Folk Ontology and the Meta-Problem of Consciousness: Commentary on Weisberg-Physicalism

Uriah Kriegel

Forthcoming in *Journal of Consciousness Studies*
(book symposium on J. Weisberg, *Explanatory Optimism about the Hard Problem of Consciousness*)

Josh Weisberg develops a form of physicalism which attempts to (a) show why there is no *ultima facie* explanatory gap between consciousness and the physical world, while (b) making us see why there nonetheless is a *prima facie* explanatory gap. The former constitutes a solution to the problem of consciousness, the latter a proposal regarding the *meta*-problem of consciousness (the problem, roughly, of understanding why there is a problem of consciousness to begin with). Together, they are intended to produce a *debunking* explanation of the explanatory gap, and a comprehensive approach to the philosophical puzzlement generated by the existence of consciousness in an otherwise purely physical world. I will argue that Weisberg-physicalism is less successful on the meta-problem of consciousness than on the first-order problem and will sketch an alternative approach, one that traces the problem of consciousness back to the structure of our folk ontology.

1. 21st-Century Physicalism

At the beginning of the 21st century, David Chalmers (2003) divided physicalism into two main forms: those that reject the existence of an "explanatory gap" (Levine 1983) between the phenomenal and physical realms ("type-A physicalism") and those that accept there is an explanatory gap but deny there is any ontological gap behind it ("type-B physicalism"). The "explanatory gap" itself was construed as follows: there is an explanatory gap between realms P and Q iff a perfectly reasoner with complete knowledge of the P-facts could not deduce the Q-facts – in other words, iff there are no a priori connections between P-facts and Q-facts.

This conception of the physicalist space of options seems to me doubly problematic. First, it forces the physicalist into an embarrassing choice somewhat artificially. On the one hand, type-A physicalists, who deny any explanatory gap, naturally end up looking like disguised eliminativists, because that which they manage to make room for in their world never looks sufficiently like phenomenal consciousness proper. On the other hand, type-B physicalists, who propose to assert physicalism while essentially ignoring the explanatory gap, end up looking like

ideologues clinging to their physicalist faith despite being unable to see how there could be such a thing as phenomenal consciousness in a purely physical world.

The second problem with Chalmers' taxonomy is that the explanatory gap, understood as lack of a priori connection between the physical and the phenomenal, may be too weak to capture the full force of the problem of consciousness. Making the case for this is far larger a task than I can take up here, but I have long felt a discomfort with the characterization of the problem surrounding consciousness and physicalism in terms of an explanatory gap. A deeper source of our instinctual resistance to physicalism, it seems to me, is our sense that consciousness and matter are simply "cut of different cloth" (compare the "just too different" objection to naturalistic moral realism in metaethics – see Enoch 2011: 4). Saying that the phenomenology of sadness, say, *just is* this or that neural process feels like a *category mistake*, a silly mistake akin to saying that justice is nothing but cement mixed with wood fiber, say, or that the number 7 is "nothing over and above" tomato sauce. In all these cases, we clearly don't need empirical inquiry to establish non-identity – proper grasp of *what we're talking about* seems sufficient. Thus what makes consciousness problematic is not *just* that we can't seem to derive a priori phenomenal facts from physical facts, but also that we seem able to derive a priori that the phenomenal is something over and above the physical.

Obviously, making the case for this category-mistake conception of the problem of consciousness would be a complicated matter.

Given these two points, what the physicalist should really be doing, it seems to me, is (a) identify something about conscious experiences that makes them appear categorically different from physical objects, but (b) provide a fully physical explanation of this appearance that would then function as a *debunking* explanation. In a debunking explanation, one tries to show that what *causes* the belief, intuition, or appearance that *p* is not the putative fact that *p*, but some other, unrelated set of facts; and in consequence, that we have no reason to *endorse* the appearance/belief/intuition that *p*. The goal of the 21st-century physicalist should be to present this kind of debunking explanation of the appearance of category mistake. That would both (a) make us understand the robust, resistant appearance of a category mistake in assimilating consciousness to the physical and at the same time (b) recommend that we reject this appearance.

In his attractive new book *Explanatory Optimism about the Hard Problem of Consciousness*, Josh Weisberg wisely takes exactly this approach (Weisberg 2024). Weisberg offers a developed account of the "appearances to be explained" in the first half of the book, then articulates a novel debunking explanation of these appearances in terms of "automated compression theory" in the second half. In the next section, I present the essentials of these proposals. Later, I will part ways with Weisberg's conception of the appearances-to-be-explained.

2. Weisberg Physicalism

As noted, the first half of Weisberg's book develops a sophisticated account of that which generates the appearance of an explanatory gap. Weisberg's proposal is that phenomenal properties appear specially problematic because they seem Directly-accessible, Indescribable, Simple, and Contingently-connected (DISC). This group of four features divides into two sub-groups, though: Direct-acessibility and ISC-iness.

Suppose you feel sad. Your sad feeling has many properties: occurring on a Wednesday, lasting an hour and a half, making it hard for you to concentrate on work, as well as the phenomenal property of feeling *like this*. I say "like this" because it's hard to put in words that specific feeling of sadness (in literal, descriptive language). Indeed, it seems quite *impossible* to informatively describe the feeling of sadness to someone who never felt it. We can describe how the feeling of sadness relates to other things, but how it feels intrinsically seems indescribable.

Part of the reason for this apparent indescribability is that the feeling of sadness appears to lack any internal structure. There may be some feelings which are composites of other feelings. If you feel sad about not getting the prize for best consciousness paper of the year, but happy that your best friend got it, you might experience a bittersweet feeling. This bittersweet feeling *could* be described by bringing out its composite structure. But phenomenal properties cannot decompose indefinitely, so this kind of compositional description must bottom out in the citing of "phenomenal simples" lacking any internal structure. Feeling sad seems like a good candidate, but even if it's not itself a phenomenal simple, phenomenal properties can be described intrinsically only insofar as they can be decomposed into a reservoir of phenomenal simples. These make all phenomenological description possible but are themselves indescribable.

So, if feeling sad is a phenomenal simple, then although we can describe how feeling sad relates to other things, we cannot describe how feeling sad feels intrinsically. But this is what is essential to the feeling of sadness: how it feels, intrinsically. Accordingly, when describing its relations to other things we don't describe what the feeling of sadness *is*, but only how it relates to other things. This suggests that part of our conception of the feeling of sadness is as something whose essence is given by its intrinsic feel, with its relational profile being merely *contingent*.

Thus the feeling of sadness is apparently ISC-y: indescribable, simple (undecomposable), and contingently connected to other things (its essence being its intrinsic feel). The other crucial feature of phenomenal properties, for Weisberg, is their apparent direct-accessibility. This feature ostensibly pertains not to the intrinsic character of the feeling of sadness but to our epistemic relation to it. Not only I feel sad right now, I am also *aware* that I feel sad, and my awareness of my sadness seems doubly unmediated. First, it does not seem mediated by inference: I do not *infer* that I feel sad from my awareness of something other than the sadness.

No, it's the sadness itself of which I'm aware. Second, says Weisberg (2024: 55), my awareness doesn't seem mediated by *causal processes*. For instance, there isn't any temporal gap between the onset of my sad feeling and my becoming aware of it.

What explains the DISC-y appearance of feeling sad? Our cognitive system is in the first instance an information processing system, and two features in particular that enhance the performance of such systems interest Weisberg: automation and compression. The former will explain the appearance of direct accessibility, the latter the appearance of ISC-iness.

Automation arises in the context of the speed-for-accuracy tradeoff faced by our cognitive system. Is this a stick or a snake in the grass? The herpetologist needs to take her time making sure it's a snake, so for her there is a premium on accuracy. The caveman wading through the forest needs to make sure he's not bitten, so for him a quick-and-dirty (read: fast-if-inaccurate) snake-detector is more valuable. An automatic process is one that puts a premium on speed rather than accuracy in this way. Clearly, many cognitive tasks are carried out by modular processes which are entirely automatized: informationally encapsulated sub-personal processes that process information quickly and relatively effortlessly (from the subject's perspective).

Compression arises from another design tradeoff – between the costs of transmitting and storing information, on the one hand, and the costs of acquiring a more complex symbol system ("language") to convey this information. I can tell you that Jimmy is the son of the sister of my father. But if you and I decide to introduce the symbol "cousin," I'll be able to compress this into "Jimmy is my cousin." This is a short symbol easier to store and transmit and very easy to decode. The only cost is that we have to learn a bigger language. Since the symbol is so useful, given how often we'll use it to compress information, the tradeoff recommends itself. In contrast, suppose I want to tell you "Johnny is the son of the colleague of the accountant of . . . [ten minutes later] . . . of the grandson of Mahatma Gandhi." If we introduce the term "bloomp" into our language, I could compress all this information into "Johnny is Gandhi's bloomp," which is also easier to store and transmit. But the problem is that here the costs of acquiring the concept of a bloomp are going to dwarf the benefits of using it – neither you nor I may know of any other bloomp! Now, while the cases of "cousin" and "bloomp" are easy, the cognitive system must have a *general* strategy for deciding when to compress and how much. What matters to Weisberg is that such information compression is clearly a central aspect of good practice in information processing and is likely rife in our cognitive life.

How do automation and compression explain the DISC-y appearance of phenomenal properties? The automation-based explanation of direct accessibility (Weisberg 2024: 96-7) is an adaptation of an account due to David Rosenthal (1993), according to whom our awareness of our conscious states seems immediate because (i) it is not mediated by inference and (ii) although it *is* mediated by a causal process, we are not *aware* of this mediating process, with the result that our awareness seems causally unmediated to us. What Weisberg's account adds is a

story about why *it stands to reason* that we be unaware of the causal process mediating awareness of conscious states: because this process has been automated as part of the brain's information-processing optimization drive.

Next, Weisberg's compression-based explanation of ISC-y appearances. When we compress "child of sibling of parent" into "cousin," we generate an appearance of a single uniform relation – cousinhood – where in fact there is a structure involving a particular arrangement of three different relations. Still, here the compressed information is readily tractable, so we can easily decode ("decompress") the information compressed into "cousin" – in a way we couldn't easily with "bloomp." According to Weisberg (2024: 100-2), phenomenal concepts – the Mentalese terms for phenomenal qualities – are more like "bloomp" than "cousin" in this respect. If we could decompress them, the result would be a description of the phenomenal properties they pick out. But since in this case we're unable to decompress, we are unable to produce a description. Moreover, these concepts involve so much compression that we have lost track of the relational information they compress, which makes them seem simple to us. For instance, the web of neural relations that have been compressed into the Mentalese "feel sad" is so prodigious that we not only cannot retrace the underlying complex structure, it seems positively lacking in any internal structure.

3. DISC and the (Meta-)Problem of Consciousness

Weisberg's physicalism involves the conjunction of two explanatory claims:

> [*ACT Explains*] Automated Compression Theory explains why phenomenal properties appear to be Directly-accessed, Indescribable, Simple, and Contingently-connected.
> [*DISC Explains*] Phenomenal properties' appearing to be Directly-accessed, Indescribable, Simple, Contingently-connected explains why consciousness is philosophically problematic.

We can see *ACT Explains* as addressing the problem of consciousness, and *DISC Explains* as addressing the "meta-problem" of consciousness (Chalmers 2018). Very roughly, the former is the traditional problem of understanding how consciousness fits into the natural world, the latter the problem of understanding what it is about consciousness that makes its fitting into the natural world so singularly problematic.

There are surely fair questions to raise about whether automation and compression really explain the DISC-y features as they occur in phenomenal qualities. But here I want to bracket such questions and focus more on the prospects of *DISC Explains*.

I argued above that the problem of consciousness is not just the explanatory gap, but the fact that phenomenal consciousness seems categorically distinct ("cut of a different cloth") from

physical matter. I now want to argue that a property appearing DISC-y would not make it appear categorically different from physical properties. For many of these features are shared by such paradigmatically physical properties as mass and charge. Indeed, mass and charge are notoriously describable only in terms of the relations they bear to other fundamental physical properties, not *intrinsically*. These relations are, moreover, contingent: There are possible worlds where determinate masses interact differently with each other and with determinate charges, for instance. As physical fundamentals, mass and charge are also *simple* properties, in that they're not decomposable into collections of more basic properties. Thus mass and charge appear clearly ISC-y.

This leaves direct accessibility as the only DISC feature with any hope of explaining why phenomenal properties appear categorically different from physical properties. Mass and charge certainly don't seem directly accessible. At the same time, it's not only access to the phenomenal quality of feeling sad that has been automated in such a way that we're unaware of the causal process allegedly mediating between the sadness and the awareness of it. My visual access to the sphericality of soccer balls has been automated in the same way, such that I am likewise unaware of the causal process leading from the spherical ball to my awareness of it. I open my eyes and *bam* – here is a sphere. Likewise for other shapes. Thus it appears macrophysical properties of material objects seem directly accessible as well. No inference need mediate our awareness of them, and the causal process that mediates it has been automated out of our awareness.

I conclude that DISC features don't do much to explain the appearance of a categorical difference between the phenomenal and the physical. And it is this apparent categorical difference, I have claimed, that makes physicalism seem like a category mistake akin to saying that justice is cement mixed with wood fiber or that the number 7 is tomato sauce.

It might be objected that even if some paradigmatic physical properties are ISC-y and others are D-y, none are *both* ISC-y and D-y. However, what DISC-iness is supposed to explain is the appearance of categorical difference between phenomenal and physical properties. If ISC-iness is not suitable for the purpose, and nor is D-iness, it's unclear why combining them would do the trick. It'd be like saying that while cement cannot amount to justice, and nor can wood fiber, once you mix cement with wood fiber the result is justice.

In fairness, Weisberg never claimed that DISC-iness explains the appearance of categorical difference, only that it explains the appearance of an explanatory gap. More exactly, in fact, Weisberg's case for *DISC Explains* consists in arguing that Chalmers' "five ways" to dualism – the zombie argument, inverted qualia, epistemic asymmetry, the knowledge argument, and the unanalyzability of the concept of consciousness – can be traced back to the DISC-iness of phenomenal properties (Weisberg 2024: 60-2). I am somewhat doubtful that they can, but anyway for me the problem of consciousness is not in the first instance something that has been manufactured by a battery of arguments due to professional philosophers. There is

also, prior to all that, the pretheoretic sense of the distinctness of conscious awareness from physical matter. That is why we are all *instinctual* dualists (Bloom 2004). It is *adherence* to physicalism (including my own!) that is theoretically based, not *resistance* it. The resistance, I claim, is due to a pretheoretic sense that the two things are categorically distinct, in a way that makes their assimilation feel like a category mistake.

4. A Different Direction: Folk Ontology

If the fundamental problem around phenomenal properties is that they seem categorically different from physical properties, making physicalism feel like a category mistake, then perhaps underlying the problem is the categorial scheme operative in our spontaneous way of thinking about the world. Professional ontologists have as part of their job to identify the ontological categories to which belong the myriad entities of our world. Presumably, however, there is also a "folk ontology" built into our natural, pre-philosophical conception of the world. I would like to suggest that (i) this folk ontology places the phenomenal and the physical in different ontological categories, and (ii) this is the real reason the way consciousness fits into the natural world is so problematic.

I start with a preliminary elucidation of "ontological category." Our perceptual and cognitive systems engage in constant categorization. Thus, a brain encountering a certain stimulus may categorize it as a duck or as a rabbit depending on various factors, and the downstream treatment of information about the stimulus will be different if it is categorized one way or the other. Having categorized the stimulus as a duck or as a rabbit, the little brain may, if circumstances call for it, go on to further categorize it as a mammal or as a bird, but also, in certain circumstances, simply as an animal. Our conceptual scheme divides reality through a complex, multilayered, nested web of categories: rabbit is a species of the genus mammal, mammal a species of the genus animal, animal a species of living, and so on.

It is a controversial question in metaphysics whether there exists a "highest category," an all-encompassing *summum genus* of "being qua being" of which *everything* is an instance. And even if there is such a category, it is an open question whether our cognitive system would have any need to keep track of it, since it'd make no discriminations. In contrast, it is *not* controversial that there are "second-to-highest categories": categories which are species of only one genus, if there is a *summum genus*, or of no genus, if there isn't one. Plausibly, a compact set of such second-to-highest categories would be useful for the little brain in imposing initial, fundamental order on reality. Such second-to-highest categories – the categories which are most-general-but-still-discriminatory – I think of as paradigmatically *ontological* categories. But other sufficiently general or abstract categories (e.g., third-to-highest) may be counted as "ontological" categories (and we shouldn't expect a bright line in nature between categories that count as ontological and categories that don't.)

We may call the categories operative at this level of generality in pre-philosophical cognition "*folk*-ontological categories." The idea is that our conceptual scheme has familiar medium-level categories, such as *Table* and *Butterfly*, but also high-level ones, such as *Physical* and *Mental*, and this high-level part of our conceptual scheme may be said to constitute our "folk ontology."

A word is due on the "our" in "*our* folk ontology." One open question in this area is whether folk-ontological categories vary across cultures, periods of history, and so on, or are more invariant across human cognition. I wouldn't want take a stand on such (ultimately *empirical*) questions here. What does seem clear, I think, is that such highly abstract categories are *comparatively* invariant: we don't expect them to vary with class and age, across professions, and so on. The distinction between object and property ("substance" and "accident"), for instance, does not seem to be operative in my cognitive system only because I am a middle-class academic living in Texas. A struggling musician from Madrid probably cognizes the world in terms of objects and properties too. Whether it also characterized the folk ontology of ancient Japan, or of Ice-Age hunter-gatherers, are more difficult questions which we'll have to bracket here.

The hypothesis I would like to put forward is that our folk ontology divides reality into three (folk-)ontological categories: the physical, the mental, and the abstract. As elements in a useful categorization scheme, these function as mutually exclusive and jointly exhaustive categories: things are supposed to belong at least *and* at most to one of them. Thus, cement and tomato sauce are categorized as physical at this level of generality, the feeling of sadness and a thought of an octopus are categorized as mental, and justice and the number 7 as abstract. In consequence, claims like (a) justice is cement mixed with wood fiber, (b) 7 is tomato sauce, and (c) the feeling of sadness is such-and-such neural process all strike the little brain as category mistakes. This, I conjecture, is the source of the problem of consciousness. We may formulate the key thesis as follows:

> [*Folk Ontology Explains*] The fact that our folk ontology divides reality in such a way that the phenomenal and the physical belong to mutually exclusive ontological categories explains why phenomenal consciousness seems categorically different from physical matter.

Call this the *folk-ontology approach to the meta-problem of consciousness*.

Folk Ontology Explains is an alternative to Weisberg's *DISC Explains*, which I argued is ill positioned to recover the sense of categorical difference given that mass and charge are ISC-y while vision accesses the material objects around us in a seemingly direct way just as much as introspection does our conscious experiences. *Folk Ontology Explains* does much better on this, since it's *designed* to capture the appearance of categorical difference rather than the appearance of an explanatory gap. Consider that when students are first introduced to the idea

of physicalism, they "hear" it as implying that there are no selves and no thoughts or feelings – nothing mental. For them, claiming that everything is physical *automatically* excludes there being anything mental. If we were to force the problem of consciousness as experienced by them into an inconsistent triad, it might be this:

[T1] Everything is physical.
[T2] Something is mental.
[T3] If everything is physical, then nothing is mental.

Here T1 is motivated by a broadly scientific-naturalistic outlook, T2 by the direct awareness each of us has of their own conscious experiences, and T3 by folk ontology. In this inconsistent triad, rejecting T1 amounts to anti-physicalism, rejecting T2 amounts to eliminative physicalism, and the "comfort zone" of most philosophers of mind – adopting physicalism without denying the reality of consciousness – requires rejecting T3 and the folk ontology behind it. What creates the tension between T1 and T2, then, is the fact that T3 is so deeply embedded in our conceptual scheme.

It is instructive to compare the areas of philosophy concerned with the status of abstracta, such as metaphysics and philosophy of mathematics. Here the claim that everything is concrete – "nominalism" – virtually never takes the form of "reducing" abstracta to concreta. Instead, it takes an eliminativist form: the typical nominalist claim is that *there are no numbers* (Field 1980). The idea that numbers might exist but be "nothing but" concrete entities has almost never been defended, presumably because it seems like a simple category mistake. The assumption is that if everything is concrete, then nothing is abstract (though see Maddy 1981 and Zemach 1985 for exceptions). This is part of the same folk ontology, I claim, according which if everything is physical then nothing is mental.

I have argued that *Folk Ontology Explains* is more plausible than *DISC Explains*. But for Weisberg, *DISC Explains* has a huge advantage, insofar as it paves the way to a debunking explanation of the relevant sense of categorical difference. What does *Folk Ontology Explains* imply on the question of debunking? I close with preliminary discussion of this matter.

Recall that in a debunking explanation of the belief, intuition, or appearance that *p*, we show that what causes this belief/intuition/appearance is not the fact that *p* but some unrelated fact *q*. (This isn't intended to show that our belief/intuition/appearance is false, but only that if it's true it'd be just luck – we have no good *reason* to accept the belief/intuition/appearance.) To debunk our folk-ontological commitments, then, would be to show that our folk ontology was not formed responsively to the mind-independent structure of reality, but on the basis of some independent factors. If one instead argues that folk ontology divides reality as it does because of certain divisions in reality itself, the result would be a vindicating rather than debunking explanation of the sense of categorical difference.

One possible *vindicating* explanation might be in terms of space and time. It might be argued that the world really does include three fundamentally different kinds of being: beings that exist both in space and in time (notably material objects), which we call physical; beings which exist in time but not in space (consciousness), which we call mental; and beings that exist neither is space nor in time (e.g., numbers), which we call abstract. Combined with a realist account of space and time, this line of thought might suggest that our folk ontology tracks a real division between three fundamentally different ways of existing.

But *debunking* explanations may be envisaged as well. For instance, it might be argued that our folk ontology is ultimately based on the fact that we have three cognitively isolated ways of accessing reality: When we access a part of reality through *perception*, we call it physical; when we access a part of reality through *introspection*, we call it mental; and when we access part of reality through *reason*, or "rational intuition," we call it abstract. In this picture our folk ontology is actually *epistemologically* grounded, rather than grounded in the mind-independent structure of reality. After all, it is not something about *the world*, but something about *us*, that makes the world seem to divide into three mutually exclusive and jointly exhaustive ontological categories. Moreover, if the folk-ontological categorization is grounded in epistemic-access differences, nothing guarantees that the parts of the world therein accessed are exclusive of each other in mind-independent reality (nor, for that matter, does it guarantee that they are *exhaustive*). For all we know there is a single part of reality which is accessed in two cognitively isolated ways – perception and introspection – and that is what creates the appearance of distinctness. In this way, *Folk Ontology Explains*, *if* interpreted debunkingly, paves the way to reductive physicalism.

As we can see, then, *Folk Ontology Explains* can play out either as a debunking or as a vindicating explanation of our sense of categorial distinctness between consciousness and matter. I am not sure, for my part, which direction is more plausible, though I share Weisberg's "philosophical desire" for physicalism to come out more belief-worthy than anti-physicalism. My claim here has been that this is the terrain on which inquiry should be pursued.[1]

References

- Bloom, P. 2004. *Descartes' Baby*. New York: Basic Books.
- Chalmers, D.J. 2018. 'The Meta-Problem of Consciousness.' *Journal of Consciousness Studies* 25: 6-61.
- Chalmers, D.J. 2003. '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.
- Enoch, D. 2011. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford: Oxford University Press.
- Field, H. 1980. *Science without Numbers*. Princeton: Princeton University Press.
- Levine, J. 1983. 'Materialism and Qualia: The Explanatory Gap.' *Pacific Philosophical Quarterly* 64: 354-361.

- Maddy, P. 1981. 'Sets and Numbers.' *Noûs* 15: 495-511.
- Rosenthal, D.M. 'Higher-Order Thoughts and the Appendage Theory of Consciousness.' *Philosophical Psychology* 4: 155-167.
- Weisberg, J. 2024. *Explanatory Optimism about the Hard Problem of Consciousness*. London: Routledge.
- Zemach, E.M. 1985. 'Numbers.' *Synthese* 64: 225-239.

[1] For comments on a previous draft, I am grateful to Torin Alter and Jacob Berger.