of Levine’s option one and apply it to targetless HOT cases. Here’s one very important passage from my book (2012, p. 97): “It is admirable that Rosenthal so clearly wishes to make room for an appearance/reality distinction with regard to our own mental states. I agree with the notion that our introspective states are fallible and may misrepresent our “selves” and our mental states. But this distinction applies at the introspective level, not within first-order world-directed conscious states. If there is an inner analogy to an illusory or hallucinatory first-order conscious state directed at an outer object, it must be a conscious state (= introspection) directed at a mental state. But then this is not a case of an appearance/reality difference between an unconscious HOT (or MET) and a mental state M. This is again why we should reject Rosenthal’s endorsement of Levine’s option one for misrepresentation cases. A lone unconscious HOT without its target is not a case of fallible introspection.”

In any case, I remain very puzzled by any defense of option one. As Levine says, “Doesn’t this give the game away? . . . Then conscious experience is not in the end a matter of a relation between two (non-conscious) states” (2001, 190). Once again, and especially with respect to targetless HOTs, I don’t see how an *unconscious* HOT *alone* results in a conscious mental state. Moreover, according to Rosenthal, HOTs themselves have no qualia. And, again, if I am right that Rosenthal and Weisberg really have introspective cases in mind here, then these examples are irrelevant to any possible misrepresentation between an *unconscious* MET and a lower-order M.

**Introspection** (Van Gulick and Seager)

Van Gulick and Seager raise some very interesting questions about my view of introspection which are also interconnected with my treatment of misrepresentation.

Van Gulick alleges that I hold the following two arguably inconsistent views:

(A) Introspective error involves a mismatch between a HOT (second-order) and a further HOT\* (third-order) directed at HOT.

(B) Conscious mental states always involve an appropriately integrated conscious mental state (CMS), whether this is a first-order CMS or a second-order introspective CMS.

He reasons that if A is true then B seems false since there could not be a mismatch in the case of a second-order introspective CMS. If B is true, then A is false since I cannot then account for introspective error between HOT and HOT\*.

The short answer to this charge is that I do not hold (A). I am not quite sure why Van Gulick thinks that I do. On my view, introspective error involves a mismatch between a second-order conscious MET or HOT and a first-order M. This not only better fits the phenomenology of introspection (i.e. consciously thinking about a mental state M), but also reflects my basic description of introspection on the WIV. For example, I say “when I introspect my perception [or any mental state], a first-order mental state is rendered conscious by a complex higher-order state. Thus introspection involves two states: a lower-order noncomplex mental state that is the object of a higher-order conscious complex state” (2012, p. 58). So Van Gulick’s charge is easily answered.

Nonetheless, I wish to elaborate further on why, on my view, there is only the one “complex” integrated CMS during introspection. This also anticipates Seager’s interesting criticism below.

HOT theory is primarily concerned with explaining how a mental state becomes *intransitively* conscious. On my view of introspection, there is only the one “complex” integrated CMS during introspection because only the second-order conscious MET (or HOT) is intransitively conscious and thus a very intimate relationship between MET and MET\* must obtain analogous to when a first-order M is intransitively conscious. However, since there is no such intimate relationship between M and a *conscious* MET (or HOT) in introspective cases, then there is room for error in these cases. The conscious MET is *transitively conscious* of M but M is not itself intransitively conscious. In the introspective case, M is the *object of* a transitively conscious state. In introspection, it is true that M is conscious but only as the object of a conscious MET directed at it whereas, in the first-order case, M is intransitively conscious and directed at the outer-world. Like cases of outer-directed hallucinations or illusions where the objects either don’t exist at all or are really different from the way they appear, when one has a conscious MET in introspection the objects (or target mental states) might not exist or are different from the way they appear. This is the main reason why the higher-order CMS (i.e. conscious MET with MET\*) is, we might say, “detached” from M whereas an unconscious MET is part of the same CMS as M.

Seager finds a related difficulty with my view, namely, that the entire HOT theory account of introspection is false insofar as it misrepresents the core phenomenology of introspection. He thinks that, on the HOT account, when I introspect I should become conscious of a thought which is about some mental state.

But here again I think that the intransitive/transitive distinction is very helpful. Again, when a mental state M is the object of introspection, the introspective state is transitively conscious of M. The conscious HOT itself is still intransitively conscious but its target isn’t, so I don’t see the problem here either. Thus, it is incorrect or at least misleading to say that “when I introspect I should become conscious of a thought which is about some mental state.” The conscious HOT (or MET) in question is accompanied by an *unconscious* thought (HOT\* or MET\*) about that conscious HOT.

Seager pushes even further regarding what happens when one transitions from a first-order conscious state to an introspective state, on my view. He says that what happens is bizarre. The component of S2 (MET) which carries the first order content (the part of S2 which corresponds to S1) is destroyed, the metapsychological part of S2 remains, and a new, yet higher order metapsychological (meta-metapsychological) state appears within S2. This new WIV structured state, call it S3, engenders the coming to consciousness of the metapsychological content which makes this episode into a case of conscious introspection. Meanwhile, the first order content remains conscious but no longer is part of a WIV structured state. If this picture of introspection is correct, Seager continues, then it endorses two distinct ways that a mental state can be conscious. That is, both the WIV mechanism of consciousness and the original HOT theory mechanism are capable of engendering consciousness. Seager thinks that this makes the WIV approach seem otiose and epicyclic.

I don’t think that S2 (MET) is typically destroyed at all in the transition, but it is an open and good question as to how such a transition might actually occur in the brain for example. If I am right that unconscious HOTs (or METs) need not occur in the prefrontal cortex (PFC), then the transition to conscious HOTs will bring in the PFC in a way that is directed at preexisting unconscious HOTs. This, in turn, will thus require further cognitive and neurophysiological “distance” or “detachment” from M to be introduced, which helps to explain the greater fallibility at this point. Indeed, as we all agree, one could even generate an introspective state with any target at all. For example, I may be consciously thinking about or searching for a lost memory or nonexistent desire. As for Seager’s point that my view “endorses two distinct ways that a mental state can be conscious,” this is fine but I fail to see why this is a problem for any theory of consciousness. As I said above, the conscious MET is transitively conscious of M but M is not itself intransitively conscious (only the MET is intransitively conscious). So, in the introspective case, M is the object of a transitively conscious state. There has always been room for two different ways for a state to be conscious in this sense, but the focus tends to be on intransitive state consciousness. Indeed, wouldn’t a non-HOT theorist also have to allow for these two ways as well?

Finally, Seager tells us that “…the natural response is to recall that introspection involves two distinct activities. One is the activity of categorizing or conceptualizing our mental states. This task requires that one apply to one’s own experience concepts of various kinds of mental states (as, for example, fear versus anger, etc.). This is an error prone activity. The second kind of introspective activity is the simple appreciation of the stream of consciousness as it occurs. This is still a cognitive activity but it is not one demanding any judgments which could be either right or wrong. (Sometimes I think that the HOT approach mistakes consciousness simpliciter with this reflective but non-categorizing appreciation of the stream of consciousness.)”

I do not see that mistake here. What Seager calls the “first” kind of introspection is what a HOT theorist typically means by “introspection.” What he calls the “second” kind of introspection is an “unconscious HOT” though we would insist that conceptualization also occurs at that level. This is important to keep in mind and we want to be sure that we do not conflate “introspection” with “unconscious HOTs,”

**Lower-Order Concepts; HOGS** (Van Gulick)

Van Gulick mentions another possible problem for my view. Are higher-order thoughts really necessary for concept application? Can complex concepts be applied in unconscious mental states? Van Gulick also explains that “higher-level" can be ambiguous between “metapsychological” and “referring to more complex features of the world.” The former tends to be used by philosophers whereas psychologists tend to use the latter.

Van Gulick is correct that there is sometimes ambiguity between how we philosophers might use the term “higher-order” as opposed to many psychologists. I certainly do not mean to imply that “higher-order” (in my sense) refers to complex features of the world. I also certainly do not suppose that all cases of feedback loops in the brain indicate the presence of HOTs. Van Gulick also asks, for example, about semantically based priming effects and subliminal perception. The answer is that there most certainly can be concepts, even some complex concepts, applied in unconscious mental states. So HOTs are not the *only* place where concepts can reside or be applied. However, it is the fact that there are HOTs (with their constituent concepts) directed at mental states which renders otherwise unconscious states conscious. (See sec 4.5.6 in my book for a bit more on this topic. Note also that I address the more fundamental issue of mental content and concept acquisition in other parts of my book.)

Van Gulick also notes that I do refer, albeit briefly, to his “HOGS” (higher-order global states) theory at several points in *The Consciousness Paradox*, noting that it offers an alternative account of how conscious states might involve an “implicit reflexive awareness” of themselves. Van Gulick says that even if Kriegel’s focal-peripheral distinction will not do the job that self-representationalists hope, there may be other ways to defend and explain the claim that everyday conscious states involve some form of self-reflexive awareness, and that when we are in such states we are in some way aware not only of external objects and first-order mental states, but also of our own self-awareness. Van Gulick mentions HOGS theory in this context. According to Van Gulick, when a sensory state is recruited into a larger dynamic state that is the momentary substrate of a subject's conscious experience, the content of that sensory state is transformed into a phenomenal or experiential mode that embodies aspects of self-awareness in its intentional structure. For my own part, I find it difficult to understand Van Gulick’s frequent use of the expression “implicit meta-intentionality” or “reflexive self-awareness” unless we (at least often) think of it as an unconscious or conscious part of a conscious whole. So I am not quite sure of his answer to numerous questions: Is the implicit reflexive aspect itself conscious? If so, how does it differ from Kriegel’s view? How does it avoid problems with Kriegel’s view? How does it avoid circularity or regress worries? What happens during introspection? Is it explicitly nonreductionist? Are there state-parts? Is there a representational relation between the reflexive self-awareness and the rest of the state? Do HOGS have two contents? Do HOGS have two attitudes? How do HOGS acquire mental content? I am not aware of Van Gulick having answered too many of the above questions, but perhaps I am mistaken. Of course, I wouldn’t expect him to answer the above questions in his commentary on my book.

**The Transitivity Principle** (Seager)

In my book, I frame the transitivity principle as follows:

(TP) A conscious state is a state whose subject is, in some way, aware of being in. (p. 28).

Seager first explains that he does not feel the intuitive pull in favor of TP. He says he has to work himself up into a rather peculiar contemplative state to begin to feel TP’s attractions. If, for example, he imagines himself “reflectively considering” what is passing before his mind, or “self-consciously introspecting,” he says he can understand that it would be very strange to be enjoying a conscious mental state without being aware of it. But that is a very unusual state for anyone to get themselves into and most if not all non-human animals never enter into such states. Moreover, Seager thinks that certain other kinds of conscious states weaken the case for TP, namely, super intense experiences that require total attention, for example, experiences of great fear or total engagement in playing hockey at full speed.

Some of the above makes me worry that Seager might be conflating “introspection” with “unconscious HOTs,” at least to some extent. He uses terms such as “reflectively considering” and “self-consciously introspecting.” But of course when one has an *unconscious* HOT, one’s consciousness is directed at the outer world, according to HOT theory. The intuitive pull of TP comes from the common sense idea that we are *in some sense* aware of our conscious states. We should *not* of course read this as holding that we are always *introspectively aware* of our conscious states. Along the same lines, Seager also says that he surely has many conscious states made up of consciousness of worldly things: objects, properties, events, etc. which are, in consciousness, presented to the mind. It seems altogether natural to call these states, states of consciousness. I agree but this only means that one is more frequently having world-directed conscious states than introspective states, which is perfectly consistent with HOT theory since it is only when one introspects that one has *conscious* HOTs. Indeed, I use examples like Seager’s hockey case against Kriegel’s insistence that each outer-directed conscious state is accompanied by an inner-directed peripheral conscious awareness.

Nonetheless, perhaps Seager is correct that we shouldn’t treat TP as *entirely* independent evidence for HOT theory. But I still think that it is one intuitive way of understanding state consciousness which can at least be thought of as a good reason to opt for some kind of higher-order theory of consciousness. TP is also typically endorsed by higher-order perception theorists and self-representationalists. Seager explains thatthe unwavering devotion to TP strikes him as somewhat odd or too swift. He thinks there is a more carefully worded and simpler claim which is an obvious alternative to TP. It is what he calls the *principle of phenomenal contribution*:

(PC) A conscious state of subject S is a state which contributes (some of its) content to what it is like to be S at that time.”

I honestly do not see why Seager thinks that PC is *more* intuitively obvious and simpler than TP, especially given the somewhat awkward phrasing “…which contributes (some of its) content to…” PC actually seems more complex to me in terms of its formulation. Further, PC does not really provide a natural way to answer the key question of state consciousness, namely, what makes a mental state a conscious mental state? Moreover, we must keep in mind that TP is not the only rationale for HOT theory. For example, there is the common “argument by elimination” strategy used by many writing in this area (and in my chapters two and three). So what’s the better alternative? Why? It is also unclear to me whether PC is meant as a definition of ‘conscious state’ or as some other substantial characterization. The ‘aware of’ language in TP is intentionally meant to be neutral as to whether or not this awareness itself is conscious. As is well understood, this is one way that HOT theory avoids the charge of circularity or regress. TP lends itself more clearly to an account of state consciousness. PC, however, may be more susceptible to circularity and regress charges since we find ‘conscious state’ defined or explained in terms of ‘what it’s like.” There is nothing wrong with, say, holding that the Nagelian “what it’s like” sense of ‘conscious state’ is ultimately what we are aiming to explain. I agree with this and have said so myself while introducing different senses of ‘conscious.’ But PC seems more complex than that and is not then very helpful in terms of offering an account, especially a reductionist account, of state consciousness.

At one point, Seager says that it is notoriously controversial exactly *how or why* having a HOT would make a state conscious. But this strikes me as an entirely different problem, much closer to the so-called “hard problem” which I address at length in chapter four (section 4.4). I do use the central notion of conceptualization as part of my answer to the hard problem, but I also urge that the “real” hard problem may have more to do with concept acquisition, a topic I address at length in chapter seven.

Of course, some will no doubt always remain dissatisfied and will want to ask a further question: *why* does the higher-order application of concepts give rise to conscious experience? But this, I suggest, is not a legitimate question. We have already reached the rock-bottom brute fact about the way that conscious minds work. The Kantian idea that concepts make our experience of the world possible is a plausible view about the nature of conscious experience. I do not think that it makes sense to ask why this is so. Notice, however, that this solution is unlike reductionist accounts in physicalistic terms and so is immune anyway to Chalmers’s criticism about the plausibility of those theories. There is no comparable problem for HOT theory about how a specific brain activity produces conscious experience. Chalmers’s criticism that functional explanations are inadequate because one can always ask, “Why is the performance of these functions accompanied by experience?” (1995, p. 203) is equally beside the point. HOT theory is not a functional explanation that merely addresses the easy problems of consciousness. In any case, HOT theory contends that a reductionist theory of consciousness can be provided in mentalistic terms in a way that can solve the hard problem.

**Conceptualism and HOT theory** (Seager)

In order to respond to Seager in this section, I reproduce below my definition of conceptualism (CON) and my criteria for concept possession (CONPOSS):

(CON) Whenever a subject S has a perceptual experience e, the content c (of e) is fully specifiable in terms of the concepts possessed by S.

(CONPOSS):Whenever a subject S has an empirical concept C that is applied to some object (or property or relation) in experience e, S must at minimum (a) be able (to some extent) to *discriminate* instances of C’s from non-C’s, (b) be able (to some extent) to *recognize* or *identify* instances of C by virtue of at least some central feature of the objects or properties in e, and (c) be able to include the concept C in at least some intentional states that S has.

In criticizing HOT theory and the WIV, Seager says that it seems ridiculous to suppose that a stream of insanely complex conceptually articulated thoughts are lurking behind all sorts of experience. However, part of the entire point of chapter six is that there really is no need for such “insanely complex” thoughts. Many of these thoughts need only have very coarse grained concepts and it must be emphasized that our focal consciousness is fairly limited in most cases. This is also relevant as a response to Seager’s worries about “heavy demands on cognitive load” and “re-duplication of content” on HOT theory. So, first, these demands need not be as great as Seager portrays and we must keep in mind that HOTs are typically unconscious so it is not as if our conscious minds would overload. I also present several other replies to this sort of objection in section 6.3.4 (“HOT Theory and the Complexity Objection”). Second, in my book (pp. 46-47), I offer several reasons for why HOTs might have been useful from an evolutionary point of view. For example, the early presence of actual unconscious HOTs can be understood as key stepping stones to introspective consciousness. Such an evolutionary history is presumably mirrored in the layered development of the cortex. Moreover, as Rolls (2004) points out, actual HOTs allow for the correction of plans that result from first-order processing. HOTs enable correction of errors made in first-order linguistic or in nonlinguistic processing.

Seager also presents what he takes to be a difficulty for conceptualism, namely, cases of gradual change blindness: How are we to describe subjects who notice no change while perceiving and introspecting a video with gradual color changes (e.g. from red to purple)? According to HOT theories, during introspection we are aware of the thought about the visual experience in addition to being aware of the visual experience. So the visual input is changing and if the visual system is functioning normally, then at the end of the video subjects are experiencing purple (surprise can be elicited if the scene at the start of the video is immediately then shown). Introspectively, subjects report no change in experience. How can this be? What concepts are applied to the visual experience over the course of the experiment?

But, as Seager knows, there are other similar cases, such as a pair of figures with slightly differing spot arrangements, where one doesn’t notice a difference between what would seem to be different conscious experiences (even perhaps when the two images are side by side). But not noticing a very slight, even if regular, *continuous* change seems just to be a special instance of this type of case and is no threat to HOT theory or conceptualism as far as I can see. Some differences between concepts are so minor that they will not result in differences in our conscious experience. Or one may just not *notice* or *remember* the difference between two conscious experiences. One can have two different conscious states over time and not realize *that* they are different, but I fail to see how this refutes CON. It hardly matters much just how long such a time gap is. I suppose it is probably true that we would be more likely to notice changes in one’s visual experiences over shorter periods of time but there is no guarantee of that.

Although each commentary raised some important questions and requests for clarification, I conclude that the WIV remains a defensible version of HOT theory and that there is reason to be optimistic about solving the consciousness paradox. In some cases, I think that my commentators might have overlooked certain other parts of the book which can at least be used to form a solid basis for a response. In other cases, they have forced me to go somewhat beyond the text or to clarify points raised in the book. One cannot address everything even in a book, but of course they are also limited by space consideration in their commentaries.

**References**:

Baars, B. and Gage, N. (2010). *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*. Second Edition. Oxford: Elselvier.

Chalmers, D. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies* 2: 200-219.

Gennaro, R. ed. (2004). *Higher-Order Theories of Consciousness: An Anthology*. Amsterdam and Philadelphia: John Benjamins.

Gennaro, R. J. (2012). *The Consciousness Paradox*. Cambridge, MA: MIT Press.

Kreigel, U. (2009). *Subjective Consciousness*. New York: Oxford University Press.

Levine, J. (2001). *Purple Haze*. New York: Oxford University Press.

Rosenthal, D.M. (2005). *Consciousness and Mind*. New York: Oxford University Press.

Rolls, E. (2004). A Higher Order Syntactic Thought (HOST) Theory of Consciousness. In Gennaro 2004.