Reply to Commentators:
Where "Verbal Disputes in the Theory of Consciousness" (Really) Goes Wrong

Joseph Gottlieb

Let me begin by thanking Aaron Henry and John Schwenkler for the opportunity to have my article "Verbal Disputes in the Theory of Consciousness" be a part of the *Brains* symposium series. I would also like to thank Jonathan Farrell, Josh Weisberg, and Assaf Weksler for their insightful commentary.

I want to do something that is perhaps unusual in a forum like this. Shortly after publication of my article, my credence in its core thesis—viz. that the dispute between proponents of Higher-Order (HO) and First-Order (FO) theories of consciousness is verbal—began to drop. While I still think the paper provides something of value to the literature, I now think my central conclusion is wrong. The dispute is *not* verbal. Or if it is, I have not shown as much.[1]

Now both **Farrell** and **Weisberg** concur with this assessment; they both agree that the dispute between HO and FO theorists is not verbal. (I am assuming here that they do not simply think that *my* argument for this claim fails.) By contrast, on the whole **Weksler** is sympathetic with my conclusion. However, I think all three commentators miss the crux of the matter. While **Farrell** and **Weisberg** are right to say that I am wrong, they both miss the fundamental reason why I am wrong. So while I will have plenty to say about each commentary, my focus in what follows will be drawing out what I take to be my most fundamental mistake. Although this diagnosis amounts to a *mea culpa*, I nonetheless think it will also help further illuminate both the scope of HO and FO theories, and where the disagreement between the respective camps lie.

But before we get to that, I want to begin with some general remarks about the use of 'what it is like'-talk (WIL-talk) and its role both in discussions of consciousness and in particular how HO theorists tend to use it. **Weisberg** contends that 'what it is like'-talk is technical. I doubt this. While **Weisberg** is right to say that *The Philosophical Review*—the journal where Nagel introduced the phrase—is a 'technical journal', this by no means indicates that the phrase itself is a technical one. And yes, I concur that it can at times be difficult to get undergraduates precisely what the phrase means, but here too this does not mean a whole lot, as there is much of philosophy that is difficult for undergraduates to understand, and it's hardly the case that this is so because all of it (or nearly all of it) is 'technical'. But this aside, as Farrell (2016) has pointed out, the issue isn't simply that Nagel didn't *explicitly signify* that he was employing technical terminology, the issue is also that the phrase itself doesn't involve technical terms, as evidenced by the fact that non-philosophers have used the phrase to talk about consciousness well before Nagel, Sprigge, and others ever brought into the philosophical discussion. As Farrell puts it, "the use of WIL-talk to talk about consciousness was not an innovation of philosophers." More can be said here, but I'll leave this issue aside for now.

[1] All page citations without a year are to "Verbal Disputes…"

Weisberg makes another intriguing point about how we should view WIL-talk in philosophical discussion of consciousness. He says the following:

> Nagel's phrase is a rough pointer at that *prima facie* problem. It provides a way to indicate to other researchers that, "hey, I'm talking about the problem under philosophical discussion, even if I later dismiss that problem." What folk disagree about, among other things, is what the real problem is. But the phrase provides a way of opening the debate. It says far less, at least at the outset, than any of the detailed analyses do. Those are possible ways to cash out the phrase, each contentious and motivated by various theoretical, phenomenological, and philosophical concerns. But the initial, more minimal pointing allows for us to get started with a meaningful debate. And that, to me, suggests that we're all speaking a common language but disagreeing on how to precisify and extend that language in theory-driven ways.

I think we should distinguish between how WIL-talk *should* be used and how theorists—in particular, HO theorists—*in fact* use it. On the former normative point, I now tend to agree with **Weisberg**: WIL-talk is a good pointer, a way of signifying the phenomena of interest, but not much more. Yet on the latter descriptive point, the problem here—as I try to illustrate in the paper—is that many HO theorists use WIL-talk as not just counting in favor of their theory, but (for some) actually entailing it! This is what I call, following Stoljar (2016), the 'emphatic argument' for HO theories. What this shows is that many HO theorist are *not* just doing "minimal pointing", as **Weisberg** suggests. Instead, they are using WIL-talk to uncover substantive claims about the nature of consciousness. Yet to make sense of that practice, we need to figure out precisely what HO theorist's mean when they use WIL-talk. And of course, that is exactly one of the things I try to do in my paper.

Now there is a further question about whether HO theorist in fact speak HO-English (as modeled on the Operator View of WIL-sentences), and whether FO theorists speak FO-English (as modeled on the Affective View), as I contend. Regarding the latter, I actually now think it is quite hard to get a uniform language to fit for all FO theorists, given the wide variety of theories that fall within this camp. In this regard, **Weksler** asks whether it is charitable to interpret Block in particular as speaking FO-English. Here **Weksler** raises Block's analysis of subjects with visual extinction, and his claim that we can still make sense of such a subject having conscious experiences that are 'for him' in the relevant sense. **Weskler** points out that "Block's account makes no appeal to feelings, or to affective relations between the subject and the face representation (or the experience)", and therefore, that it is unclear whether it is charitable to claim that Block is speaking FO-English. **Weksler** is right that Block makes no mention of these things, as I concede on page 333. Yet this is not the point of the example. The point of the example is to illustrate how Block thinks of what I take to be a key component of WIL-talk, viz. the 'for the subject' phrase. As I argue, speakers of FO-English will, in accordance with the Affective View, interpret the 'for the subject' phrase as simply indicating that one of the relata of the affective relation is a subject. And that's what we see in Block's example. But **Weksler** does have a point here, which he nicely draws out with his suggestion of 'IA-English': the appeal to affective relations, as part of FO-English, isn't obviously doing much work in my argument at all (cf. pg. 329, fn. 18). What's doing the work, again, is how

the 'for the subject' phrase is understood: HO theorists, in whatever language they speak, interpret it as implying that subject is aware of the mental state, while FO theorists, in whatever language they speak, do not.

This brings me to a point made by **Farrell.** In a particularly incisive commentary, **Farrell** notes, amongst other things, that what I claim is the disputed sentence is not in fact the disputed sentence. I claim that the disputed sentence—the sentence that results from combining Transitivity with the Nagelian Conception (TNC) is this:

NT A mental state M is like something for its subject S only if its S is in some way aware of M (pg. 325)

But **Farrell** points out that the actual result is not NT but:

NT* There is something it is like for S to be in M only if S is aware of M

Farrell is right here. So let's grant what's due. But then **Farrell** adds that the argument won't work if we focus on NT* instead of NT,[2] since I get Stoljar's Operator view wrong. And here I am less sure: the argument may still go through.

To see this, consider first the two disambiguations of the Operator View provided by **Farrell:**

O1 There is something it is like for S to be in M iff S's being in M seems some way to S.

O2 There is something it is like for S to be in M iff M seems some way to S.

Now **Farrell** points out that whereas I adopt O2, what is licensed by Stoljar's Operator View is actually O1. Fine; let's run with that. What matters now is how exactly we should understand Transitivity, and as a result any disputed sentence we form from it via TNC. This takes some explaining.

I state Transitivity as follows: a mental state is conscious only if one is in some way aware of that mental state. For sake of clarity, let's call this *state-relational transitivity* (cf. Gottlieb 2016). Yet as I allude to at the beginning of my paper (pg. 320, fn. 1), this is not the only way to understand Transitivity. At its core, Transitivity expresses a constitutive connection between *some* form of inner awareness and being (phenomenally) conscious. So state-relational transitivity is one way of expressing this idea, since being aware of M is a form of inner awareness. Yet another way is this: conscious mental states are mental states we are aware of *ourselves* as being in. Again for the sake of clarity, let's call this *self-relational transitivity*.[3]

Now in the paper, I assume state-relational transitivity as both a simplifying assumption, and because this is how Transitivity is perhaps most frequently understood, and because many HO

[2] Here I am setting aside the issue of whether NT* entails NT, something **Farrell** seems to deny.

[3] This is not actually my preferred way to state self-relational transitivity, but since it is closer to how most HO theorists do it, I follow suit here. For discussion, see Gottlieb (forthcoming).

theorists in fact hold it. However, state-relational transitivity is for many just a quick-and-dirty way of getting at Transitivity; at least close enough to mark their disagreement with FO theorists, since FO theorists reject it. But in fact what many HO theorists ultimately want—and here I count Berger (2014), Brown (2015), and Rosenthal (2011) too—is self-relational transitivity.

Here's why this matters. First, all FO theorists reject self-relational transitivity, not just state-relational transitivity. This is because its implementation will plausibly require some form of HO representation, given that FO representations are limited to environmental and bodily features. Second—and critically in light of **Farrell's** comments—I think O1 is apt here. Combining TNC and self-relational transitivity, we get something like this:

NT** There is something it is like for S to be in M only if S is aware of herself as being in M.

Now, if we opt for NT** as the new disputed sentence, and add in O1 (not O2) as part of HO-English, we can get a new undisputed sentence, viz.:

U2* S's being in M in will seem some way to S only if S is aware of herself as being in M.

And here U2* should be undisputedly true: to speak of S's *being in M* as seeming some way plausibly entails that S is aware of herself as being in M. To be aware of being in M would, I take it, require being aware of that which is in M, i.e. S. And this is precisely what self-relational transitivity says.[4] There is more to say about **Farrell's** rich commentary, but for sake of getting to my main point, I will press ahead.

To get at where I think I really go wrong, I'll begin with the FO side of things, and unlike my paper where I take Ned Block as the core representative of FO theory, here I will focus on Michael Tye's variant of FO theory. The reason for this will become clear as we proceed.

Tye (2014a, 2014b, 2014c) is a *property-complex theorist*. I'll understand this as the conjunction of two claims. First: phenomenal or qualitative character is *identical* to a property-complex, which may or may not be instantiated in the subject's environment.[5] Second: to have an *experience* with a property-complex as its phenomenal character, the subject must have a mental state of the right sort that *represents* this property-complex.[6] So on this view, experiences 'have' phenomenal characters much in the same way a predicate has a semantic value (e.g. a property); not by instantiating it, but by representing (Tye 2014b: 85).

[4] Incidentally, all of this fits better with Rosenthal's characterization of how he understands WIL-talk (as quoted on pg. 330): "As many, myself included, use that phrase, there being something it's like for one to be in a state is simply its seeming subjectively that one is in that state."

[5] As I am using the term, 'phenomenal character' refers to what types an experience by what it is like for the subject to have it.

[6] There is a good sense in which Mark Johnston (2004) qualifies as a property-complex theorist, although he would deny this second condition, since he denies that experiences represent. I understand Dretske (1995) as being a property-complex theorist in Tye's sense.

As an illustration, suppose that Lena sees a green apple, and sees the apple as green. She has a veridical experience ('VE'). Lena then sees a red apple, but sees the apple as green. She has an illusory experience ('IE'). Later, it visually seems to Lena as if there is a green apple on her desk, but there is no apple at all. She is having a hallucinatory experience ('HE'). These experiences, let's suppose, are phenomenally identical. So, given the 'common kind claim', the Property-Complex theory says that they all represent the same property-complex, e.g.:

$\lambda x(x \text{ is green} \wedge x \text{ is round} \wedge x \text{ is an apple})$

Property-complexes—here, the property of being an x such that x is green and x is round and x is an apple—are conjunctive or structural properties that have simple properties (like being green and being round) as parts (Tye 2014c: 306). What distinguishes Lena's VE and HE is whether this property-complex is instantiated. In a HE, the property-complex is wholly uninstantiated.

Now here is the key takeaway for present purposes: as alluded to above, not everything that represents a property-complex counts as an experience.[7] Property-complexes, being constituted by ordinary physical properties, can be represented unconsciously. This is why the mental state needs to be of 'the right sort' to constitute an experience. For Tye, the mental state needs to be a representation that plays the right functional role. It has to be suitably 'poised' to impact beliefs and desires (Tye 2014b: 86; cf. Tye 2000). What makes Tye a *FO* theorist is that this existence condition on conscious experience—that of a mental state's being suitably poised—is supposed to be cashed out without adverting to any form of HO awareness or representation.

But here's the thing: given that not everything that represents a property-complex counts as an experience, and given that phenomenal character is *identical* to property-complexes, it follows that there is a sense in which unconscious perceptual states can have phenomenal character too. This might sound odd, but it's not when we remember what phenomenal character is for Tye: ordinary physical properties that, if they are instantiated anywhere, are instantiated in the subject's environment. You might have qualms with that claim, but my point is that *if* you endorse it, the upshot is not terribly odd, given the existence of unconscious perception. (The latter point being something HO theorists almost universally endorse.)

The reason I belabor all these points is that I think they help underscore where I go wrong in "Verbal Disputes…". A central part of my argument there turns on what I claim are two uncontested sentences (pg. 334):

U1: 'A mental state will be such that its subject feels some way in virtue of being in it only if its subject is in some way aware of it.'

U2: 'A mental state seems some way to its subject only if its subject is in some way aware of that mental state.'

[7] Here I am using 'experience' where experiences are by definition conscious. This is counter to some uses in the HO-theoretic landscape, but in the present context it should be harmless.

Putting aside the point about U2* made above in response to **Farrell**, the original idea was that the FO theorist would hold that U1 is equivalent to Transitivity, while the HO theorist would hold that that U2 is equivalent to Transitivity, and critically that HO and FO theorists agree that U1 is *false* and that HO and FO theorists agree that U2 is *true*. My efforts in this regard focused on defending the claim that HO theorists think U1 is false. It is here where I think the crux of my mistake lies. For I now think that, on HO theory, U1 is *true*—or at least, there is no reason, within the HO platform itself, that U1 *must* be false. This means that even if we take (for instance) **Farrell** to be wrong about the actual disputed sentence, or anything else, the present point is enough to block my conclusion. If the HO theorist merely *can* say that U1 is true, that should be enough to show my argument fails. In that sense, this point is fundamental. And I think the aforementioned discussion of Tye's FO theory helps illuminate why this is indeed the case.

Consider first why one might think the HO theorist is committed to saying that U1 is false. Here I discuss Rosenthal's take on Block's stance on visual distinction. Rosenthal denies that such subjects have conscious experiences in the Nagelian sense, because the subjects lack the requisite HO awareness. But curiously (as I note on pg. 335), he allows that while the relevant states of such subjects lack 'thick phenomenality', they do have 'thin phenomenality':

> One . . . consists in the subjective occurrence of mental qualities, while the other kind consists just in the occurrence of qualitative character without there also being anything it's like for one to have that qualitative character. Let's call the first kind *thick phenomenality* and the second *thin phenomenality*. Thick phenomenality is just thin phenomenality together with there being something it's like for one to have that thin phenomenality. (2002a: 657)

Rosenthal (2010) also tells us, in line with his *quality space theory* (QST), that unconscious perceptual states—again, 'unconscious' in the sense of not being HOT-targeted—have the very same 'mental qualities' that occur in conscious perception. Similarly, I noted (pg. 338) that Kriegel allows that (what are for him) unconscious perceptual states (as in cases of absent-minded perception) have *qualia*.

The lesson I drew from all of these consideration is that, given the ordinary meanings we associate with terms like 'phenomenality' and 'qualia', the HO theorist must think that U1 is false. After all, if first-order states have qualia with 'thin phenomenality', what else could this mean but that the subject *feels* some way (in a broad sense) in virtue of being in these FO states, even if (per the Operator View) the state isn't conscious because the subject isn't aware of it'?

But this lesson is wrong, and Tye's theory helps illustrate why. My view now is that we can think of a mental state's having Rosenthal's 'thin phenomenality' and Kriegel's 'qualia' as akin to the unconscious representation of property-complexes on Tye's theory. Remember: for Tye, phenomenal character, or 'phenomenal properties', are identical to property-complexes. And *when* a mental state *with the right functional profile* represents them, they will individuate the resultant experience phenomenally, so that if what it is like for a subject S to be in an experience E1 differs from what it is like for S to be in an experience E2, E1 and E2 will have distinct phenomenal characters and so represent distinct property-complexes. But it does not follow

from this that when a mental state of the *wrong sort* (e.g. one that that does *not* have the right functional profile) represents these properties that the subject feels anyway in virtue of being in that state. Likewise, for the HO theorist, when a first-order state has (what Kriegel calls) qualia or (what Rosenthal calls) mental qualities, a subject need not feel anyway in virtue of being in it. They *can* feel some way, but only if the first-order state is itself represented in the right way.

So while I think it was a mistake for Rosenthal to even concede to Block that non-HOT-targeted FO states have 'thin phenomenality' since this is potentially misleading, it's really no more misleading than Tye allowing that unconscious perceptual states can, by his lights, represent phenomenal properties—which, upon reflection, is not that misleading at all. Now for someone like Rosenthal, mental qualities—here the counterpart to Tye's phenomenal properties—*are* instantiated by our first-order states. Yet this makes no difference, and the details show why. For example, consider Rosenthal's (2010) QST. Here mental qualities (like 'mental redness' or red*) are extrapolated from a subject's perceptual discriminatory capacities, with the quality space of mental qualities being homomorphic to the external perceptible properties a subject can discriminate, which are themselves charted into their own quality space. On QST, for a subject to engage in the behaviors captured by its external quality spaces (e.g. how red is discriminated by our visual systems to be less similar to green than orange), it must be able to be in (first-order) perceptual states that resemble and differ from one another in ways that mirror how the perceptible properties resemble and differ from one another (Rosenthal, ibid: 377). And the way this is done, on QST, is to say that our perceptual states instantiate mental qualities, like red*. But, according to Rosenthal, a perceptual state can have red* absent it being conscious.

Of course, QST is an interesting theory of 'qualitative character' and first-order content in its own right, one that is very different from tracking theories. But the point here is how QST can intersect with the HO theorists take on consciousness and the truth-value of U1. We can think of the situation like this. The HO representation—for Rosenthal, a HOT—determines *whether* a first-order mental state is conscious. This provides an existence condition. But the *location* of that first-order state in the relevant quality space provides an identity condition.[8] In this way, we have a situation analogous to Tye's FO theory. Which property-complex a mental state represents determines its identity conditions, but whether a given mental state is conscious at all is a matter of whether it is poised. None of this requires saying that U1 is false. First-order states with mental qualities (or 'qualia') have thin phenomenal *only* in the sense that, if they were conscious, those qualities would determine the respect in which the state was conscious.

[8] This is complicated a bit by the fact that on Rosenthal's particular HO theory, a HOT, in the absence of a FO state instantiating these mental qualities, is still *sufficient* for a subject to undergo a conscious experience (e.g. 2009: 249). I think this causes its own set of problems—*pace* Block (2011), it's not internal incoherence--but again the point here concerns what the HO theorist *must* say, not simply what a particular HO theorist happens to say.

References

- Berger, J. (2014). Consciousness is Not a Property of States: A Reply to Wilberg. *Philosophical Psychology*, 27 (6): 829 - 842.
- Block, N. (2011). The higher-order approach to consciousness is defunct. *Analysis*, 71 (3): 419-431.
- Brown, R. (2015). The HOROR Theory of Consciousness. *Philosophical Studies*, 172 (7), 1783-1794.
- Dretske, Fred. (1995). *Naturalizing the Mind*. MIT Press.
- Farrell, J. (2016). 'What It Is Like' Talk Is Not Technical Talk. *Journal of Consciousness Studies*, 23(9– 10), 50– 65.
- Gottlieb, J. (2016). Transitivity and Transparency. *Analytic Philosophy* 57 (4):353-379.
- Gottlieb, J. On Ambitious Higher-Order Theories of Consciousness. Forthcoming in *Philosophical Psychology*.
- Johnston, M. (2004). The Obscure Object of Hallucination. *Philosophical Studies* 120 (1-3): 113-83.
- Rosenthal, D. (2009). Higher order theories of consciousness. In B. McLaughlin and A. Beckermann (eds.), *Oxford Handbook of the Philosophy of Mind*, 239–52. Oxford: Clarendon Press.
- Rosenthal, D. (2010). How to Think about Mental Qualities. *Philosophical Issues* 20 (1): 368-393.
- Rosenthal, D. (2011). Exaggerated Reports: Reply to Block. *Analysis* 71 (3):431-437.
- Stoljar, D. (2016). The Semantics of 'What it Is Like'- Sentences and The Nature of Consciousness. *Mind*, 125(500), 1161– 1198. https://doi.org/10.1093/mind/fzv179
- Tye, M. (2000). *Consciousness, Color, and Content*. MIT Press.
- Tye, M. (2014a). Transparency, qualia realism and representationalism. *Philosophical Studies* 170: 39–57.
- Tye, M. (2014b). Speaks on Strong Property Representationalism. *Philosophical Studies* 170 (1): 85-86.
- Tye, M. (2014c). What is the Content of a Hallucinatory Experience? In Berit Brogaard (ed.), *Does Perception have Content*? Oxford University Press.