

Self-Representationalism and the Explanatory Gap

Uriah Kriegel Arizona

Forthcoming in J. Liu and J. Perry, *Consciousness and the Self: New Essays* (CUP)

Introduction/Abstract

According to the self-representational theory of consciousness – *self-representationalism* for short – a mental state is phenomenally conscious when, and only when, it represents itself in the right way. In this paper, I consider how self-representationalism might address the alleged explanatory gap between phenomenal consciousness and physical properties. I open with a presentation of self-representationalism and the case for it (§1). I then present what I take to be the most promising self-representational approach to the explanatory gap (§2). That approach is threatened, however, by an objection to self-representationalism, due to Levine, which I call the *just more representation* objection (§3). I close with a discussion of how the self-representationalist might approach the objection (§4).

1. Self-Representationalism

In *Subjective Consciousness: A Self-Representational Theory* (henceforth, SC), I develop and defend a specific version of self-representationalism. Self-representationalism can be formulated as follows:

(SR) Necessarily, for any mental state M, M is phenomenally conscious iff M represents itself in the right way.

Different versions of SR can be obtained by unpacking ‘in the right way’ in different ways. My own version construes ‘the right way’ as ‘non-derivatively, specifically, and essentially.’¹

What motivates SR, at least to me, is a certain conception of the structure of phenomenal character. As I look at the blue sky, I undergo a conscious experience, and there is a bluish way it is like for me to undergo that experience. This ‘bluish way it is like for me’ is the experience’s phenomenal character. As Levine (2001) notes, there is a conceptual distinction to be drawn between two components of this ‘bluish way it is like for me’: (i) the bluish component, and (ii) the for-me component. I call the former *qualitative character* and the latter *subjective character* (Kriegel 2005, 2009). To a first approximation, the experience’s bluish qualitative character is what makes it the experience it is, but its for-me-ness is what makes it an experience at all. A better, if initially less clear, approximation is this: my experience is the experience it is because it is bluish-for-me, and is an experience at all because it is somehow-for-me (or qualitatively-for-me).² Thus qualitative character is what varies among conscious experiences, while subjective character is what is common to them.

Many philosophers have assumed that the core of the problem of consciousness is qualitative character, but an interesting result of the above conception of the structure of phenomenal character is that it is actually subjective character that is more central (Levine 2001; Kriegel 2009). Although it is important to understand what accounts for the differences among conscious experiences, it is more central to the problem of consciousness to understand what distinguishes conscious experiences from non-conscious mental states. According to Levine and me, the deeply mystifying feature of phenomenal consciousness is that when I have a conscious experience, the experience does not occur only *in me*, but also *for me*. There is some sort of direct presence, a subjective significance, of the experience to the subject. This is of course not uncontroversial, but I will not argue for it here. What I want to focus on is the inference *from* this conception of the structure of phenomenal character *to* self-representationalism.

Self-representationalism is essentially an account of subjective character: it claims that a mental state has subjective character just in case, and because, it represents itself in the right way.³ The argument for this can be thought of as proceeding in three stages. Here I will only sketch the argument; for details, see Ch.4 of SC.

First, for a conscious experience to be not only *in me*, but also *for me*, I would have to be *aware* of it. The awareness in question need not be particularly focused or

attentive. But there must be some minimal awareness of a mental state if the state is to be described as exhibiting ‘for-me-ness.’ So we can reason as follows:

- 1) Necessarily, for any mental state M and subject S , such that S is in M , M is phenomenally conscious iff M has subjective character (is *for* S).
- 2) Necessarily, for any mental state M and subject S , such that S is in M , M has subjective character (is *for* S) iff S is aware of M in the right way. Therefore,
- 3) Necessarily, for any mental state M and subject S , such that S is in M , M is phenomenally conscious iff S is aware of M in the right way. (1,2)

This is the first stage of the argument. It takes us from phenomenal character to awareness.

The second stage employs crucially a pair of relatively uncontroversial lemmas, to the effects that (a) being aware of something is a matter of representing it and (b) representing something is a matter of being in mental state that represents it:

- 4) Necessarily, for any entity X and subject S , S is aware of X in the right way iff S represents X in the right way. (Lemma)
- 5) Necessarily, for any entity X and subject S , S represents X in the right way iff there is a mental state M^* , such that (i) S is in M^* and (ii) M^* represents X in the right way. (Lemma) Therefore,
- 6) Necessarily, for any mental state M and subject S , such that S is in M , M is phenomenally conscious iff there is a mental state M^* , such that (i) S is in M^* and (ii) M^* represents M in the right way. (3,4,5)

This is the second stage, which takes us from awareness to representation.

The third stage takes us from representation to self-representation. It does so by first setting up a dilemma – are the conscious state and its representation numerically distinct or numerically identical? – and then offering considerations in favor of the latter horn. Thus:

- 7) For any mental states M and M^* , either $M=M^*$ or $M \neq M^*$. (Excluded middle)

- 8) Necessarily, for no mental state M and subject S , such that S is in M , M is phenomenally conscious iff there is a mental state M^* , such that (i) S is in M^* , (ii) M^* represents M in the right way, and (iii) $M \neq M^*$. Therefore,
- 9) Necessarily, for any mental state M and subject S , such that S is in M , M is phenomenally conscious iff there is a mental state M^* , such that (i) S is in M^* , (ii) M^* represents M in the right way, and (iii) $M = M^*$. (6,7,8)

The conclusion, Proposition 9, is equivalent to SR. The negation of Premise 8, while not equivalent to the so-called higher-order theory of consciousness, is a commitment of that theory.⁴ What is needed to complete the argument are considerations that support Premise 8.⁵

In Ch.4 of SC I offer a battery of considerations against higher-order theory, hence in favor of Premise 8. I cannot go through all of them, and anyway many are familiar from the literature. But the consideration which is least familiar, yet which personally has been most persuasive to me, can be put thus: for-me-ness is internal to the phenomenology of conscious experience – it is a component of phenomenal character, after all – and this cannot be accommodated by higher-order theory, only by self-representationalism. There are two parts to this.

The first part is the claim that for-me-ness is internal to the phenomenology – that it is itself a conscious phenomenon. This seems to me self-evident. The very reason to believe in the for-me-ness of experience is fundamentally phenomenological: it is derived not from experimental research, nor from conceptual analysis, nor from any other sources, but rather from a certain first-person impression. This suggests that for-me-ness is phenomenologically manifest.

The second part is the claim that only SR can accommodate the phenomenological manifest-ness of for-me-ness. The reasoning here is this. If the for-me-ness of a conscious mental state M is itself conscious, then the mental state that represents M , i.e. M^* , must be a conscious mental state. If M^* is numerically identical to M , as per SR, it is predictable that M^* be conscious, since M is conscious and $M^* = M$. But if M and M^* are numerically distinct, as per higher-order theory, M^* 's being conscious is not only inexplicable, but in fact leads straightforwardly to an infinite regress: M^* 's being conscious requires the postulation of a third-order M^{**} , and so on.

This argument is developed in much greater detail in SC, Ch.4. It amounts to splitting Premise 8 in the above argument into two parts:

8a) Necessarily, for any mental state M and subject S , such that S is in M , M is phenomenally conscious iff there is a mental state M^* , such that (i) S is in M^* , (ii) M^* represents M in the right way, and (iii) M^* is conscious.

(Phenomenological observation)

8b) Necessarily, for no mental state M and subject S , such that S is in M , M is phenomenally conscious iff there is a mental state M^* , such that (i) S is in M^* , (ii) M^* represents M in the right way, (iii) M^* is conscious, and (iv) $M \neq M^*$.

(On pain of infinite regress)

Together, 8a and 8b entail 8. With 8 in place, and given our starting point in 1 and 2 and the relatively uncontroversial lemmas in 4 and 5, we obtain 9. Call this the *master argument for self-representationalism*.

An immediate challenge self-representationalism faces is making sense of the notion of self-representation. This has two central parts: explicating what is involved in self-representation, and explaining how it can occur in the natural world. These two parts interact, of course: we require an explication of self-representation that primes it for naturalization.

The natural approach to this challenge is to consult naturalistic theories of mental representation, and suggest that whatever natural relation they identify as underlying mental representation – informational, teleological, or what have you – is the kind of relation that mental states can sometimes bear to themselves. Call the kind of self-representation this would be *crude self-representation*. The problem with crude self-representation is that when we actually consult such theories as Dretske's (1981, 1988), Millikan's (1984), and Fodor's (1990), we find that they identify natural relations that are anti-reflexive: nothing can bear them to itself.⁶ The core of the problem is that these relations typically involve causal processes and relations, and those are often anti-reflexive.⁷

In Ch.6 of SC, I offer an account of self-representation intended to make it consistent with naturalistic accounts of mental representation. To a first (and rough) approximation, the story is this. First, there is a distinction to be drawn between direct and indirect representation. For example, I might represent a house by representing its

façade. In this case, I represent the façade directly and the house indirectly. Secondly, for M to self-represent is for M to have two parts, M1 and M2, such that M2 represents both (i) M1 *directly* and (ii) M *indirectly*. (M2 represents M *by* representing M1.) Thirdly, a naturalistic theory of mental representation can have two parts: accounting for *direct* representation in terms of the natural relation identified by the best naturalistic theory, and accounting for *indirect* representation in terms of the combination of the relevant natural relation and some *representation-transmission* relation R that holds between what is represented directly and what is represented indirectly. For example, I might represent a façade in virtue of bearing the right teleo-informational relation to it, and represent the house of which it is a façade in virtue of (i) bearing that relation to the façade and (ii) the façade bearing R to the house. The result casts a self-representing mental state as a mental state with two parts, such that one part bears the right natural relation to the other part and this second part bears the representation-transmission relation to the whole of which they are both parts. More precisely: M represents itself iff there are states M1 and M2, such that (i) M1 is a proper part of M, (ii) M2 is a proper part of M, (iii) M1 bears the right natural relation to M2, and (iv) M2 bears R to M. To distinguish it from crude self-representation, call this *subtle self-representation*.

Once this relatively specific structure has been identified, we may seek brain structures and processes that implement it: neurophysiological structures we have good reasons to describe as involving two parts one of which bears the right natural relation to the other and the other bearing the right representation-transmission relation to the whole of which they are both parts. Although an endeavor of this sort is extremely speculative at present, I indulge in it in Ch.7 of SC. With the aid of several empirical claims, the speculative hypothesis I arrive at is this: phenomenally conscious states are brain activations neurally synchronized with activation in the dorsolateral prefrontal cortex (dlPFC). Thus there is an intimate connection between the property of being phenomenally conscious and the property of being neurally synchronized with dlPFC activation. The latter is the neural *correlate* of consciousness, but more excitingly, it might also turn out to be the neural *reducer* of consciousness.

2. Explanatory Gaps and Explanatory Sequences

Any theory of consciousness that ends up suggesting a reduction of phenomenal consciousness to something like neural synchronization with dlPFC activation runs up against Levine's (1983) explanatory gap: how could something as majestic as phenomenal consciousness be *nothing but* those brute and blind processes unfolding inside the darkness of the skull? Something about phenomenally conscious states seems left unexplained by anything familiar from neuroscience. Levine would say that it is the precisely the *subjective character* of conscious states, their for-me-ness, that evades explanation by neuroscientific means. In this section, I sketch an approach to the explanatory gap that it would be natural for the above self-representational theory to take.

Consider a sorites series that takes you from a yellow circle to a red circle. As you are force-marched through the series, any pair of adjacent circles are visually indistinguishable to you, yet the first and last circles are very much distinguishable. In other words, when the steps in a sequence of this sort are small enough, the relation of visual indiscriminability will hold between the two sides of each step but not between the start and end points of the sequence. The relation of explainability – or perhaps just reductive explainability (as distinguished, say, from causal explainability) – might exhibit similar behavior, though perhaps for different reasons (not because it is vague). A series of claims can be envisaged, such that every claim $n+1$ is a reductive explanation of claim n , but there is no reductive-explanatory relation between the first and last claims.

Indeed, the discussion in the previous section can be modeled as such a sequence of proposed explanatory steps. Consider:

Step 1: explain phenomenal consciousness in terms of subjective character

Step 2: explain subjective character in terms of a certain type of awareness

Step 3: explain this type of awareness in terms of representation

Step 4: explain the relevant kind of representation in terms of self-representation

Step 5: explain the naturalistic possibility of self-representation in terms of subtle self-representation

Step 6: explain subtle self-representation in terms of synchronization with dlPFC activation

Each step seems to involve an explanatory move that does not strike us immediately as outlandish: the gap between explanandum and explanans does not seem obviously unbridgeable. So the relevant explanation relation *does* hold within each step.⁸ Yet the explanatory gap looms ominously when we consider, in one intellectual act as it were, phenomenal consciousness and synchronization with dlPFC activity.

The explanatory gap arises, on this line of thought, because of an unwarranted expectation that a complex sequence of explanations could be appreciated in one intellectual act. When we look at water and H₂O, a single intellectual act would leave us equally puzzled about how it is that the right interlocking of an oxygen atom and two hydrogen atoms could make something *wet*. It is a general feature of the relationship between the manifest image and the scientific image that structures and processes from the latter do not illuminate ones from the former in such a direct way. The illumination is not provided in a single encompassing act of apprehension. Rather, it is appreciated indirectly through patient consideration of a sequence of local explanations too long or complex to grasp at once (see Pollock and Cruz 1999).⁹

Note as well that we are familiar, in everyday life, with two kinds of understanding. Sometimes we understand something in a purely intellectual, somewhat cold-blooded manner. On other, relatively rarer occasions, we understand something in a more visceral way, where we feel like we can *see* the truth (or plausibility) of some notion. Indeed, it sometimes happens that we understand something first in the cold-blooded manner and suddenly in the more visceral way. The latter experience of understanding is much more phenomenologically impressive, and is also more *satisfying* and more confidence-imbuing. But it is also rarer, and there is no reason to suppose that it is always available: there may be areas where the human cognitive system does not have the resources that would allow us to undergo the experience of this more visceral variety of understanding. We must there rest content with the phenomenologically lamer variety of understanding – and remember that it is still a variety of *understanding*.¹⁰

Taking these considerations into account, one may suggest that the explanatory gap is an illusion grounded in the attempt to take in a complex sequence of explanations in a single intellectual act.¹¹ The sequence may simply be too complex for us to do so successfully, in a way that summons the visceral phenomenology of understanding. But the other variety of understanding, the more ‘cold-blooded’

variety, can still be enjoyed when we consider patiently the sequence of explanatory steps presented above, perhaps precisely because we do experience the visceral variety whenever we consider any single step in the sequence.

On this interpretation of the line of thought under consideration, there is no genuine explanatory gap between phenomenal consciousness and synchronization with dlPFC activity. There *is* in fact a reductive explanation of the former in terms of the latter. It is just that this reductive explanation is not such as to elicit in us a visceral phenomenology of understanding, only an intellectual one.

The analogy with the sorites series points to a different interpretation, however. In that series, the red circle *really is* dissimilar to the yellow. The two are *not* visually indistinguishable. It is just that the continuity between them, which would otherwise be surprising, can be appreciated through the series. If we take the analogy at face value, phenomenal consciousness *really is* unexplainable in terms of synchronization with dlPFC activity. There is no reductive explanation to be had of the former in terms of the latter. What there is, however, is a sort of intellectual domestication of consciousness without reductive explanation of it. This admittedly elusive intellectual domestication may allow us to accept that consciousness reduces to neural processes even though it is not reductively explainable in terms of them.

To appreciate this approach, consider Chalmers' (2002) distinction between type-A and type-B materialism. There are several ways to draw the distinction, but in the present context the most natural one is the following. The dualist who argues from the explanatory gap appears to reason as follows: there is an explanatory gap between phenomenal properties and physical properties; if there is an explanatory gap between phenomenal and physical properties, then there is an ontological gap between them as well; therefore, there is an ontological gap between phenomenal and physical properties. The difference between type-A and type-B materialism is that the former rejects the first premise and the latter the second one. According to type-A materialism, any appearance of an explanatory gap between phenomenal and physical properties is strictly illusory. According to type-B materialism, the explanatory gap is genuine, but nothing ontological follows from this, and to suppose otherwise is to conflate the epistemic and the ontological.

To clarify the logical geography further, let us say that F is *epistemically reducible* to G iff there is no explanatory gap between F and G; and that if there is an explanatory gap between F and every other property, such that F is not epistemically

reducible to anything, then F is *epistemically primitive*. Correspondingly, let us say that F is *ontologically reducible* to G iff there is no ontological gap between F and G; and that if there is an ontological gap between F and every other property, such that F is not ontologically reducible to anything, then F is *ontologically primitive*. We may then formulate the dualist argument as follows:

- 1) Phenomenal properties are epistemically irreducible to physical properties;
- 2) If phenomenal properties are epistemically irreducible to physical properties, then phenomenal properties are ontologically irreducible to physical properties; therefore,
- 3) Phenomenal properties are ontologically irreducible to physical properties;
- 4) If phenomenal properties are ontologically irreducible to physical properties, then phenomenal properties are ontologically primitive; therefore,
- 5) Phenomenal properties are ontologically primitive.

Here type-A materialism denies Premise 1 and type-B materialism denies Premise 2, while Premise 4 is denied by certain types of so-called neutral monism (Chalmers' 'type-F monism'), namely, those that posit a third type of property, neither physical nor phenomenal, and attempt to reduce phenomenal ones to it.¹² Accepting all three premises, by contrast, leads one to dualism.¹³

Sociologically speaking, most materialists today are type-B. They concede an explanatory gap between consciousness and physical properties, but insist that consciousness is nonetheless ontologically reducible to physical properties. In general, this seems like a safe strategy, founded as it is on the widely recognized heuristic that we should avoid deriving ontological conclusions from exclusively epistemological premises. What is the dualist reason for upholding Premise 2, then?

The canonical presentation of the case against type-B materialism, and its denial of Premise 2 in the above argument, is due to Chalmers and Jackson (2001). The key notion in their reasoning is that of *epistemic transparency*. What this amounts to is not entirely clear, but it seems to suggest a connection between two facts, events, or phenomena that allows us to see, when we already know that one occurred, why the other one should occur as well. If subject S can see why *p* should be the case given that *q* is the case, then the connection between *p* and *q* is epistemically transparent to S. (For Chalmers and Jackson, epistemic transparency is achieved

through a priori entailment: if p entails q , then the connection between them is epistemically transparent. Thus a priori entailment is sufficient for epistemic transparency, but it is less clear whether it is supposed to be necessary for it as well. In any case, it a priori entailment does not seem to be *definitional* of epistemic transparency.)

According to Chalmers and Jackson, ontological reduction requires epistemically transparent connections between reduced and reducer. Even in the case of water and H₂O, although their identity is a posteriori, it is nonetheless epistemically transparent, in that a subject who knew all the non-identity truths about water and all the non-identity truths about H₂O would be in a position to establish the identity of water and H₂O. However, Chalmers and Jackson argue, once the connection between reduced and reducer is epistemically transparent, the reduction is not only ontological but also epistemic: one is in a position to *explain* the facts about the reduced in terms of the facts about the reducer.¹⁴ Chalmers and Jackson's reasoning may thus be summarized as follows: 2.1) We are justified in holding that one property ontologically reduces to another only if the connection between them is epistemically transparent; but 2.2) When the connection between two properties is epistemically transparent, we are also justified in holding that one *epistemically* reduces to the other; therefore, 2.3) We are justified in holding that one property ontologically reduces to another only if we are also justified in holding that one *epistemically* reduces to the other; so, 2) If phenomenal properties are epistemically irreducible to physical properties, then they are ontologically irreducible to them as well.¹⁵

Type-B materialists typically deny Premise 2.1 in this argument, insisting (e.g.) that the reduction of water to H₂O is opaque (see Block and Stalnaker 1999). However, one way to interpret the line of thought I have been examining is as resisting Chalmers and Jackson's reasoning rather by rejecting Premise 2.2 and allowing that epistemic transparency can arise even in the absence of epistemic reduction. On this view, a sequence of explanatory steps may be such that there is a genuine explanatory gap between the first and last items in the sequence, but the continuity that can be traced between them through consideration of the intermediary steps generates epistemic transparency in the entire sequence: the connection between the first and last items in the series is epistemically transparent to any subject who can follow each explanatory step in the series. Thus because every step in the series is an instance of reductive explanation, and we can follow the sequence, the identification

(or ontological reduction) of phenomenal consciousness to synchronization with dlPFC activity is epistemically transparent; but because the relation of reductive explainability is not transitive – whether for reasons of vagueness or some other reasons – phenomenal consciousness is not reductively explainable in terms of synchronization with dlPFC activity. This seems to be the correct analogy with the sorites series of circles.

On this interpretation, the explanatory gap between phenomenal consciousness and synchronization with dlPFC activity is real, in that we cannot understand why it is – indeed, how it could be – that phenomenal consciousness is nothing but synchronization with dlPFC activity. Phenomenal consciousness is genuinely epistemically irreducible to synchronization with dlPFC activity. Nonetheless, it does not follow that an ontological reduction of consciousness to synchronization with dlPFC activity must be epistemically opaque, leaving us with no insight into why it should be that consciousness is nothing but synchronization with dlPFC activity. On the contrary, by tracing a sequence of reductive explanations step by step, we can come to appreciate why it should be that consciousness is nothing but synchronization with dlPFC activity, say – even though contemplating the notion that it is in a single intellectual act produces in us only the phenomenology of incredulity.

We have here the reduction without reductive explanation – reduction in the face persistent explanatory gap – that is the hallmark of type-B materialism. But unlike typical type-B materialism, which embraces ontological reduction as brute and epistemically opaque, and is to that extent widely acknowledged to leave something to be desired, the present variety of type-B materialism offers epistemically transparent ontological reduction, and merely denies that epistemic transparency must bring in its train epistemic reduction. It takes its inspiration for this denial from the analogy with sorites series, and pointing out the sorites-like behavior of (reductive-)explanatory relations.

Let us distinguish, then, between type-B1 and type-B2 materialism. The former is the more common variety, embracing epistemically opaque ontological reduction. The latter is the variety suggested by the present interpretation of the line of thought explored in this section, the variety that exploits the sorites-like behavior of reductive explanation. What I am proposing here is in effect a self-representational variety of type-B2 materialism.¹⁶

To conclude. I started this section with an analogy between a sequence of (reductive) explanations leading from phenomenal consciousness to the neural process of synchronization with dlPFC activity, on the one hand, and a sequence of visually indistinguishable pairs of circles leading from a yellow circle to a red one, on the other. I then offered two interpretations of the analogy. On the first interpretation, the explanatory gap between phenomenal consciousness and synchronization with dlPFC activity is illusory: there is no explanatory gap between the two, but the appearance of such a gap arises from the unwarranted expectation that we undergo a visceral phenomenology of understanding upon contemplating the start and end points of the explanatory sequence; this is a form of type-A materialism. A tighter analogy is offered by the second interpretation: the explanatory gap is genuine – we really do not understand how phenomenal consciousness could be nothing but synchronization with dlPFC activity – but consciousness is nonetheless ontologically reducible to synchronization with dlPFC activity, and moreover ontologically reducible in an epistemically transparent manner, thanks to the sequence of reductive explanations connecting the two; this is a form of type-B2 materialism. I am happy with either interpretation, but find that the second one is vastly more satisfactory, insofar as it manages to respect rather than dismiss the force of the explanatory gap intuition.

3. Levine's 'Just More Representation' Objection

This self-representational approach to the explanatory gap can be resisted in two main ways. One is to deny the general claim that a series of reductive explanations can underlie an epistemically transparent physicalistic reduction of consciousness even in the absence of reductive explanation of consciousness in physical terms. The other is to claim that, however the general issue turns out, one of the six individual steps of reductive explanation I described in the previous section fails.

The most acute criticism of self-representationalism that takes this form is developed by Levine (2006), who argues that subjective character, the for-me-ness of experience, cannot be recovered by self-representation because the kind of awareness involved in subjective character cannot be accounted for in terms of the notion of representation at play in the relevant type of self-representation. This is to reject Step 3 in the explanatory sequence (explaining awareness in terms of representation).

For Levine, self-representation cannot account for subjective character, because just as something needs to bestow for-me-ness – a subjective significance – on any old representation, so something has to bestow that subjective significance on *self*-representations. As long as self-representing representations represent themselves in the same sense in which other-representing representations represent things other than themselves, it is not clear what would make the former ‘for the subject’ even though the latter are not (Levine 2006: 194):

Subjectivity, as I described it earlier, is that feature of a mental state by virtue of which it is of significance for the subject; not merely something happening within her, but ‘for her’. The self-representation thesis aims to explicate that sense of significance for the subject through the fact that the state is being represented. But now the question arises: how is that representation itself of significance for the subject?

The answer to this question cannot be, of course, that a self-representing representation is of significance to the subject because it represent itself to be self-representing. That would quickly lead to an infinite regress. The suspicion Levine raises is that there may not be a way to answer his question without invoking phenomenality.

Certainly what makes a representation ‘for the subject’ cannot be what it represents. It cannot be that when a representation represents *x*, it is not for the subject, but when it represents *y*, it is. And at a first pass, it might seem that this is precisely what self-representationalism claims. It claims that what makes some representations ‘for the subject’ is that what they represent is themselves. Yet the fact that a state represents itself rather than something other than itself does not dissolve the mystery involved in it representing what it does *to me*, i.e., in a subjectively significant way. Much more plausible is that representations endowed with subjective character are of a categorically different kind from other representations, and thus that what gives them their subjective character is an aspect not of *what* they represent, but of *how* they represent what they do – not their *object* of representation, but their *manner* of representation.

Note well, however: the heart of Levine’s objection is *not* that representations have subjective character in virtue of how they represent and not what they represent. For this is something that standard versions of self-representationalism can

accommodate. Compare ‘I am happy’ and ‘my mother’s nieceless brother’s only nephew is happy.’ Both statements represent me as happy, but there is a sense (perhaps elusive, perhaps not) that the former does so *essentially* whereas the latter *accidentally*. In specifying what makes a ‘suitable’ self-representation – the kind of self-representation that bestows subjective character – it is natural for the self-representationalist to insist that only the essential variety is relevant. Only essentially self-representing states are ‘for the subject’ and hence phenomenally conscious. Merely accidentally self-representing states are not. The point is that *what* is represented in both essential and accidental self-representation is the same, so what accounts for the fact that only the former involves subjective character must be the *manner* of representation (*how* what is represented is represented).

The heart of the objection is therefore not the what/how (object/manner) distinction. Rather, it must be the thought that there is no way to account for the right *manner* of representation in non-subjective, non-phenomenal terms. Even if a certain non-phenomenal specification of the right manner were extensionally adequate, such that no counter-example could be found to the thesis that necessarily, a mental state has subjective character iff it self-represents in that manner, we would still have on our hands an explanatory gap between subjective character and this non-phenomenal specification of the relevant manner. It would still be unclear how this specific kind of self-representation, understood in non-phenomenal terms, could give rise to the distinctive kind of awareness of one’s conscious experiences that is imbued with subjective character. Thus as long as representation is understood in non-phenomenal terms – certainly as long as it is understood in purely physical terms – it does not help to appeal specifically to *self*-representation.

The problem with self-representation, then, is that it is *just more representation*. As Levine (2006: 195) puts it:

Somehow, what we have in conscious states are representations that are intrinsically of subjective significance, ‘animated’ as it were, and I maintain that we really don’t understand how that is possible. It doesn’t seem to be a matter of more of the same—more representation of the same kind—but rather representation of a different kind altogether.

The awareness we have of our conscious experiences, in virtue of which they are ‘for us,’ involves a kind of acquaintance with those states that brute representations simply

do not seem to replicate, not even when they are representations of themselves. For a self-representation as for an other-representation, we can always ask: why is there something it is like *for me* to have this representation? Call this the *just more representation objection*.

We can, of course, countenance a phenomenal notion of representation that casts some representations as inherently subjective (see Loar 1987, 2002; Horgan and Tienson 2002). With this phenomenal notion of representation, one could certainly attempt to account for our awareness of our conscious experiences in terms of their manner of representing themselves: they represent themselves *phenomenally*. If a mental state represents itself phenomenally, then given that phenomenality involves for-me-ness, it would be clear what makes the state represent itself *for me* and not merely *in me*. But the result would be a kind of non-reductive self-representationalism, a kind of self-representationalism that does not even *attempt* to account for phenomenal consciousness in non-phenomenal terms.

Conversely, there may be *a* notion of awareness that can be accounted for in representational terms – essentially, a *non-phenomenal* notion of awareness.¹⁷ However, the kind of awareness we have of our conscious experiences, in virtue of which they are ‘for us’ in the relevant sense, is phenomenologically manifest, being as it is a component of phenomenal character. So it is a phenomenal notion of awareness. Thus even if we account in self-representational terms for a *non-phenomenal* awareness, that would not help us account for the for-me-ness of conscious experience, since the latter is constituted by a *phenomenal* awareness.

The upshot seems to be this: while we can account for the phenomenal kind of awareness in terms of a phenomenal kind of (self-)representation, and can account for non-phenomenal awareness in terms of a physicalistically unproblematic, non-phenomenal kind of (self-)representation, there appears to be no way to account for phenomenal awareness in terms of a non-phenomenal notion of representation. For appealing to a non-phenomenal notion of (self-)representation in the context of explaining the phenomenal notion of awareness falls prey to the just more representation objection, leaving unclear what it is about self-representations that renders them (but not other-representations) *for me*.

This undermines the self-representational approach to the explanatory gap presented in the previous section. For suppose it is true that epistemically transparent ontological reduction of phenomenal consciousness can proceed without closing the

explanatory gap (without epistemic reduction), namely, in case it is mediated by a sequence of more local reductive explanations. Still, this kind of epistemically transparent ontological reduction, although possibly available in the case of water and H₂O, is not available for consciousness and physical phenomena, because the reductive explanation of awareness in terms of representation (in Step 3 of the above explanatory sequence) fails. This is where the explanatory sequence is derailed because this is the step that happens to involve the attempt to transition from the phenomenal sphere to the non-phenomenal sphere. But the more general point is that *some* step in the explanatory sequence that leads from phenomenal consciousness to physical phenomena (such as synchronization with dlPFC activity) has to involve such a transition, and therefore *some* step in the explanatory sequence is bound to fail.¹⁸

If this is right, then the self-representational approach to the explanatory gap from §2 fails. The deep reason for this failure seems to be that while reductive explanation may exhibit sorites-like behavior, phenomenality does not: there is no way to get from the non-phenomenal to the phenomenal in a series of sorites-like steps. Because of this, some step in any relevant explanatory sequence must involve a discrete leap from the non-phenomenal to the phenomenal. Whatever step that is, we can expect reductive explanation to fail there, due to the explanatory gap, thus vitiating the sequence of reductive explanations that enables an epistemically transparent reduction. Of course, if phenomenality *could* be shown to exhibit sorites-like behavior, then the worry would dissipate and the self-representational approach to the explanatory gap might be viable after all. It is just that going from something non-phenomenal to something phenomenal in a sorites-like series of steps seems on its face utterly implausible.

4. Self-Representationalism and Epistemic Opacity

I think this is the deepest objection to self-representationalism. In fact, I am persuaded by Levine that there is something fundamentally mysterious about for-me-ness that is not removed simply by citing self-representation. Levine is right that the question of subjective significance applies with equal force to self-representation as to other-representation. For a self-representing state too, we can ask what it is about the state

that makes it represent itself *to me*, rather than merely represent itself *in me*. In this section, I present the reaction I think a self-representationalist ought to have to Levine's objection; the reaction is more concessive than confrontational.

The first thing I want to point out is that although I am keen to defend type-B2 self-representationalism, self-representationalism as such admits of many varieties, including type-A, type-B1, and dualist varieties.

Recall that according to dualism, consciousness is ontologically irreducible to any other properties, and is therefore ontologically primitive. An irresponsible kind of dualism would maintain that the phenomenal realm is completely dissociated and insulated from the physical realm. A responsible dualism would connect the phenomenal realm to the physical realm via laws of nature – probably causal laws – that dictate what phenomenal property instantiations are caused (under what conditions) by what physical property instantiations. Because phenomenal properties are ontologically primitive, these laws of nature would be themselves primitive. As a result, phenomenal properties would supervene on physical properties with *nomological* necessity, but not with *metaphysical* necessity.

Given the distinction between phenomenal and non-phenomenal notions of (self-)representation, dualist self-representationalism can come in two varieties. The first combines metaphysical supervenience of phenomenal consciousness on self-representation, construed phenomenally, with nomological supervenience of self-representation, so construed, on microphysical properties. The second, more interesting variety combines nomological supervenience of phenomenal consciousness on self-representation, construed *non-phenomenally*, with metaphysical supervenience of self-representation, so construed, on microphysical properties. In the former, the primitive laws of nature connect microphysical properties with a phenomenal self-representation, which is seen to be part of the phenomenal structure of consciousness. In the latter, there is a kind of self-representation that is fully reducible to the microphysical, and it is only this kind of self-representation that *causally* brings about phenomenal consciousness, in accordance with some primitive laws of nature.¹⁹

Thus, a type-E (say) dualist, who holds that phenomenal properties are causally inert, could be a self-representationalist, and in two kinds of ways. Type-E1 self-representationalism is the view that self-representation is part of the phenomenal structure of consciousness, which nomologically supervenes on the microphysical.

Type-E2 self-representationalism is the view that a microphysically reducible self-representation is the causal basis (hence nomological supervenience base) of phenomenal consciousness.

The point is that neither type-E1 nor type-E2 self-representationalism is threatened by Levine's just more representation objection. This is not surprising, since the objection is not *meant* to threaten them. But it does bring out the difference between self-representationalism as such and self-representationalism as an attempt to address the explanatory gap. The following two theses are obviously different:

- (T1) Self-representationalism neutralizes the explanatory gap.²⁰
- (T2) Self-representationalism is true.

Levine's objection threatens T1, but not T2. It is thus not an objection to self-representationalism, strictly speaking, but to something else.

This is important, because the master argument for self-representationalism (from §1) can be readily reframed in such a way that it does not *require* that the relevant kind of inner awareness be recovered by self-representation. Clearly, it can be deployed in pursuit of type-E1 self-representationalism if the 'right way' cited in Premise 4 is allowed to be the *phenomenal* (or *subjective*) way. But it can also be understood as arguing for type-E2 self-representationalism: the premises in this argument involve a modal operator, but while it is natural to interpret the modal force of those premises as metaphysical, the argument can be reframed as involving rather nomological necessity (without commenting on whether it is *merely* nomological necessity). The result would be this:

- 4*) Nonomologically-necessarily, for any entity X and subject S, S is aware of X in the right way iff S represents X in the right way.

With this weakened premise in place, and leaving all other premises untouched, we can obtain the following weakened conclusion:

- 9*) Nonomologically-necessarily, for any mental state M and subject S, such that S is in M, M is phenomenally conscious iff there is a mental state M*, such that (i) S is in M*, (ii) M* represents M in the right way, and (iii) M=M*.

This guarantees that *at least* type-E2 self-representationalism is right. The weakened master argument thus concedes that self-representationalism may not recover for-me-ness, or subjective character, but insists on the following two points: (a) self-representationalism can at least *accommodate* this for-me-ness; (b) no other theory of phenomenal consciousness can accommodate it. This is not everything a self-representationalist might want, but it is not all that weak a conclusion either.

Of course, not only dualist versions of self-representationalism fail to neutralize the explanatory gap; type-B1 self-representationalism does as well.²¹ And so a self-representationalist might consider reverting to type-B1 self-representationalism in light of the just more representation objection, conceding that the reduction of consciousness to self-representation (construed non-phenomenally) is epistemically opaque – due to the epistemic opacity of explaining awareness in terms of representation. Thus someone who is impressed with both the weakened master argument for self-representationalism and the just more representation objection could still embrace the disjunction of type-B1 and dualist self-representationalism. Both are forms of self-representationalism that cohabits with a persisting explanatory gap.

What would lead one to prefer type-B1 self-representationalism over dualist self-representationalism is, of course, an antecedent commitment to physicalism. Consider what Perry (2001) calls ‘antecedent physicalism,’ the view that physicalism should be our default position – we should be physicalists pending reasons not to be (physicalism is innocent until proven guilty, if you will). Someone who is impressed with both the weakened master argument for self-representationalism and the just more representation objection, but also embraces antecedent physicalism, would be naturally led to what we may call ‘antecedent type-B1 self-representationalism.’

For my part, this is indeed where I find myself led. I have already indicated why I am impressed with the weakened master argument for self-representationalism and the just more representation objection. As for antecedent physicalism, it should not be confused with physicalism as an unargued-for article of faith, nor with physicalism as an attitude rather than a thesis (Ney 2008), both of which do not call for argumentation.²² An argument for antecedent physicalism is needed, but the argument needed is not nearly as strong as the argument needed to establish all-things-considered physicalism. What it calls for is a *prima facie* rather than *ultima facie* case for physicalism. This is a burden we can certainly meet. Thus, citing

Occam's razor as a reason to adopt a single type of properties over a duality thereof, while an underwhelming argument for all-things-considered (*ultima facie*) physicalism, is a perfectly cogent argument for antecedent (*prima facie*) physicalism.²³

Of course, antecedent type-B1 self-representationalism does allow that, given appropriate reasons, one might have to relinquish type-B1 self-representationalism. So if one were inclined to reject epistemically opaque reduction as incoherent, or as otherwise necessarily false, say for Chalmers and Jackson's reasons, one would have to reject type-B1 self-representationalism. Thus someone who is impressed with both the weakened master argument for self-representationalism and the just more representation objection, but also accepts Chalmers and Jackson's argumentation, would be naturally led to dualist self-representationalism.

The dialectical upshot seems to me to be this. The issue of whether there is a case for self-representationalism and the issue of whether there is a case for physicalism are orthogonal, since one can be a self-representationalist without being a physicalist or a physicalist without being a self-representationalist. The problem of the explanatory gap is relevant to the issue of whether there is a case for physicalism, not to the issue of whether there is a case for self-representationalism. Given the 'just more representation' consideration, epistemically transparent reduction of phenomenal consciousness seems elusive. Whether *some* reduction may nonetheless be achieved depends on whether there is another kind of reduction to be had. That is, it depends on the viability in general of epistemically opaque reduction (or whether it may sometimes make sense for us to believe that feature F reduces to feature G even though we cannot quite see how it could). However this debate is resolved will determine which of type-B1 self-representationalism and dualist self-representationalism is more plausible.²⁴

Conclusion

As just mentioned, the problem of the explanatory gap is not directly relevant to the issue of whether there is a case for self-representationalism. In a weakened form, the master argument for self-representationalism (presented in §1) does not require that the subjective character of experience be recovered by self-representation, only that it be accommodated. At the same time, one might have wished that self-

representationalism would neutralize the explanatory gap (that would certainly constitute a major advantage of the view), which turns out to be unlikely. Although consideration of the sorites-like behavior of explanatory sequences inspires initial confidence, upon closer examination the prospects dim as the failure of reductive explanation of awareness in terms of representation comes to the fore. Nonetheless, there may yet be hope for a physicalist variety of self-representationalism, namely, if either epistemically opaque reduction turned out to be possible or phenomenal consciousness turned out to exhibit the sorites-like behavior that reductive explanation does. The former would prop up type-B1 self-representationalism, the latter type-B2 self-representationalism. My inclination, on the basis of the entire array of considerations examined here, is to adopt what I have called antecedent type-B1 self-representationalism.²⁵

References

- Block, N.J. and R. Stalnaker 1999. ‘Conceptual Analysis, Dualism, and the Explanatory Gap.’ *Philosophical Review* 108: 1-46.
- Brentano, F. 1874. *Psychology from Empirical Standpoint*. Edited by O. Kraus. English edition L. L. McAlister. Translated by A. C. Rancurello, D. B. Terrell, and L. L. McAlister. London: Routledge and Kegan Paul, 1973.
- Buras, T. 2009. ‘An Argument Against Causal Theories of Mental Content.’ *American Philosophical Quarterly* 46: 117-130.
- Chalmers, D.J. 1996. *The Conscious Mind*. Oxford and New York: Oxford University Press.
- Chalmers, D.J. 2002. ‘Consciousness and its Place in Nature.’ In D.J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford and New York: Oxford University Press.
- Chalmers, D.J. and F.C. Jackson 2001. ‘Conceptual Analysis and Reductive Explanation.’ *Philosophical Review* 110: 315-361.
- Dretske, F.I. 1981. *Knowledge and the Flow of Information*. Oxford: Blackwell.
- Dretske, F.I. 1988. *Explaining Behavior*. Cambridge MA: MIT Press.
- Fiala, B. Ms. ‘The Phenomenology of Explanation and the Explanation of Phenomenology.’
- Fiala, B. Forthcoming. **TITLE**. PhD Dissertation (University of Arizona).
- Fodor, J.A. 1990. *A Theory of Content and Other Essays*. Cambridge MA: MIT Press.

- Horgan, T. and J. Tienson 2002. ‘The Intentionality of Phenomenology and the Phenomenology of Intentionality.’ In D.J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford and New York: Oxford University Press.
- Kriegel, U. 2009. *Subjective Consciousness: A Self-Representational Theory*. Oxford: Oxford University Press.
- Levine, J. 1983. ‘Materialism and Qualia: The Explanatory Gap.’ *Pacific Philosophical Quarterly* 64: 354-361.
- Levine, J. 1993. ‘On Leaving out What It’s Like.’ In M. Davies and G. Humphreys (eds.), *Consciousness: Psychological and Philosophical Essays*. Oxford: Blackwell.
- Levine, J. 2001. *Purple Haze: The Puzzle of Consciousness*. Oxford and New York: Oxford University Press.
- Levine, J. 2006. ‘Awareness and (Self-)Representation.’ In U. Kriegel and K. Williford (eds.), *Self-Representational Approaches to Consciousness*. Cambridge MA: MIT Press.
- Loar, B. 1987. ‘Subjective Intentionality.’ *Philosophical Topics* 15: 89-124.
- Loar, B. 2002. ‘Phenomenal Intentionality as the Basis for Mental Content.’ In D.J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford and New York: Oxford University Press.
- Millikan, R.G. 1984. *Language, Thought, and Other Biological Categories*. Cambridge MA: MIT Press.
- Ney, A. 2008. ‘Physicalism as an Attitude.’ *Philosophical Studies* 138: 1-15.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. Cambridge MA: MIT Press.
- Pollock, J.L. and J. Cruz 1999. *Contemporary Theories of Knowledge* (second edition). Oxford: Rowman & Littlefield.
- Rosenthal, D.M. 1990. ‘A Theory of Consciousness.’ ZiF Technical Report 40, Bielfeld, Germany. Reprinted in Block et al. 1997.
- Rosenthal, D.M. 2005. *Mind and Consciousness*. Oxford: Oxford University Press.

¹ I explain what these qualifications mean toward the end of Chapter 4 of the book. Their exact nature will not matter here, so I will not go into it here.

² The latter is a determinate of which the former is a determinate. As is common, what makes X the X it is is a determinate of what makes it an X at all.

³ Here the ‘because’ must be understood as denoting a constitutive rather than causal explanation. That is, it is not the ‘because’ of ‘I am a bachelor because I never met the right woman,’ but the ‘because’ of ‘I am a bachelor because I am an unmarried man.’

⁴ For higher-order theory, see (most notably) Rosenthal 1990, 2005.

⁵ The argument is in actuality a little more complicated than this, because there are in fact three possible views here of what makes mental states phenomenally conscious: (a) that each is represented by itself; (b) that each is represented by a numerically distinct state; (c) that some are represented by

themselves and some by numerically distinct states. Ruling out (b) is thus insufficient. Ruling out (c) is part of the argument for (a). For details, see SC Ch. 4.

⁶ I argue for this in Ch. 6 of SC.

⁷ Certainly the causal relation of 'x causes y' is anti-reflexive, since nothing can cause its own occurrence, but other, subtler causal relations are often anti-reflexive as well, and as I argue in SC Ch.6, those adverted to by Dretske, Millikan, and Fodor in fact are.

⁸ In the first step, it is based on the first-person impression that motivates the conception of the structure of phenomenal character presented above; in the second and third steps, it is based on some kind of a priori reasoning; in the fourth step, on various forms of philosophical argumentation (namely, those that undermine higher-order theory); in the fifth step, on another kind of philosophical argumentation (from the prior commitment to naturalization); and in the sixth, by a combination of empirical and speculative considerations. Note that with the exception of the very last step, all steps in this reasoning can be understood as broadly a priori.

⁹ Again, compare the proof of Fermat's Last Theorem. Having read through the 100-page proof, one does not find oneself in a position to enjoy the experience of direct grasp of *why it should be* that $x^n + y^n \neq z^n$ whenever x , y , and z are non-zero integers and $n > 2$. But with sufficient acumen in Number Theory, one might just find oneself in a position to trust that the theorem does in fact hold.

¹⁰ I should state that almost everything I know about the connection between the explanatory gap, on the one hand, and the nature of understanding and its phenomenology, on the other, I learned from my friend and student Brian Fiala (see Fiala Ms, Forthcoming).

¹¹ I am assuming here a tight connection between explanation and understanding. One possibility is that explanation be construed as that which produces an appropriate state of understanding. Another is to reverse the order of explication here and construe understanding as the state which an appropriate explanation is supposed to licit. Either way, we can treat explanation and understanding as correlatives.

¹² For example, on certain interpretations, the view that there are protophenomenal properties, which are neither physical nor phenomenal, but which phenomenal properties can be reductively explained in terms of, would entail that phenomenal properties are not ontologically primitive, even though they are irreducible to physical properties.

¹³ This could be either what Chalmers calls type-D dualism or what he calls type-E dualism. The former denies the causal closure of the physical domain, the latter the causal efficacy of phenomenal properties. A final type of view Chalmers discusses is type-C materialism, but as he notes (correctly, in my opinion), that type of view is fundamentally unstable and collapses under different assumptions to one of the other types of view.

¹⁴ Thus, a subject who is in a position to establish the identity of water and H₂O is also in a position to explain the water facts in terms of the H₂O facts.

¹⁵ We might also put the argument in support of Premise 2 above as follows: reduction is always epistemically transparent; ontological reduction without epistemic reduction would be ontologically opaque; therefore, ontological reduction must be accompanied by epistemic reduction; and therefore, if phenomenal properties are epistemically irreducible to physical properties, then phenomenal properties are ontologically irreducible to physical properties.

¹⁶ It might be objected that type-B2 materialism is incoherent, on the grounds that epistemic transparency goes along with epistemic reducibility, such that as long as there is an explanatory gap between two phenomena, there cannot be any epistemically transparent connection between them. It is more natural, according to the objector, to suppose that just as reductive explanation turns out to exhibit sorites-like behavior, so does epistemic transparency: perhaps there is epistemic transparency within every individual step of the explanatory sequence, but the entire sequence is not epistemically transparent. My response is that although the notion of epistemic transparency is technical, so that one could define it any way one wished, including in such a way that its logical behavior is pegged to that

of reductive explanation, conceptually tying the two notions in this way would render Chalmers and Jackson's sub-argument question-begging. The argument is that ontological reduction requires epistemological reduction because ontological reduction requires epistemic transparency and epistemic transparency delivers epistemological reduction. But if 'epistemic transparency' simply *means* more or less the same as 'epistemic reduction,' then we still have no argument for the claim that ontological reduction requires epistemic transparency/reduction.

¹⁷ Consider: Tim says to Tom 'You know, Tom, 328.57 is greater than 174.16,' and Tom replies 'Yes, Tim, I'm aware of that.' Taking this exchange at face value, Tom is giving voice to his standing, tacit, non-phenomenal state of awareness that $328.57 > 174.16$. This non-phenomenal notion of awareness may very well admit of a self-representational treatment, even where the notion of representation is not phenomenally construed.

¹⁸ As we just saw, it is possible to adopt a notion of awareness that pushes the crucial step to Step 2, and it is possible to adopt a notion of representation that would push the crucial step to Step 6 (or perhaps 5). But in an explanatory sequence that leads us from phenomenal consciousness to synchronization with dlPFC activation, *some* step must involve a transition from the phenomenal to the non-phenomenal.

¹⁹ There is also the possibility of combining (i) metaphysical supervenience of phenomenal consciousness on phenomenal self-representation, (ii) metaphysical supervenience of non-phenomenal self-representation on microphysics, and (iii) nomological supervenience of phenomenal self-representation on non-phenomenal self-representation. I am of two minds about how motivated this kind of dualist self-representationalism is.

²⁰ I use the term 'neutralizes' to cover two possibilities: that the explanatory gap is bridged, and that the explanatory gap becomes something we can live with, say because we have an epistemically transparent reduction that illuminates why there is an explanatory gap, as in the approach to the explanatory gap sketched in §2. Thus both type-A and type-B2 materialism 'neutralize' the explanatory gap, even though only the former *bridges* it (or 'closes' it).

²¹ Type-B1 materialism insists on the ontological reducibility of consciousness to physics but accepts that the reduction is epistemically opaque and leaves the explanatory gap untouched. As Chalmers and Jackson (2001) note, the metaphysical supervenience it posits between consciousness and physics is as epistemically primitive as the nomological supervenience posited by dualism: there is no explanation of it, merely brute assertion.

²² This is true of Ney's physicalism-as-an-attitude because the latter is not truth-apt, and argumentation – in the relevant sense, involving the notion of validity (where true premises would necessitate true conclusion) – is only called for where a truth-apt statement is at stake.

²³ Indeed, this seems to be the view that Levine should adopt. Since he endorses the thought of embracing ontological reduction without epistemic reduction (Levine 1993), and at the same time seems to hold that self-representationalism comes closest to meeting the explanatory burden of physicalism and is the climax of physicalist attempts to address the explanatory gap (Levine 2001), he should certainly embrace something like antecedent type-B self-representationalism (of one of the two kinds).

²⁴ It is worth keeping in mind that, given antecedent physicalism, in this debate the physicalist is playing defense and the dualist offense. Thus as long as the debate is unresolved we are free to adopt the physicalist position.

²⁵ For comments on a previous draft, I am greatly indebted to Brian Fiala, ongoing interaction with whom has influenced the paper. I am also grateful to Shaun Nichols for another set of comments on an earlier draft and have benefited from interactions with Stephen Biggs, David Chalmers, Jennifer Corns, and Sebastian Watzl.