How Can We Solve the Meta-Problem of Consciousness?

David J. Chalmers

I am grateful to the authors of the 39 commentaries on my article "The Meta-Problem of Consciousness". I learned a great deal from reading them and from thinking about how to reply.[1]

The commentaries divide fairly nearly into about three groups. About half of them discuss potential solutions to the meta-problem. About a quarter of them discuss the question of whether intuitions about consciousness are universal, widespred, or culturally local. About a quarter discuss illusionism about consciousness and especially debunking arguments that move from a solution to the meta-problem to illusionism. Some commentaries fit into more than one group and some do not fit perfectly into any, but with some stretching this provides a natural way to divide them.

As a result, I have divided my reply into three parts, each of which can stand alone. This first part is "How can We Solve the Meta-Problem of Conscousness?". The other two parts are "Is the Hard Problem of Consciousness Universal?" and "Debunking Arguments for Illusionism about Consciousness".

How can we solve the meta-problem? As a reminder, the meta-problem is the problem of explaining our *problem intuitions* about consciousness, including the intuition that consciousness poses a hard problem and related explanatory and metaphysical intuitions, among others. One constraint is to explain the intuitions in topic-neutral terms (for example, physical, computational, structural, or evolutionary term) that do not make explicit appeal to consciousness in the explanation.

In the target article, I canvassed about 15 potential solutions to the meta-problem. I expressed sympathy with about seven of them as elements of a solutions: introspective models, phenomenal concepts, independent roles, introspective opacity, immediate knowldge, primitive quality attribution, and primitive relation attribution. I summed up my own preferred path to a solution as follows:

> We have introspective models deploying introspective concepts of our internal states

[1] Thanks to François Kammerer for editorial help and to Kammerer and Wolfgang Schwarz for feedback.

> that are largely independent of our physical concepts. These concepts are introspectively opaque, not revealing any of the underlying physical or computational mechanisms. Our perceptual models perceptually attribute primitive perceptual qualities to the world, and our introspective models attribute primitive mental relations to those qualities. We seem to have immediate knowledge that we stand in these primitive mental relations to primitive qualities, and we have the sense of being acquainted with them.

This is not in itself a solution to the meta-problem, because in many respects it is more like an explanandum (what needs to be explained) than an explanans (a substantial explanation). But it can help point us in the direction of a solution by pointing to what needs to be explained. Perhaps it is not too hard to see why a cognitive system would have introspective models and independent introspective concepts involving introspective opacity. Somewhat more work is needed to explain the attribution of primitive qualities and relations, and the crucial sense of acquaintance remains very much in need of explanation. So I am especially interested to see if any of the commentaries help with this project.

By my count about 20 of the commentaries focus on solutions to the meta-problem of consciousness. Strategies include evidential and Bayesian strategies (Kammerer, Schwarz, Clark et al), the phenomenal concept strategy (Diaz Leon, Papineau), the attention schema (Graziano, Dewhurst and Dolega), revelation and qualitative inaccuracy (Michelle Liu, Schriner, Williford, Pereboom), underestimating the physical (McClelland, Strawson), dual systems approaches (Drescher, Fiala and Nichols, Haoying Liu, Storm), the access problem (Dennett), evolution of sensations (Humphrey), and control (Klein and Barron). There are also some cultural and sociological approaches which I address in “Is the Hard Problem Universal?”.

In what follows I will examine these potential solutions. I don’t want to set myself up as the grand arbiter of solutions to the meta-problem, so my discussions will often be short, just with a few key thoughts. I have said more in cases where I think the solution might be especially promising or where the discussion interfaces directly with material in the target article. At the end I will examine where things stand in the quest for a solution.

Evidential and Bayesian strategies

Among the most promising strategies explored in the commentaries are closely related evidential and Bayesian strategies laid out by François Kammerer, Wolfgang Schwartz, and jointly by Andy Clark, Karl Friston, and Sam Wilkinson. Kammerer argues for the *evidential approach*, on which problem intuitions arise from cognitive system's need to represent mental states as evidence. Wolfgang Schwarz outlines a *sensor variable* framework, itself a repackaging of his earlier "imaginary foundations" approach, on which problem intuitions arise from the use of states akin to sensor variables to represent the sensory states that serve as the evidential foundations of Bayesian cognitive processes. Clark et al endorse a verson of Schwarz's imaginary foundations view, spelt out in terms of a hierarchy of Bayesian processes in which some mid-level representations are held with something close to certainty.

I will start with Schwarz's sensor variable approach. There is something immediately appealing about this framework. Computational vision systems use sensor variables to represent inputs to their systems. Say a digital camera has a 1000 by 1000 array of pixels each of which can take 100 values each. Then the system will represent this with a 1000 by 1000 matrix of 100-valued variables. It need not represent anything more about the pixels themslves, such as their physical realization. And this large matrix will serve as a foundation for all later processing in the system. The matrix of sensor variables will ground inference to the physical world, but it will not be represented as part of the physical world. In this way it is interestingly analogous with the way sensory experiences seem to function in us. If a system uses something like sensor variables, perhaps reflection on them might lead a system to treat them much as we treat sensory qualities or sensory experiences, with associated puzzlement about how these qualities fit into the physical world.

An initial point is that at least in many cases where the human system uses analogs of sensor variables, they are unconscious. Stimulation of retinal rods and cones in effect gives rise to an array of sensor variables that is used by later processes. However, they do not correspond to anything in our experience or in our phenomenal intuitions. We have no access to these sensor variables, so on their own they cannot explain phenomenal intuitions. To handle this point, Schwarz could add a condition about *access* to his story: phenomenal intuitions are associated with accessed sensor variables.

A more serious problem is that sensory experiences are (or seem to be) deeply representational: visual experiences represent colors, shapes, and locations of objects in the external world. Sensor variables are not. At most they represent goings-on in a two-dimensional array. If sensor variables constituted our sensory experience, our experience would presumable be of some sort of “qualia array”: perhaps a distribution of colors or brightnesses in a two-dimensional manifold. But our sensory experience seems to be nothing like this. Perhaps there could be other creatures who have primary access to sensor variables and whose experience is like this. But this is not our situation, so sensor variables do not seem well-matched for an explanation of how things seem to us.

To handle this point, we need to move from pure sensor variables to representations of the external world. Schwarz talks briefly about this sort of move in “Imaginary Foundations” and in the current article, proposing a two-tiered Bayesian model with traditional sensor variables serving as inputs to the first tier and representations serving as inputs to the second. Such a model seems truer to the human cognitive system. Sensory experiences now correspond to the second-tier inputs, which are representational and as a bonus may be accessible too. On the other hand, in these second-tier inputs there is no obvious role for sensor variables. In the newer paper, Schwarz suggests briefly that sensor variables might be reused in our models of the external world. But this does not fit the human case well. Experienced colors and shapes have only an extemely loose and indirect relationship to retinal sensor variables. Insofar as there is overlap between sensory models and world models, it seems to be features of the world model (colors and shapes) that are reused in our sensory experiences. Visual experiences are or seem to be representations or experiences of colors and shapes. But now we have moved a long way from sensor variables.

Could one adapt Schwarz’s discussion to thes states, even if they no longer involve sensor variables per se? Of course a computer vision system might have states like these. For example familiar Marr-based systems build up layers of representation of the external world, and some layers (the 2.5D and 3D sketch) involve representation of external colors, shapes, and locations in a way that is reminiscent of sensory experience. These representations could be set up to be inputs to second-tier processing, and perhaps the earliest layer a system has access to. They could also serve as a foundation for much of what comes later. But the analogy with sensor variables is now quite unclear. Because these states involve so much reuse of world variables, it is no longer clear why they should seem so independent of the external world. One idea is that these states involve *representation* of the external world, and that when the system becomes aware of these states, representation is represented to the system as a special relation that is not part of the physical world. But at this point, sensor variables and the like are not explaining phenomenal intuitions. Rather, the crucial work is done by the way we model representation.

I think there is still value in Schwarz’s imaginary foundations idea, once stripped of a strong dependence on the sensor variable idea and once combined with an appeal to personal-level access.

On the modified picture, we need to have accessible states to treat as foundations or as evidence, and those states will seem special and nonphysical (and will be imaginary or illusions, if one is an illusionist). These evidential states will be representational states: representing objects in the environment as having certain colors and shapes ("Red square there"), or if want evidential states to be especially secure, representing ourselves as representing objects as having certain colors and shapes ("It looks to me as if there's a red square there"). This modified picture is close to Kammerer's evidential strategy.

Kammerer argues that (1) we should expect some mental states to be metacognitively treated as passive (belief-independent) evidence, (2) these states will be treated as self-presenting and as revealed, so (3) these states will generate problem intuitions. In effect, where Schwarz uses evidence to ground a version of the phenomenal concepts strategy (phenomenal concepts are independent of physical concepts), Kammerer uses it to ground a version of the immediate knowledge strategy (phenomenal states seem to be presented to us directly and self-evidently).

In more detail: Kammerer tells a persuasive story about why it is useful for a sophisticated cognitive system to be an evidential system – not just in the sense of using some states as evidence for other states (which any Bayesian system might do, even subpersonally), but in the sense of metacognitively representing certain states as evidence for other states. Here in effect he puts significant weight on access at the personal level. He argues that this will allow more flexible and reflective process of belief fixation as well as better sharing of evidence with others and rational assessment of the beliefs of others. He goes on to argue that some evidence should be treated as "passive evidence", holding independently of our beliefs.

Kammerer then makes a case that passive evidence states should be treated as self-presenting, where a state is self-presenting if being in that state provides evidence that we are in those states (for example, experiencing red gives evidence that we are experiencing red).[2] Finally and crucially, Kammerer argues that at least in humans, states that are treated as self-presenting will seem to be objects of acquaintance (presentation and revelation), and that this will lead to anti-physicalist intuitions.

I am persuaded that we need to treat some states as evidence, and it's natural for some states to be treated as passive evidence. My doubts concern the steps that follow. These steps turn on our using an "evidence-resemblance mechanism", which involves a tacit commitment to a resemblance theory of representation. We represent sensory states as resembling the external

[2] Scott Sturgeon (199x) uses the self-presenting nature of experiences (they serve as their own "canonical evidence") to support a version of the phenomenal concept strategy.

states they represent. On this picture, Kammerer argues, passive evidence states will be treated as providing evidence for whatever states they maximally resemble, so they provide evidence for themselves. If they are self-presenting in this way, it is impossible for this evidence to be misleading. We will treat these evidential states as being automatically the way the evidence presents them as being, i.e. as being revealed to us. If they are revealed to us, they cannot be physical, becuse they are not presented as physical. So if we treat them as revealed, we will treat them as nonphysical.

Now, I am doubtful about resemblance theories of representation, and I am doubtful that even our cognitive systems are committed to them. I also don't see the resemblance theory should apply more to passive evidential states than to any other states. I also don't quite see how we get from their being the way the evidence presents to their being fully revealed. We can accept that X is perfect evidence for X without accepting that X is fully revealed. So I'm not sure that this story really explains revelation and anti-physicalist intuitions.

Still, Kammerer notes that the evidence-resemblance story is just one highly speculative approach for using the evidential strategy to approach the meta-problem. There may well be others, and I agree that these are worth pursuing. In effect, the early steps of the evidential strategy make a strong case that we should expect certain states to be treated as evidence in our models of ourselves. I think there is also a reasonable case that some will be treated as basic evidence and as objects of a sort of immediate knowledge, needing no evidence beyond themselves. On the other hand, it is also plausible that we treat our believing something as basic introspective evidence; but familiarly we do not get nearly such strong problem intuitions for belief. So the residual question remains why those states that are treated as objects of immediate knowledge should also be treated as objects of acquaintance or revelation, and why they should support problem intuitions.

A somewhat related strategy is developed by Clark, Friston, and Wilkinson, who endorse a version of Schwarz's imaginary foundations claim. They put the central weight on explaining the certainty of our judgments about consciousness, offering an explanation in terms of a hierarchy of Bayesian predictive processes. On their view, qualia (or representations of how things look) may be objects of mid-level representations held with a high degree of certainty. They say this level of representation is useful in part to deal with scenarios where one is not certain about the external world. It makes sense to have a layer of relative certainty to be able to use even when external facts are in doubt.

A version of this strategy may help us to explain why we have beliefs about consciousness that are held with relative certainty. We might say that it helps explain *certainty intuitions*. But it's not obvious how explaining certainty intuitions this helps to explain problem intuitions more broadly. After all, we are certain of many things (mathematics, beliefs) that don't generate these intuitions. Clark et al tell the beginning of a story about how these intuitions could generate an appearance-reality distinction, with experiences on the appearance side. That's a start, but this is a good way from familiar gap intuitions. Perhaps Clark et al could tell a story about how these states of mid-level certainty help to generate these too, but the details are not laid out here.

One general limitation on the evidential approach is that it seems to work much better for the meta-problem concerning perceptual experience rather than experience in general. Perceptual experiences are certainly treated is basic evidential states. This is not so clear for other experiences, including say the experience of cognition or of action. Perhaps we treat these as introspectively basic evidence, but they do not seem to play anything like the same widespread evidential role that perceptual evidence plays. So while it is easy to make the case that any cognitive system would need to have a layer of perceptual states that it treats as basic evidence, it is much less clear that it needs a layer of cognitive states and the like that serve the same purpose.

The phenomenal concept strategy

Esa Diaz Leon and David Papineau are both advocates of the phenomenal concept strategy, and both advocate elements of it as a potential solution to the meta-problem.

Diaz Leon gives a masterful treatment of the connection between the phenomenal concept strategy and the meta-problem. She makes a strong case that the advocate of PCS should treat metaphysical intuitions differently from explanatory, modal, and knowledge intuitions. They can accept the latter three (explanatory gap, conceivability of zombies, Mary's new knowledge) and not treat them as illusory at all. They will reject certain inferences from there (ontological gap, possibility of zombies, knowledge of new facts) but the problem there lies in the inference, not in any sort of illusion. By contrast they cannot accept the correctness of metaphysical intuitions (e.g. consciousness is nonphysical). However, Diaz Leon argues that these intuitions are not as strong as the others, and may also involve some sort of bad inference, such as the headless-woman inference. In this way one could argue (though Diaz Leon does not put it this way) that the PCS is committed only to a very weak illusionism on which errors about consciousness arise only from bad inferences.

On this version of PCS, the core non-inferential problem intuitions that need explaining are explanatory, knowledge, and conceivability intuitions. Diaz Leon argues that these can all be explained by the cognitive isolation of phenomenal concepts from physical concepts. In the target article I echoed a claim I made in "Phenomenal Concepts and the Explanatory Gap", arguing that the phenomenal concpts that this strategy yields will either be too "thin" to explain the gaps or too "thick" to be physically explained. Diaz Leon notes that the "thin" claims in the 2007 paper turned on assuming realism about consciousness, which can't be assumed in the current context where we are interested in explanations of the problem intuitions that are consistent with illusionism. I think Diaz Leon is correct about this dialectical point: "thinness" in the sense of not accounting for the genuine realist explanatory gap is no objection to a PCS-based explanation of the problem intuitions.

Still, in the target article I criticized the PCS-based explanations of problem intuitions in terms of cognitive isolation (or independent roles) as well as explanations in terms of indexical and recognitional concepts for somewhat different reasons that I think still apply here. First, there is the familiar point that our concepts of belief seem to be cognitive isolated in the same way without generating problem intuitions of anything like the same strength. Second, there are many other indexical and recognitional concepts that are cognitively isolated without generating the problem intuitions.

Diaz Leon does not address the point about belief, but she addresses the point about indexical and recognitional concepts. She argues that phenomenal concepts differ from most indexical and recognitional concepts by not having any associated descriptive content. At this point Diaz Leon is in effect going beyond simply explaining problem intuitions in terms of cognitive isolation and appealing to further features of phenomenal concepts. I'm not sure how much this helps with the indexical case, as it's arguable that "I" and "now" lack any significant descriptive content. It's also not obvious to me that phenomenal concepts have less descriptive content that belief concepts. Speaking as a realist, I find it very implausible that phenomenal concepts lack descriptive content. On the face of it they have a rich and substantive content that characterizes their referent far more richly than any indexical does. It's this content that is implicated in Mary's rich knowledge of what it's like to see red. Diaz Leon says this rich knowledge can be explained in terms of recognitional abilities to recognize the same phenomenal type, but I don't think this nearly suffices to explain the knowledge in part for reasons famliar from discussions of the ability hypothesis. So while I think Diaz Leon's strategy of appealing to further special features of phenomenal concepts is an important one, I don't think the no-descriptive-content strategy used here succeeds.

Diaz Leon also suggests that she does not really need these further features to explain the problem intuitions. Cognitive isolation alone can explain the key intuitions: the conceivability of zombies, Mary's new knowledge, the a priori/explanatory gap. It's true that these are the intuitions that drive the classic argument against materialism (so they're highly relevant for the type-B defense of materialism). But I don't think these exhaust our problem intuitions. For a start, we have formally analogous gap intuitions for indexicals, but the gap seems much more substantive and robust in the phenomenal case. For example, the explanatory gap intuition involves an a priori derivability gap, but this doesn't come close to exhausting it. On the face of it the explanatory gap involves a chasm between different realms. There are also metaphysical intuitions about the qualitative nature of consciousness, and about its intrinsic and nonstructural character. There are episteimc intuitions about our seeming to be acquainted with consciousness or it seeming to be revealed to us. All of these need to be explained for a proper accounting of problem intuitions and their strength. Cognitive isolation may be a start on the project of explaining problem intuitions, but I think much more is needed.

We could put things by saying that there are both *negative* problem intuitions, which are those tied to the derivability gap (one can't derive phenomenal knowledge from physical knowledge, zombies are consistent), and *positive* problem intuitions which involve a positive characterization of consciousness (consciousness has a substantial qualitative nature, it involves acquaintance or presentation of certain natures, Mary gains substantial knowledge on leaving the room). The classic phenomenal concept strategy appealing to cognitive isolation does a reasonable job in explaining the existence of negative problem intuitions (I don't think it explains their truth, but that's not the main issue here). But it does not do nearly such a good job in explaining positive problem intuitions. Perhaps there is some way to augment the strategy to do that, but I don't think this has yet been done.

Papineau asks why I don't take the a priori derivability gap to be the explanation of the problem intuitions. One reason is that as above I think there are problem intuitions that go well beyond the derivability gap. I think those associated with the derivability gap (Mary's inability to know what it's like, and so on) are just an important subclass of problem intuitions. But a more basic reason is that the derivability gap is much more explanandum than explanans. The derivability gap is just a small generalization and abstraction from the conceivability of zombies and so on, and provides a correspondingly small explanation of them. And it simply raises another version of the meta-problem question: why do we think (or talk as if) we have these features that are not a priori derivable from physical features? That question is about as hard as the original question about zombies, Mary, and so on, and roughly the same answers apply.

Where Diaz Leon thinks that metaphysical intuitions such as the intuition that consciousness is nonphysical are less ubiquitous and less important than derivability-associated intuitions, Papineau thinks they are far more ubiquitous and important. He thinks it is these distinctness intuitions that are responsible for the explanatory gap, though he doesn't address the question of what explains them. I am somewhere between Diaz Leon and Papineau here: I think all the intuitions are important, and a number of different intuitions may play a role in generating the explanatory gap. Likewise, I don't think the hard problem of consciousness involves just the derivability intuitions or just the metaphysical intuitions. All of them play a role. The derivability intuitions play a particularly central role in my arguments against materialism—but the hard problem should not simply be identified with the arguments against materialism. It is much more general than that.

Papineau as well as Katalin Balog note that my article gives short shrift to the type-B materialism that they favor. I acknowledge it as a form of weak illusionism late in the paper, but I do not consider it at length. The reason is simply that I have argued against it at length in other work. If I had discussed it further here it would have been mainly to set it aside on similar grounds. Still, the type-B materialist certainly has their own distinctive line on the meta-problem, one that tends to lead to weak illusionism.

My main criticism of weak illusionism was that it doesn't help with the hard problem: why is there something it is like to be us. Papineau disagrees, saying that if we understand the hard problem as the metaphysical problem intuitions, it turns precisely on on the claims that consciousness seems intrinsic, non-physical, and so on. Here, I would say that the hard problem should certainly not be identified with the metaphysical intuitions. It is also not exactly the same as the derivability gao either, as I noted above. It is an explanatory problem: how can we explain why there is something it is like to be us? The hard problem is no doubt connected to the derivability problem as well as the metaphysical intuitions. But as I said in the target article, one can reject the metaphysical intuitions, holding for example that consciousness is physical, and the hard problem remains hard. Simply rejecting the metaphysical intuition doesn't made it easier to explain why there's something it's like to be us.

Now, to be fair, the type-B materialist may say that they don't need to use weak illusionism per se to solve the hard problem. They have their own way of doing that in terms of phenomenal concepts, psychophysical identities, and the like. I will argue against them, but my grounds will be similar to those I have offered before. So I don't claim to be offering any new reasons to reject type-B materialism in this paper. Still, the type-B materialist is not really offering any new reasons in their favor of their view either. By contrast I think strong illusionism offers a distinctive route of its own to addressing the hard problem, with distinctive arguments in its favor that arise from the meta-problem. That's why I focused more on it here.

The attention schema theory

Michael Graziano defends his attention schema theory, which says that our intuitions about awareness arise from internal models of attention, as an approach to the meta-problem. Joe Dewhurst and Krzysztof Dolega also defend it, in part by combining it with the predictive processing framework. In the target article I had two main criticisms of AST. First, as it stands it does not give a solution to the meta-problem. Second, a general solution to the meta-problem will involve more than an attention schema.

On the first point, Graziano expands on his previous sketchy remarks that suggest a treatment of the meta-problem in terms of introspective opacity. He says our model of attention contain no information about its physical properties, so as far as one can tell from the attention schema, it lacks physicality. If the system makes claims about itself on this basis, it will claim to have a subjective, nonphysical grasp of objects.

In effect, Graziano's approach requires the brain to make the headless woman fallacy (we don't see the woman's head, so we seeing her as having no head): it moves from not representing attention as physical to representing it as nonphysical. It is easy to see why it should do the former. It is harder to see why ti should do the latter. Usually failing to represent something as X does not lead to representing it as non-X. If I fail to represent an object I am looking at as heavy, I need not reflect it as light. I might simply fail to represent its weight. Likewise, one would normally expect that failing to represent attention as physical would not lead to representing it as nonphysical. It would just mean being neutral on the question of physicality, or not addressing it.

Graziano responds to a similar complaint by Kammerer by maing a stronger claim. It is not just that we fail to represent physical features of consciousness. Instead, we intuitively understand consciousness as something for which physical features are irrelevant. It is as if our representation of consciousness has no room for physicality. Still, I am not sure this suffices to explain intuitions. When we think about computation, or algorithms running on a computer, we do not represent physical features. More strongly we have an intuition that those physical features are irrelevant to the computation as a computation. But still, we do not have the intuition that it is nonphysical. So Graziano's approach still requires us to commit the headless-woman fallacy in a way that needs explanation.

Dewhurst and Dolega respond to the problem in a similar way. They say that it is because the attention schema leaves out certain details that it represents awareness as something mysterious and ethereal. Representing the mechanistic details of attention would make the model less efficient and less precise. As a result, the system does not represent details, and the resulting "sparse" representation yields problem intuitions. Here my response is as before. Leaving out details about X doesn't usually lead to representing something as not X. For this to happen, the system must commit a version of the headless woman fallacy. Much more needs to be said to explain why consciousness seems to be nonphysical, as opposed to not seeming to be physical.

On the whether what explains consciousness is the *attention* schema: Graziano notes that awareness correlates best with attention. But I am not sure that correlation is the decisive factor here. For example: I am inclined to think that our intuitions about visual experience arise from a *vision schema*. This is an introspective model of vision and how it works, no doubt vastly simplified but useful. Our intuitions about perceptual experience in general arise from a *perception schema*, intended as a model of perception in general. And so on.

Now, the correlation between perceptual experience and perception is not perfect, because there is also unconscious perception. The model leaves this out. The model claims to be representing all of perception all the same. It uses perceptual experience to do that. The model is in effect committed to all perception being conscious. Still, it is a (distorted) model of perception, rather than a (relatively accurate) model of conscious perception.

Still less is it a model of attentive perception. Representations of perceptual awareness are not just a specialized tool for handling attention. They are a general model of perception, tied to the quite general need for the brain to model itself. They rule out unconscious perception, but that isall in the interest of havig a usable model.So (our model of) visual awareness is really not a model of visual attention. If it is a model of anything, it is a model of vision. Similarly, I would say that (our model of) awareness in general is not a model just of attention. It is a model of representation.

Revelation and primitive quality attribution

The *revelation* thesis holds roughly that when we have a conscious experience, we are in a position to know its nature. Or more precisely, when a subject has a conscious experience (and certain further conditions obtain), the subject is in a position to know all of its essential and/or intrinsic properties. The *revelation intuition* is the intuition that the revelation thesis is true.

In the target article I brought up the revelation intuition as part of the sense of acquaintance and said that it may play a role in generating our problem intuitions, especially the intuition that consciousness is irreducible.

A number of commentators give a central role to the revelation intuition. Chris Schriner and Kenneth Williford think it can play a central role in explaining our problem intuitions, and that denying it opens the way to weak illusionism and type-B materialism. Michelle Liu thinks that the revelation intuition is best explained by its truth, and that illusionists cannot give a satisfactory explanation of the intuition. She criticizes Derk Pereboom's account of the intuition in terms of qualitative inaccuracy (itself the illusionist version of the more neutral strategy I call primitive quality attribution). Pereboom defends his qualitative inaccuracy thesis against some related criticisms.

I confess that I do not have the revelation intuition as strongly as some other intuitions. It is not obvious to me antecedently that consciousness cannot have further intrinsic or essential properties that are not revealed in introspection. It is not out of the question that it has some underlying surprising intrinsic nature. The intuition I have is that consciousness has certain distinctive properties that I *do* know about in introspection, not that it has no further distinctive properties that I do not know about. This may involve a positive partial revelation claim (I know some of the nature of consciousness), but it does not involve a negative full revelation claim (I know the full nature, so there is nothing I do not know). I think it is this positive intuition that generates the problem intuitions in me, and not the negative intuition that is built into the revelation thesis.

As a result, I am inclined to be skeptical about Williford's claim that the revelation intuition is responsible for all the problem intuitions. I do see how the revelation intuition, if accepted, *could* generate the intuitions (especially metaphysical intuitions about nonphysicality and related explanatory intuitions). But I am skeptical that it is needed to explain the intuitions and that it is the best explanation of the intuitions in us. I think partial revelation theses can do that. For example, intuitively, we have knowledge of a certain qualitative character of consciousness. This knowledge seems to go beyond what knowledge of the physical world yields, and we can conceive of this qualitative character being absent in a physicaly identical system. Nothing here requires the full revelation thesis.

The revelation thesis is perhaps more relevant to further steps such as the move from conceivability to possibility (as Williford suggests), though it is not obvious that these rest entirely on the thesis either. Perhaps it plays a role in the Kripkean thesis that appearance is the same as reality when it comes to consciousness, which blocks the appeal to a posteriori necessities. But the anti-materialist argument can also be run with more sophisticated conceivability-possibility principles (for example using the two-dimensional semantic framework) that are entirely consistent with a posteriori necessities, and with denying the appearance-reality thesis. So I think that denying the revelation thesis does not undermine all relevant anti-materialist arguments, though it may undermine some of them. In any case, here we have moved beyond what explains problem intuitions to the separate question of what we infer from them.

As Williford outlines, the revelation intuition does provide a nice potential explanation for the headless-woman move from consciousness not seeming physical to its seeming nonphysical. Again I am skeptical that this intuition is operative in my own case, but perhaps there are others in whom it operates. Like Diaz Leon, I find the intuition of distinctness somewhat weaker than some of the other intuitions. I am happy enough to use the knowledge intuition, say, as a basic premise in an argument, but I would never use the distinctness intuition as a basic premise. At least for me the distinctness intuition seems downstream from other intuitions, including explanatory intuitions and perhaps other metaphysical intuitions, including the intuition that consciousness has a robust qualitative character. But I can allow that there are others such as Liu, Schriner, and Williford in whom the revelation intuition plays a more central role.

Schriner also thinks the revelation thesis is both central and false, and uses it to support a weak illusionism. He suggests that it explains the explanatory gap: "If we lack introspective access to the constitutive ontology of conscious experiences, there is no gap to explain between qualia and material properties." But as before, full revelation is not needed for a gap. Partial revelation is enough for that. Perhaps Schriner is suggesting that the partial revelation thesis is false—he says that we have *no* access to the nature of experience. This is a stronger claim, but one worth entertaining. It is less clear how partial revelation on its own can explain the problem intuitions, though. On the face of things, it can only do this alongside certain strong claims about what seems to be revealed: nonstructural qualitative character, for example. Then I suspect that it is these claims that will be doing the work.

Schriner also gives a role to presentation and duality intuitions. He explains presentation intuitions throgh the idea that internal maps can be iconically configured, and we should expect this to yield a sense of presence. I'm not so sure. We have many maps in the brain, but few of them yield a sense of presence. So something more is needed to explain where the sense comes from. Still, all of these intuitions are worth focusing in investigating where our problem intuitions come from.

Michelle Liu focuses on the question of what explains the revelation intuition. She suggests that the best explanation is its truth: consciousness really does reveal it nature to us. She considers an alternative illusionist explanation in the spirit of Pereboom's qualitative inaccuracy thesis and argues that it does not work. This argument primarily involves a regress argument against illusionism, which I consider in “Debunking Arguments For Illusionism about Consciousness”.

Pereboom’s qualitative inaccuracy thesis is an illusionist thesis rather than a neutral solution to the meta-problem, but it corresponds to the more neutral thesis I called “primitive quality attribution” This thesis says that we attribute primitive qualities or primitive to our mental states, while the qualitative inaccuracy thesis adds that our mental states do not have these qualities.

In the target article I criticized the primitive quality attribution thesis and Pereboom’s qualitative inaccuracy thesis for not solving Kammerer’s *resistance problem*: why is illusionism so hard to accept? A similar thesis is arguably true of colors: we attribute primitive color qualities to things in the world. But color illusionism on which these qualities are not really present is easy accept. What is the difference?

Pereboom says the answer is that (i) we have no independent check on introspection, unlike perception, and (ii) illusions of phenomenal consciousness would have to be phenomenally conscious. I am skeptical about (i). I don’t need an independent check on color to seriously entertain the idea that color qualities are not present. Even prior to a check, I can quickly grasp that the view might be true, whether or not it is true. It is much harder to do that for consciousness. As for (ii), certainly one will get into trouble if one assumes that all illusions are phenomenally conscious, but it is obviously part of the illusionist view to reject that assumption. I don’t really see why it should be so hard to accept all at once that we have unconscious color representations and unconscious introspective representations, given Pereboom’s view. Pereboom suggests that this is hard to conceive or imagine, but I can certainly conceive that this is true of some other system such as a zombie. It is hard to accept that it is true of me, but I don’t think that this is because of any incoherence or unimaginability. It just seems certainly false.

I continue to think primitive quality attribution may be a central part of a solution to the meta-problem, but I think more work is needed to solve the resistance problem. An appeal to primitive relation attribution (where we attribute primitive relations to the qualities) may help, especially where the relation is something like acquaintance. This brings us back to explaining the sense of acquaintance, which is the key problem we have not yet solved.

Underestimating the physical

Some physicalists explain problem intuitions by saying not that we overestimate the phenomenal but that we underestimate the physical. Strawson and others have argued that we mistakenly focus on the physics-al, or the structural, when in fact the physical is much richer than this and includes intrinsic character that may involve or explain consciousness. In the target article I said that this may help explain problem intuitions about a physical-phenomenal gap, but does not really explain others that are not cast in terms of the physical per se. For example, it does not help explain a physicsal-phenomenal gap, a functional-phenomenal gap, or a structural-phenomenal gap. But all of these are interesting gaps that need explaining and that are subject to the meta-problem program.

Tom McClelland responds to this critique by saying that intuitions about a physical-phenomenal gap are important and argues that the ignorance hypothesis, holding that we are ignorant of the nature of the physical, can explain them.

I am not sure that McClelland and I have a major substantive disagreement here. I agree that physical-phenomenal gap intuitions are important and I agree that the ignorance hypothesis can help explain them. But I think the other intuitions are also important and also in need of explanation.

For example, the functional-phenomenal gap plays a major role in my original statement of the hard problem. There the key intuition was that explaining functions does not suffice to explain experience. Nothing here mentions the physical and underestimating the physical does not bear on explaining it. McClelland responds by saying that this intuition is only a problem intuition if one thinks that functionalizability is required for explanation in physical terms; for others, there is no problem. I think this is wrong. For example, many Russellian monists endorse an expanded physicalism and so don't see a physical-phenomenal gap, but still have the strong sense that there is a hard problem precisely because there is a structural-phenomenal gap or a functional-phenomenal gap. It is these gaps that force one to rely on the appeal to an expanded notion of the physical in the first place. So these gaps are quite central to the problem.

Of course which intuitions count as problem intuitions or part of the hard problem is a largely terminological matter. More substantively, I think structural-phenomenal gaps and physicsal-phenomenal gaps are interesting and worth addressing. Why do we think consciousness has a nature beyond the structural and the functional? Perhaps McClelland thinks this is just obvious. Still, the request for a topic-neutral explanation seems a reasonable one, and most of what I said in the target article about possible explanations still apply to this question. There remain many interesting forms of illusionism that deny these intuitions and are worth examining. Furthermore, an explanation of these intuitions will play a central role in explaining physical-phenomenal intuitions. Even if we underestimate the physical by taking it to be structural, this will only yield a gap if we take the phenomenal to be nonstructural. So explaining the latter will also be crucial.

As for the ignorance hypothesis, I am not unsympathetic with it in its Russellian form. I am sympathetic with Russellian monism, and this naturally goes along with the view that we are ignorant of the intrinsic nature of the physical. I'm not so sympathetic with non-Russellian versions. If we do not make the Russellian appeal to nonstructural properties, then we are left having to explain consciousness in terms of the structural properties in physics. Then we are faced with the structure-consciousness gap, about which the ignorance-of-the-physical hypothesis has little to say. Perhaps there is a more general ignorance-of-structure hypothesis that could help to close the structural-phenomenal gap, but that is a different thesis and making a case for it would take a lot of work.

Most of my reply to McClelland also applies to Galen Strawson's paper, which diagnoses underestimating the physical as "The Great Mistake". A more straightforward reply to Strawson is also available. The issue about whether physicalism should be understood as physics-alism or some different thesis is almost largely verbal, as is the issue of whether the physical should be understood as the physics-al. Furthermore, at least where physicalism is concerned, many of the verbal facts are on the side of physics-alism. Carnap and the other logical empiricists introduced "physicalism" in the 1930s precisely for a physics-alist thesis, about the primacy of physics in metaphysics. The existing term "materialism" was available but too imprecise, and they wanted one tied to physics. Likewise, a great many of the people who are interested in the current debate are interested precisely in physics-alism: can physics explain everything? And physics-alism itself is certainly a substantive and important thesis, whatever one thinks of its truth. So there is a strong case that "physicalism" should be understood as "physics-alism".

If this terminological move is accepted, Strawson's opposition to the usual framing simply drops away. He agrees with all the other nonreductionism that physics-alism is false, and his positive view looks like that of many other Russellian panpsychists. Of course Strawson is unlikely to accept this terminological move (to be fair, I don't think the verbal facts are cut and dried). But even if he doesn't, the discussion makes clear that Strawson's vigorous disagreement with the terms of the debate is largely verbal. The fact that some people in this debate use "physical" to mean "physics-al" makes no difference at all to the substantive issues.

Either way, the meta-problem remains an interesting question. Why do we think that consciousness is not physics-al? Can we explain that intuition in topic-neutral terms? Oddly, Strawson does not begin address this question. The fact that we say we are conscious and that this can't be physics-ally explained (or the fact that we make corresponding noises and inscriptions, at least) is a behavioral fact about us, and one might think it is open to physics-al or topic-neutral explanation, say in computational terms. It would be good to know whether Strawson agrees and if so whether he has any views about what the topic-neutral explanation might be.

The access problem

Daniel Dennett's commentary is almost entirely devoted to the access problem: how is it that some information becomes accessible at the personal level for action, reasoning, and report? This is an extremely interesting problem, but it is not the meta-problem. Explaining how "That object is red" or even "I see a red object" is made accessible does not yet explain why perception of redness should seem to have the distinctive properties of consciousness. Of course there may be a connection. As Dennett says, it could be that a solution to the access problem might uncover mechanisms that will also solve the meta-problem. But Dennett does not even try to make that case here. His brief ensuing discussion of deja vu doesn't try to explain any problem intuitions; instead he argues that the intuitions are incorrect, which is a different project.

I think the strategy of using access to help explain problem intuitions is worth trying. One might see Kammerer's treatment of the evidential strategy as an instance of it: passive evidence needs to be made accessible (to help in reasoning and communication), and the mechanism used for this (the evidence-resemblance strategy) also generates our problem intuitions. I don't think Kammerer's solution works, but it is the sort of idea we should be exploring. It would be good to hear similarly substantial hypotheses along these lines from Dennett.

Dennett also suggests that there need not be a single solution to the meta-problem. There may be different mechanisms that generate many problem intuitions. Perhaps, but given the many commonalities between intuitions in different domains, I would be surprised if the explanation were too disunified. In any case, I am not sure that any of the mechanisms in Dennett's account so far explain any central problem intuitions. And certainly there are many core problem intuitions that are explained by no part of Dennett's account. Eventually we need to explain these, whether by one mechanism or many.

In any case, I agree with Dennett that the access problem is important in its own right. It's an interesting hypothesis that much of our personal-level access to mental states evolved to serve human communication and reflective reasoning, and that nonhuman animals don't have anything really analogous. It's also plausible that access is at least necessary to generate problem intuitions, even if it's not sufficient. Putting those things together, it's not out of the question that nonhuman animals not only lack problem intuitions (which is no surprise), but don't have the cognitive architecture that produces them. Of course if one is a realist about consciouness, that doesn't entail that animals are not conscious. It may well be that the architecture required to subserve access for reflection and report goes well beyond the architecture required for consciousness, and that animals still have the latter (perhaps involving access for simple reasoning and action, if not for reflection and report). But at least this would require some hard thinking, in the spirit of the meta-problem challenge, about how these architectures are related and and about what function consciousness serves in non-human animals.

Dual systems and agency detection

A number of commentators try to explain problem intuitions in terms of dual brain systems or dual modes of thinking. The idea is that there is one system or mode for thinking about physical processes and another for thinking about mind or consciousness.

In two commentaries, the idea is combined with an appeal to the idea that we have special systems or processes for agency detection. Brian Fiala and Shaun Nichols develop their earlier work on this topic (Arico et al 2011), and Gary Drescher pursues a related version of the idea. The key idea is that we have systems in our brain that detect and categorize entities as agents. It is activated upon seeing a conscious person, but it is not activated upon seeing a neural process. This explains our sense that neural processes are not adequate to explain consciousness.

This explanation is intriguing, but I am not sure it is a good match for many problem intuitions. In the zombie case, for example, we conceive of a being who looks and behaves just like an ordinary conscious being. This will surely trigger agency detection, and I may well be inclined to judge that such a creature is conscious. Nonetheless, despite the strength of the agency detection, we can nevertheless conceive that they are not conscious, with all the behavior and the underlying processes the same. I can even have a zombie intuition while talking to someone: I'm sure they are conscious, but I can nevertheless conceive that they are not. It is very hard to see how agency detection explains that. We need not particularly be focusing on neural processes, which play the prime role in the story. We may well be focusing on appearances and on behavior, which trigger agency detection. Even so, the gap between these things and consciousness remains.

The agency detection story is somewhat better suited for addressing explanatory gaps between neural processes and consciousness, and the rejection of brain-consciousness identities. But I worry that it overgeneralizes. When we explain Y in terms of X, we don't typically expect X to trigger the concepts involved in Y. Descriptions of H_2O don't trigger our water concepts. Nevertheless, we don't reject water-H_2O identities. We're used to the idea that explanations take some work. We don't expect the explanans to immediate trigger concepts of the explanandum. So it would seem a bad error for us to suddenly start doing this with consciousness. Perhaps Fiala and Nichols will say that the concept of consciousness is primitive but the concept of water is not. But it is still unclear why we should suddenly impose this high standard on explanations involving primitive concepts.

As well as the agency detection story, Drescher offers an explanation of problem intuitions in terms of meta-observation or higher-order representation. He thinks that higher-order monitoring of mental states makes them conscious, so that every state we introspect is conscious and seems to be conscious. Then a refrigerator-light fallacy makes us think that many more states are conscious, including those we don't monitor. Perhaps this could help explain intuitions about the richness of consciousness, and also why the higher-order theory seems to be false. It does not get us a long way toward explaining the core problem intuitions, though. It is also unclear why higher-order monitoring should make the state it monitors seem to be conscious. On standard higher-order account this requires a third-order state to monitor the second-order monitoring and thereby reveal that the state is conscious. Of course there is room for nonstandard pictures but it would be interesting to see one spelled out.

Storm offers a different sort of dual-systems account: we have one system for understanding physical phenomena and a separate system for understanding mental phenomena. This system is subject to different rules and its concepts play different roles. This gives rise to a kind of dualism. I think something like this may well be part of the story, but again I worry that the story overgeneralizes. Everything Storm says applies equally to belief, so as he acknowledges, he cannot explain the difference in strength of problem intuitions between experience and belief. There also seem to be specialized brain systems for various other domains that do not generate the same explanatory gaps. So something special must be going on with the consciousness system, and part of the challenge is to explain what.

Haoying Liu discusses a "use-mention" strategy closely related to the two-systems strategy and the agency detection strategy. Where the agency detection strategy rests on agency detection being activated in one mode of thought but not another, this strategy rests on experience being activated in one mode of thought but not another. When we think about consciousness in the first-person way, we have certain experiences. This creates the impression that to refer to consciousness we must have these experiences. When we think about brains and the like in the third-person way, we do not have these experiences. So we infer that we are not thinking about consciousness at all.

Liu thinks this should not be called a use-mention fallacy. Nothing turns on the name, but the inference from “this thought does not involve an experience” to “this thought does not refer to an experience” at least involves something like a use-mention inference. In any case, it looks like a bad inference. There’s not in general reason to think that for a thought to refer to X, it has to use X. I take it the idea is that we have at least some basis for thinking that when X is consciousness, since ordinary thoughts that refer to consciousness use consciousness. Still, it would be a big leap for us to suppose there is no other way. Our ‘water’ thoughts may typically involve images of water and our H_2O-thoughts do not, but this does not lead us to reject the water-H_2O identity. So as with the agency detection strategy, this strategy leaves it unclear why this sort of bad inference should so strong in the consciousness case.

Miscellaneous

Two miscellanous strategies: Nicholas Humphrey appeals to his distinctive story about the evolution of sensation, and Colin Klein and Andrew Barron appeal to the apparent arbitrariness of brain-consciousness relationships and our lack of control over them.

Humphrey outlines his theory of the evolution of sensory consciousness, and along the way says a couple of things about the meta-problem. He thinks that there is a trajectory of brain evolution that leads toward an “ipsundrum” attractor state that he thinks makes consciousness seem weird and wonderful. He also suggests that in an evolutionary context, thinking of ourselves as conscious in this mysterious way makes ourselves seem more significant, which leads us to place more value on our and others lives. This is an interesting idea, though it would be nice to see a mechanism spelled out that makes clear exactly how these special states enhance fitness, in a context where most creature don’t seem to have too much trouble placing a high value on their own lives to start with.

Klein and Barron pin problem intuitions on the fact that brain-consciousness relationships seem arbitrary and out of our control. They suggest that if we had an autocerebroscope that we could use to control our brain states and thereby control our consciousness, we would not be impressed by the hard problem. For what it’s worth, I think the hard problem would persist about as strongly as ever. The autocerebroscope may bring out a useful and systematic dependence of consciousness on the brain. But this dependence is equally compatible with a dualist, who can equally use the autocerebroscope. Nothing in this systematic dependence explains why consciousness exists in the first place. To be sure, such a dependence will be central to the science of consciousness and will certainly be very useful in some sorts of explanation. But it is hard to see how it does much to deflate the core problem intuitions.

Summary

At the start of this article I outlined my own preferred approach to the meta-problem. Where have the contributions to this symposium pushed things forward?

Of the many interesting contributions in the commentaries, I think the Kammerer/Schwartz evidential strategy is intriguing, especially in explaining our need for access to certain special evidence states. This might play a role in explaining the sense of immediate knowledge with special qualities, but more work is needed to see how it explains the sense of acquaintance and various core problem intuitions.

Diaz's contribution makes a strong case that the phenomenal concept strategy (in the form that appeals to cognitive isolation or independent roles) can explain negative problem intuitions, whether or not it can explain their truth. To explain positive problem intuitions, though, it needs something more. Perhaps an appeal to acquaintance or primitive quality and relation attribution (and/or revelation and presentation) can help explain the positive problem intuitions—but then much of the work is in properly explaining the sense of acquaintance and the like.

Of the other strategies, there is something generally right about the attention schema appeal to introspective models and the dual systems appeal to different modes of thinking, but I don't think the specific appeal to introspective opacity or agency detection get us far. The other strategies are all interesting in various respects but my view is that they do not ultimately change the core issues here.

So, the residual puzzle as I see it lies especially in explaining the positive problem intuitions, and especially in explaining the sense of acquaintance along with primitive quality and relation attribution. I think these should be topic-neutrally explainable in principle, but I haven't seen a good explanation yet. I look forward to seeing further work that explains them.

References

Arico, A., Fiala, B., Goldberg, R.F. & Nichols, S. 2011. The Folk Psychology of Consciousness. *Mind and Language* 26 (3):327-352.

Chalmers, D.J. 2018. The meta-problem of consciousness. *Journal of Consciousness Studies*.

Clark, A. ; Friston, K. & Wilkinson, S. 2019. Bayesing Qualia: Consciousness as Inference, Not Raw Datum. *Journal of Consciousness Studies* 26 (9-10):19-33.

Dennett, D. C. 2019. Welcome to Strong Illusionism. *Journal of Consciousness Studies* 26 (9-10):48-58.

Dewhurst, J. & DoÅga, K. 2019. Attending to the Illusion of Consciousness. *Journal of Consciousness Studies*.

Diaz-Len, E. 2020. The meta-problem of consciousness and the phenomenal concept strategy. *Journal of Consciousness Studies*.

Fiala, B. & Nichols, S. 2019. Generating Explanatory Gaps. *Journal of Consciousness Studies* 26 (9-10):71-82.

Graziano, M. S. A. 2019. We Are Machines That Claim to Be Conscious. *Journal of Consciousness Studies* 26 (9-10):95-104.

Humphrey, N. 2019. Easy Does It: A Soft Landing for Consciousness. *Journal of Consciousness Studies* 26 (9-10):105-114.

Kammerer, Franois 2019. The Meta-Problem of Consciousness and the Evidential Approach. *Journal of Consciousness Studies* 26 (9-10):124-135.

Klein, C. & Barron, A. 2020. First-person interventions and the meta-problem of consciousness. *Journal of Consciousness Studies*.

Liu, H. 2020. On Chalmers on the meta-problem. *Journal of Consciousness Studies*.

Liu, M. 2020. Explaining the intuition of revelation. *Journal of Consciousness Studies*.

McClelland, T. 2020. Ignorance and the meta-problem of consciousness. *Journal of Consciousness Studies*.

Papineau, D. 2019. Response to Chalmers' 'The Meta-Problem of Consciousness'. *Journal of Consciousness Studies* 26 (9-10):173-181.

Pereboom, D. 2019. Russellian Monism, Introspective Inaccuracy, and the Illusion Meta-Problem of Consciousness. *Journal of Consciousness Studies* 26 (9-10):182-193.

Schriner, C. 2020. Illusionism helps realism confront the meta-problem. *Journal of Consciousness Studies*.

Schwarz, W. 2019. From Sensor Variables to Phenomenal Facts. *Journal of Consciousness Studies* 26 (9-10):217-227.

Storm, J. 2020. Why does the brain-mind (consciousness) problem seem so hard? Reflections on our mental limitations and dualistic intuitions: neuroskepticism/neurocomplementarity. *Journal of Consciousness Studies*.

Strawson, G. 2019. Underestimating the Physical. *Journal of Consciousness Studies* 26 (9-10):228-240. Schwarz, W. 2018 Imaginary foundations.

Sturgeon, Scott (1994). The Epistemic View of Subjectivity. *Journal of Philosophy* 91 (5):221-235.

Williford, K. 2020. Headlessness without illusions: Phenomenological undecidability and materialism. *Journal of Consciousness Studies*.