Debunking Arguments for Illusionism about Consciousness*

David J. Chalmers

Why do people believe that God exists? One sort of answer is in evolutionary or psychological terms. One hypothesis is that it enhanced evolutionary fitness to believe in a god, perhaps because this belief gave special motivation for action. Another is that belief in God helps to fulfill a deepseated need for meaning in our life. If we accept a hypothesis like this, many people think that our justification for believing in God falls away. After all, if our beliefs in God can wholly be explained without God, then it looks as if it would be a giant coincidence if there was actually a God that made those beliefs correct.

Why do people believe that certain actions are right and other actions are wrong? Once again, one sort of answer is in evolutionary or psychological terms. One hypothesis is that believing certain actions were right and wrong enhanced evolutionary fitness by increasing our chances of survival. If we accept a hypothesis like this, some philosophers think that our justification for believing in right and wrong falls away. Others think that as a result we should at least reject theses such as moral realism, on which there are moral facts independent of our judgments.[1] If our beliefs about right and wrong can wholly be explained without appealing to right and wrong, then it would be a giant coincidence if there were actually mind-independent facts about right and wrong that make the beliefs correct.[2]

These arguments are *genealogical debunking arguments*. They start from a premise about the genealogy of our beliefs about a domain: that is, about how those beliefs are formed. They move from here to a conclusion that debunks those beliefs in some fashion: perhaps saying that the beliefs are unjustified, or that they do not reliably track the truth, or at least that they do not reveal the truth about a mind-independent world. In recent years, debunking arguments have been especially popular in the domain of morality, but they have also been made in domains including

[0]This article is one of three responding to 39 commentaries in a symposium on my article "The Meta-Problem of Consciousness" in the *Journal of Consciousness Studies*. It is written to stand alone. I'm grateful to all the commentators for their interesting contributions, to François Kammerer for his editorial work, and to Dan Korman for feedback.

[2]See Joyce 2006 and Street 2006 for arguments of these two sorts.

religion, metaphysics, mathematics, and many more (Korman 2019 has a nice survey).

One can also make debunking arguments in the philosophy of mind. For example, if we can explain our belief that we are conscious in terms that do not require the existence of consciousness, then perhaps our belief that we are conscious will be debunked. This sort of debunking argument can be used as an argument for *strong illusionism* about consciousness: the thesis that nobody is phenomenally conscious, and that our beliefs that we are conscious involves some sort of illusion.

Similarly, if we can explain why we think consciousness is irreducible in terms that do not require anything irreducible, perhaps our belief in the irreducibility of consciousness will be debunked. If we can explain our belief that there is a hard problem of consciousness wholly in terms that do not require any problematic features of consciousness, perhaps our belief that there is a substantive hard problem of consciousness will be debunked. This sort of debunking argument can be used as an argument for *weak illusionism* about consciousness: the thesis that although we are conscious, we are wrong about some of its features and our beliefs about these features (perhaps especially those tied to its irreducibility or its problematic status) involve some sort of illusion.

All this brings out one key role for a solution to the meta-problem of consciousness: it can serve as the key premise in a debunking argument for illusionism about consciousness.[3] The meta-problem is the problem of explaining why we make the judgments that we do about consciousness, and especially why we judge that consciousness poses a problem. There is at least some reason to think that we should be able to explain these judgments in topic-neutral terms—that is, without appealing to consciousness in the explanation. If such a topic-neutral explanation is possible, we can use it to argue that our judgments about consciousness are unjustified or do not track the truth, so that we should endorse some sort of illusionism.

This brings out the exciting possibility that a solution to the meta-problem of consciousness might dissolve the hard problem of consciousness. If we can explain our judgments about why we think there is a problem about consciousness in terms that never mention consciousness, then this will leave little reason to believe that there is a real problem. More generally, if we can explain our judgments about consciousness in terms independent of consciousness, then there will be little reason to think those judgments are correct and perhaps little reason to believe in consciousness itself.

It is easy to see some sort of debunking reasoning as tacitly playing a role in the recent popu-

[3]The title should be read as a noun phrase along these lines, though other readings are available.

larity of illusionism. But debunking arguments have rarely been made explicit. I think they should be made explicit, as in my view debunking arguments are the strongest arguments for illusionism and they need to be assessed. I am not an illusionist about consciousness, and I do not think debunking arguments for illusionism succeed. But I think they are important arguments that need to be answered (if not debunked), and I think that answering them may well give us insights into the problem of consciousness that are interesting in their own right.

In the later parts of "The Meta-Problem of Consciousness", I presented two debunking arguments for illusionism and tried to at least briefly assess them. I concluded that some replies are available, but that more work needs to be done. I also discussed the pros and cons of weak and strong illusionism in their own right, providing some arguments against each.

In the current symposium, a number of commentaries focus on these debunking arguments, on possible replies, and on associated issues about illusionism. Justin Clarke-Doane and Adam Pautz focus on the nature and force of debunking arguments regarding consciousness. Hedda Hassel Mørch and Bradford Saad discuss what I called realizationism, a realist strategy which resists the arguments by holding that consciousness realizes the processes responsible for our judgments, so that explanations of these judgments are not really independent of consciousness. Katalin Balog, Joseph Levine, and Keith Frankish focus on the role of acquaintance in replying to the arguments and in approaching the meta-problem more generally. Giovanni Merlo and Michelle Liu argue against illusionism, and a number of others discuss different aspects of it. In what follows I will proceed through each of these issues.

Debunking arguments

Illusionist strategies regarding the mind–body problem have a long history, some of which I outlined briefly in the target article. These strategies often involve a tacit debunking argument: our intuitions about consciousness are explained away using some psychological mechanism, and as a result the intuition is supposed to be debunked. The debunking reasoning is rarely made explicit and general, though.

One fairly explicit and general debunking argument in the recent literature on consciousness was made by Sydney Shoemaker in his 1975 article "Functionalism and Qualia". Shoemaker is arguing that "absent qualia" cases—functional duplicates of conscious humans who lack phenomenal states entirely—are possible. Ned Block (1980) summarizes Shoemaker's argument as follows:

> "[I]f absent qualia are possible, then the presence or absence of the qualitative character of pain would make no difference to its causal consequences; and so, according to a causal theory of knowledge, we could have no knowledge of the qualitative character of pain; but given that we do have knowledge of the qualitative character of pain absent qualia are not possible."

Shoemaker's argument is structurally analogous to a well-known mathematical debunking argument by Paul Benacerraf: if numbers are abstract objects, then they make no difference to our beliefs, so (by a causal theory of knowledge) we could not know about numbers; but we do know about numbers, so numbers are not abstract objects. Shoemaker's and Benacerraf's arguments are not arguments against the *existence* of qualia and of numbers, but they are arguments against theses saying that they are irreducible in certain ways.

In "The Conscious Mind" (1996) and "The Content and Epistemology of Phenomenal Belief" (2003), I developed and replied to two Shoemaker-inspired arguments, here deployed as arguments against the possibility of zombies. The 1996 version focused on the causal theory of knowledge. I replied by denying the causal theory of knowledge, holding that we have knowledge by acquaintance of consciousness, which is stronger than knowledge by causation and does not require a causal connection. The 2003 version focused more explicitly on the idea that zombies would have the same beliefs about consciousness that we do. I responded by denying this claim: there is a constitutive connection (again stronger than any causal connection) between consciousness and our core beliefs about consciousness, so zombies do not have these beliefs.

In the target article I presented two somewhat more general debunking arguments that support (directly or indirectly) the stronger conclusion of illusionism about consciousness.[4] (Here and in what follows, when I say illusionism alone I will always mean strong illusionism.)

> 1. There is a correct explanation of our beliefs about consciousness that is independent of consciousness.
>
> 2. If there is a correct explanation of our beliefs about consciousness that is independent of consciousness, those beliefs are not justified.

[4] I also presented a third argument briefly and enthymatically: 1. If there is a reductionist explanation of our nonreductionist beliefs about consciousness, those nonreductionist beliefs are not justified. 2. There is a reductionist explanation of our nonreductionist beliefs about consciousness. Therefore 3. Our nonreductionist beliefs about consciousness are not justified. This argument has a more restricted scope than the argument from justification below, applying to a class of problem beliefs rather than phenomenal beliefs in general. It raises a few distinctive issues, but the range of replies available to both arguments overlap to a considerable extent.

3. Our beliefs about consciousness are not justified.

The conclusion is not exactly a statement of illusionism, but it tends to support illusionism: if our beliefs about consciousness are not justified, then we have no justification for believing we are conscious and as a result we should probably accept illusionism. [5]

I suggested that this argument could be resisted in a number of ways. First, there is an issue about just how "independent" should be understood to make both premises plausible. I suggested that the standard topic-neutrality claim of descriptive independence (the explanation doesn't mention consciousness) is not enough to make premise 2 plausible: there may be physical or computational explanations of our justified table-beliefs that never mention tables. A weak version of modal independence (local elements of the explanation could be present without consciousness) is not enough either: the local brain processes that explain our table beliefs can be present without tables, for example in a case of hallucination. Instead, something like causal independence (consciousness plays no causal role in the processes invoked in the explanation) and constitutive independence (consciousness does not constitute and is not constituted by elements of the explanation) is required.

Once independence is understood this way, the door is open to denying the key premises. For a start, reductionists about consciousness will happily deny the causal and constitutive independence premises, so (not unexpectedly) the argument is more of a problem for nonreductionists. But even for nonreductionists, at least three replies are available. The first two correspond to my responses to debunking arguments in the 1996 book and 2003 article. One can deny premise 1 by arguing that consciousness plays a role in constituting beliefs about consciousness. One can deny premise 2 by saying we have a special acquaintance with consciousness that justifies our beliefs regardless of causal history. Finally, one can also deny premise 1 with an appeal to realizationism, on which consciousness realizes the processes leading to judgments about consciousness, thereby yielding a causal dependence of the judgments on consciousness.

I have come to think that while the appeals to acquaintance and constitution go some way to

[5] Korman (2020) divides debunking arguments into *conditional* debunking arguments (used to argue that if certain theories of X are true, our beliefs about X are problematic) and *skeptical* debunking arguments (used to argue that our beliefs about X) are poblematic. Where the Shoemaker arguments and my earlier Shoemaker-inspired arguments are conditional debunking arguments, the current argument has the form of a skeptical debunking argument. However, once independence is unpacked appropriately (say as modal or expanatory independence), only certain theories will accept the independence premise, so this argument becomes effectively equivalent to a conditional debunking argument.

blocking debunking arguments, they do not go all the way. One can get around the reply from acquaintance by moving from an argument about justification to an argument directly about coincidence. One can get around the reply from constitution by moving from beliefs to "intuitions", which are stipulated here to be characterized in functional terms. What results is the second argument:

1. There is an explanation of our phenomenal intuitions that is independent of consciousness.

2. If there is an explanation of our phenomenal intuitions that is independent of consciousness, and our phenomenal intuitions are correct, their correctness is a coincidence.

3. If our phenomenal intuitions are correct, their correctness is not a coincidence.

———

4. Our phenomenal intuitions are not correct.

Here the premises all look somewhat plausible if there is a solution to the meta-problem. Of course one can still question premise 1 with an appeal to realizationism, and one could question premises 2 or 3 by arguing about just what is required for problematic coincidence. But in each case there is at least a prima facie case for the premises.

Regarding premise 1: The realizationist can reasonably deny causal and constititive independence as described above: on the realizationist view, consciousness plays a causal role in the formation of phenomenal intuitions in the actual world. But there remains a case for a strong sort of modal/explanatory independence. In particular, there remains a near-complete structural explanation of the intuitions (in computational terms, say), and this explanation can obtain without consciousness. This is a stronger sort of modal independence than the local modal independence discussed earlier (which was rebutted with the case of table hallucinations): it involves reproducing the global structure of a near-complete causal explanation of the intuitions rather than just local or incomplete elements, as in the table case.

Regarding premise 2: If there is a near-complete structural explanation of the intuitions that could obtain without consciousness, this gives a strong sense that if the intuitions are correct, their correctness is something of a coincidence. If our intuitions can be explained algorithmically, for example, it seems that they could easy have been realized without consciousness, so that it is a lucky coincidence that they are actually realized by consciousness in a way that makes them

correct. So the sort of modal/explanatory independence just outlined yields a case for coincidence that needs to be addressed.

Regarding premise 3: Why is coincidence problematic? One reason is that it may undermine justification. Invoking this reason in effect turns the coincidence argument into a justification-involving debunking argument, which may be hard to square with the appeal to intuition (arguably judgments are objects of justification but intuitions are not) and which will also be subject to the reply from acquaintance above (acquaintance may justify even when a phenomenal belief is true by coincidence). However, coincidence is also problematic quite independently of its role in justifying ordinary beliefs. In general it is a theoretical vice for a theory to postulate massive coincidence (such as the existence of phenomenal states that are entirely independent of an algorithmic explanation of our phenomenal intuitions but that make them correct), as it is antecedently implausible that there should be massive coincidences in the world. Of course it is not logically impossible that there are massive coincidences, but a theory that avoids them is much preferable.[6]

Of course the case for the premises is merely prima facie. It may well be possible for a realist view such as realizationism to deny the independence claims or the coincidence reasoning, but there is at least more work to be done.

Some of the work is done by Mørch and Saad, who pursue the realizationist program further to rebut the case for explanatory independence and for coincidence. First, however, I will address Clarke-Doane's and Pautz's contributions about the nature and force of the arguments themselves.

The nature and force of debunking arguments

Clarke-Doane has written extensively on debunking arguments regarding morality and mathematics. Here he discusses the relationship between these arguments and debunking arguments about consciousness. He is mainly concerned to establish a sort of parity between them. His main target is someone who thinks moral and mathematical debunking arguments fail (for certain reasons) but that consciousness debunking arguments succeed. Of course refuting that target doesn't entail that consciousness debunking arguments fail, but it might help resist the arguments by showing that consciousness is in good company.

For what it's worth, I think that consciousness debunking arguments are weaker in some respects than moral or mathematical debunking arguments and stronger in others. On one hand,

[6]Because it does not argue for an epistemic flaw in our beliefs about consciousness, the coincidence argument is neither a skeptical nor a conditional debunking argument in Korman's sense. Rather, non-epistemic coincidence arguments are a third sort of debunking argument. Field (1989) presents related coincidence arguments about mathematics.

they have a harder hill to climb, because illusionism about consciousness seems antecedently less plausible than anti-realism about morality or nominalism about mathematical objects. Relatedly, replies to debunking arguments that appeal to acquaintance perhaps have a special strength in the case of consciousness. On the other hand, there is arguably a stronger sort of modal independence in the case of consciousness than the other domains. The key claims in the moral and mathematical domains seem to be necessary and a priori, and there is less of a case that analogs of zombie worlds are conceivable or possible. In this respect I may fall into one part of Clarke-Doane's target area, by holding that debunking arguments from modal independence get more of a grip for consciousness than in the other cases.

Clarke-Doane starts by outlining two debunking arguments about consciousness,[7] and by questioning analogous debunking arguments about mathematics. His core strategy is to argue that mathematical beliefs are reliably true (safe, sensitive, and probable) and this is enough to save them from debunking. By analogy, he suggests that the reliability of phenomenal beliefs (e.g. if I hadn't been in pain, I would have reacted differently and I wouldn't have believed that I was in pain) may save them from debunking.[8]

I am not sure I agree about either domain. Under epiphenomenalism about consciousness, beliefs about consciousness may be reliably true but more fine-grained debunking worries remain. One worry is that what explains the judgments is wholly distinct from what explains their truth. Another worry (as Clarke-Doane notes) arises from the conceivability and metaphysical possibility of zombie worlds where I have the same belief that I am in pain, caused exactly the same way, without being in pain. These strong sorts of explanatory and modal independence threaten to undermine the epistemic status of my belief that I am in pain, as well as creating the worry that we are lucky to live in a world where the psychophysical laws make our judgments true. On my view, reliability is too coarse-grained a tool to get at the causal, constitutive, modal and especially explanatory issues that lie at the heart of the issues about justification and coincidence here.

Clarke-Doane responds to the modal independence worry about zombie worlds by questioning why conceivability or metaphysical possibility should be the modality that matters most for

[8]Clarke-Doane considers consciousness debunking arguments about consciousness somewhat different from mine, with "Any explanation" where I have "There is an explanation" and using knowledge (not just truth) of independence to undermine the justification of phenomenal beliefs. I'm happy with my versions. Premise 2 (which Clarke-Doane in effect questions) follows from plausible background premises saying that justification of phenomenal beliefs requires that there is an explanation of these beliefs in terms of consciousness, that there are not two independent explanations of the beliefs, and that an explanation in terms of consciousness and one independent of consciousness will be independent in the relevant sense.

debunking. He argues that there other modalities in which it is possible that the mathematical or moral truths are different (formal logical possibility, for example) and it is not clear why this modality is any less relevant to debunking.

In response: I don't take a stand on whether debunking arguments go through for morality and mathematics, but I think conceivability and possibility matter a lot, especially where coincidence is concerned. If the worlds where not-P can be ruled out a priori, then it's not very lucky that P. If not, it's at least somewhat lucky that P. Maybe there is some sense in which P can be a priori and still lucky, but this is a highly attenuated and not very problematic sort of luckiness. I'm inclined to say something similar for metaphysical possibility (perhaps with special treatment for the contingent a priori and necessary a posteriori). I don't have clear intuitions about Clarke-Doane's notion of logical possibility, but I think that if P is a priori but not logically necessary then the truth of P is not problematically lucky. If so the logical possibility of mathematical falsehoods doesn't do much to debunk mathematical truths.

Now, perhaps one could argue for symmetry between all three modal notions here, for example holding that necessity in any of these senses suffices for nonluck and that possibility in all three senses is required for luck. But that will still block the mathematical debunking argument while allowing the consciousness debunking argument to go through. So parity between these modal notions isn't enough to establish Clarke-Doane's parity between the arguments.

Of course there can still be epistemic worries about necessary and a priori mathematical truths, as Clarke-Doane notes, and perhaps these worries may go along with certain sorts of luck. For my purposes, I don't need the claim that apriority and necessity rule out any sort of luck. What matters more for the consciousness debunking argument is that non-apriority and metaphysical possibility of zombies and the like entails certain sorts of luck. Again, Clarke-Doane hasn't made a case against that.

Clarke-Doane ends by using his modal pluralism to cast doubts on modal arguments for dualism. The idea is that there may be modalities for which zombies are possible (and for which conceivability entails possibility), and modalities for which the possibility of zombies entails the falsity of physicalism, but we don't have reason to believe in a modality that satisfies both constraints. Now, I've argued at length that metaphysical possibility satisfies both constraints. Clarke-Doane says that once we acknowledge that it is one modality among many, then even if we grant that conceivability entails metaphysical possibility, we no longer have reason to accept the second constraint, because necessity of identity may fail. I don't really see this point. Kripke's argument for the metaphysical necessity of identity goes through as well as ever, even if we allow there are

other modalities where necessity of identity fails. And the arguments that conceivability entails metaphysical possibility are using the same notion of metaphysically necessary. Perhaps someone could make a case for a subtle equivocation—but that case doesn't follow from modal pluralism alone. So more is needed to block the standard arguments.

Adam Pautz is especally interested in the coincidence argument. He develops his own version of a coincidence problem, the "normative harmony" problem. Roughly, normative harmony is the fact that conscious states are associated with causal profiles that are harmonious with their normative profiles. For example, pain makes it rational to avoid stimuli, and tends to lead to one avoiding stimuli. Pautz's problem is that if consciousness is distinct from physical-functional states which are causally closed, normative harmony is a coincidence. It seems that psychophysical laws could easily have been disharmonious, as with pain-pleasure inversions where excruciating pain causes a desire for more.

Pautz argues that normative harmony poses a coincidence problem for dualists and some materialists, although identity physicalism may avoid the problem. I am not sure about the last part. Chalmers and Jackson (2001) argued that a posteriori identities can call for explanation. Harmonious identities strike me as potentially in need of explanation, and therefore as potential objects of luck.

My coincidence problem and Pautz's harmony problem are distinct although related. My problem does not turn on normative roles for conscious states, but just on their existence. The potential coincidence that needs explaining is that our phenomenal judgments are correct. There is a related normative harmony problem: how is it that an experience E is associated with a disposition to judge that we have E, a judgment that it also rationalizes. A solution to this harmony problem might be used to solve the coincidence problem, but the reverse is not obviously the case. A realizationist solution to the coincidence problem might say nothing about normative roles, for example.

Pautz thinks that a debunking argument from modal independence to coincidence does not go through. He makes this case using a "deflationary pluralist" identity theory that supports Twin Earth cases: e.g. a biological creature is conscious and a silicon creature is twin-conscious, where twin-consciousness is just as good as consciousness but is not consciousness. So the same meta-problem processes are possible without consciousness, but there is no sense of luck. Now, I'm opposed to deflationary pluralism and I think that if twin-consciousness is as good as consciousness in all respects it is consciousness. But even if deflationary pluralism is correct, I think there's a stronger sort of modal independence in the zombie case. In the silicon twins, the counterparts of

phenomenal beliefs/intuitions are true, while in zombies they are not true (they're false or at least meaningless). So the truth of these beliefs/intuitions is much more of a coincidence if zombies are possible than if twins are possible. This requires a tweak to the definition of modal independence (what matters is the possible falsity of the beliefs, not the possible absence of consciousness), but it is a reasonable one.

Realizationism as a response to debunking

Hedda Hassel Mørch and Bradford Saad discuss realizationism as a response to the debunking arguments. The idea is that although there is a topic-neutral explanation of our phenomenal intuitions in structural terms, these structurally characterized processes are in fact realized by consciousness. As a result, our judgments are not causally independent of consciousness. Still, there remains the worry that the judgments are modally and explanatorily independent of consciousness in a strong sense, which raises the worry that the correctness of the judgments is a coincidence.

Mørch and Saad both offer versions of realizationism that resist the coincidence debunking argument. I had mentioned Mørch's phenomenal powers view in the target article as perhaps the most promising version of realizationism to play this role. On a strong version of the phenomenal powers view, phenomenal states have certain causal powers essentially and a priori, and nothing else has those causal powers. For example, pain essentially carries the power to make subjects try to avoid it. It may be inconceivable that pain lacks this causal power or has a distinct causal power. This doesn't mean that pain is reducible to this causal power (as some analytic functionalists might say). It still has a phenomenal nature that goes beyond the causal power. But it means that phenomenal states can provide a particularly strong explanation of associated causal roles.

How does this strong phenomenal powers view affect the debunking argument? The idea is that any topic-neutral explanation of phenomenal intuitions will be grounded in phenomenal powers, where no nonphenomenal powers could have grounded it as well. That suggests that the correctness of the intuitions is not a coincidence. In more detail: if the topic-neutral explanation includes or entails claims about powers, this view entails that we could not have had those powers without consciousness, so modal independence is false. If it does not, then perhaps there are regularity-based or law-based worlds where the structural explanation holds with no powers and no consciousness, so that the structural explanation is modally independent of consciousness. But it is at least arguable that a world with phenomenal powers gives a better and more parsimonious explanation. The phenomenal properties explain the powers and thereby the regularities and

the laws, whereas alternative views need to postulate both underlying properties and regularities and laws. If so, then despite modal independence, the presence of consciousness underlying the structure is not a coincidence.

Mørch's phenomenal powers view is somewhat weaker than this. She does not claim that phenomenal states have their powers essentially. She thinks it is at least conceivable that they have no powers. But she thinks it is necessary and a priori that if they have powers, they have the specific powers they do. This gap raises the worry that maybe the topic-neutral structure could exist without causal powers, as in the regularity-based or law-based worlds above. But as above, it is arguable that an explanation in terms of causal powers is better and more parsimonious.

Also, Mørch does not claim that no non-phenomenal states could have had the same causal powers. If something else could have had exactly the same causal powers, then it will be contingent that phenomenal states play the roles they do in our world and plausibly it will be contingent that our phenomenal intuitions are correct. Perhaps this engenders a small amount of luck, with the amount depending on how many carriers of the powers are possible. To avoid this problem, Mørch canvases an extended view (akin to the strong view above) where all powers are phenomenal powers and there are no nonphenomenal powers.

A weaker alternative to Mørch's extended view is a view on which for all phenomenal powers, there are no isomorphic nonphenomenal powers, though there may be nonisomorphic nonphenomenal powers (playing structurally distinct roles). Then given a full topic-neutral explanation of our phenomenal intuitions that is realized by phenomenal powers, it is unlikely that this explanation could have been realized by nonphenomenal powers. Perhaps it will remain possible that there are other systems with very different structure in which phenomenal intuitions are realized and explained by nonphenomenal powers. But this view no longer has the same strong explanatory and modal independence. What explains phenomenal intuitions in us will not be explanatorily or modally independent of consciousness, and the luck problem is correspondingly diminished.

The residual question is whether it is plausible that phenomenal powers could explain phenomenal intuitions. Mørch adapts an idea of Harold Langsam's to make the case that they could. Perhaps every phenomenal state has a power to cause judgments that we are in that state, at least under appropriate conditions of inward attention. This does not obviously explain our problem intuitions ("Consciousness is hard to explain") but it may help explain phenomenal intuitions ("I am conscious", "I am in pain").

My major worry here is that it is not at all obvious how to reconcile phenomenal powers of this sort with a Russellian phenomenal power panpsychism, on which the fundamental phenom-

enal powers correspond to fundamental microphysical powers found in physics. This is Mørch's preferred version of the view and perhaps the most powerful. On this view it is hard to see how the fundamental phenomenal states involve anything like a power to cause judgments under certain conditions. Now, Mørch may say that only higher-level phenomenal states have this power. Phenomenal power panpsychism allows that there are higher-level phenomenal states (constituted by microphenomenal states) that have higher-level phenomenal powers (constituted by microphenomenal/ microphysical powers). Still, the restriction seems somewhat ad hoc: why don't all phenomenal states have this power? Also, even if the restricted claim is correct, we still need to explain how the microphenomenal powers could add up to and constitute these distinctive powers. Perhaps that will be yet another aspect of the difficult combination problem for panpsychism.

One way to avoid this problem is to embrace an interactionist version of the phenomenal powers view. On this view, phenomenal states have no special association with microphysics, and fundamental phenomenal powers will be quite distinct from microphysical powers. Of course powers like this are hard to reconcile with the causal closure of microphysics, but I set that familiar problem aside here. On an interactionist phenomenal powers view, it may be that all phenomenal states are associated with a capacity for inward attention to them and a power to cause judgments that they are present, at least when inward attention is deployed. These phenomenal powers could then play a central role in grounding a topic-neutral explanation of our phenomenal judgments.

Saad proposes another very interesting version of realizationism that takes an interactionist rather than a panpsychist form. Where Mørch's approach centers on phenomenal powers, Saad's centers on a proposal about a fundamental psychophysical law by which consciousness affects matter. Saad's idea is roughly that Pautz's normative harmony thesis has the status of a psychophysical law. That is (roughly) there is a law saying that conscious states tend to cause the effects that they most rationalize.

This is an intriguing proposal. Obviously there are many questions. For a start, conscious states often don't cause the states that they rationalize. People can be irrational, or just fail to form judgments that it would be rational to form. So the law here must involve some sort of weak disposition. One might also worry that conscious states seem to cause many effects that they don't rationalize. Another issue is that a full interactionist theory should specify effects of consciousness on microphysics, but this law does not. The rationalized effects are on mental states and actions, which may constrain microphysical states but leaves the precise microphysical effects open. Still, Saad offers his proposal just as a toy law that could be complicated and qualified in various ways to handle these issues.

The proposal is well-placed to defuse the coincidence problem for realizationism. It is plausible that conscious states rationalize judgments that we are in those states, and at least arguable that conscious states rationalize problem judgments about the irreducibility of consciousness. If so, the proposal entails that conscious states will cause these phenomenal judgments and problem judgments. Furthermore, consciousness will have a really central role in explaining them. As before it is possible to give a topic-neutral version of this explanation that does not mention consciousness. For example there may be a non-normative algorithmic version that spells out all the effects. But it is plausible that the explanation in terms of consciousness is simpler and more powerful, not least because it explains and unifies all the effects involved in the algorithm. So as with Mørch, modal independence does not entail that the role of consciousness is a coincidence. There may also be a topic-neutral normative explanation of the intuitions saying there are states that rationalize certain effects and tend to cause what they rationalize, but now it may well be that (as Saad suggests) this story can only be realized by conscious states. Either way, this view is well-placed to avoid the modal and explanatory arguments for coincidence. On this story it is not a matter of luck that consciousness realizes the meta-problem processes.

The biggest worry for this view is the usual interaction problem for interactionism, focusing especially on reconciling the view with physics. Presumably the view will deny causal closure and require special nonphysical effects on physics, and presumably it would be possible to get experimental evidence for these effects in principle. So far there is no experimental evidence for these effects, so there is at least an interesting empirical challenge here.

In any case, Mørch's and Saad's proposals both strike me as intriguing and promising realist responses to the meta-problem. It may even be possible to combine them, somewhat in the spirit of Langsam's view. We could endorse a version of the phenomenal powers view saying that experiences have the power to cause what they rationalize. That view is not so easy to reconcile with Russellian panpsychism, for reasons discussed above. But if we deploy the view as a version of interactionism rather than a version of panpsychism, the problem is avoided, albeit at cost of the usual worries about the closure of physics. The result might be an account of metaproblem processes where consciousness plays a deep and fundamental explanatory role.

The appeal to acquaintance

We have so far not said much about debunking arguments that focus on justification. Where these debunking arguments are concerned, acquaintance is often given a central role. My own

epistemological discussions of consciousness have leaned on acquaintance frequently. One key idea is that even if there is a complete explanation of our phenomenal judgments or intuitions that is independent of consciousness, this does not undermine the justification of those beliefs, which is grounded not in these mechanisms but in direct acquaintance with consciousness. This appeal can be used to undermine key premises connecting genealogy with justification, such as premise 2 in the first debunking argument above.

In the current symposium, acquaintance is discussed by Katalin Balog, Keith Frankish, and Joseph Levine. Balog and Levine are advocates of an appeal to acquaintance, while Frankish is a critic.

Frankish discusses acquaintance in the context of his "hard meta-problem", which is itself a sort of debunking challenge in the spirit of justification-based debunking arguments, though with some different details. His hard meta-problem is roughly that of why we should trust our phenomenal judgments, given that they can be causally explained without appealing to consciousness. Frankish argues that the meta-problem processes and even their correlation with and realization by consciousness cannot fully answer the question. They might explain the reliability of the judgments (and so their justification in an external sense), but this leaves open the question of how we can be confident that the conditions for justification are satisfied.[9]

Frankish argues that confidence in phenomenal judgments (or that the conditions for their justification hold) requires a separate source of justification, over and above those deriving from the meta-problem processes and the way they are correlated with and realized by consciousness. The natural candidate for this separate source of justification is acquaintance. Perhaps one could also appeal to phenomenal powers or teleological laws to help ground the justification without acquaintance, but it is not entirely obvious how this would work. Arguably, even realizationist views such as these will benefit from an appeal to acquaintance.

Frankish focuses on acquaintance as the needed source of justification, and argues that it is problematic. His first criticism is that it is not clear that acquaintance is consistent with a naturalistic worldview. It is true that it is not clear that acquaintance is compatible with a *physicalist* worldview. Perhaps this should give pause to materialists who want to appeal to acquaintance to resist the debunking arguments. But for nonreductionist realists who appeal to acquaintance, this

[9] I made a related point in *The Conscious Mind*, arguing that mere reliability does not explain justification with certainty. I am not sure the point is best put in terms of confidence that the conditions for justification are satisfied—this sounds like a sort of higher-order justification, for which issues pertaining to first-order justification recurs. For some relevant discussion of certainty and higher-order justification see Bayne 2001 and my reply in Chalmers 2003.

worry carries little force. Maybe Frankish means to suggest that acquaintance is also problematic for naturalist versions of dualism, panpsychism, and the like, but this would take considerable argument.

His second criticism is that acquaintance requires a self or subject to stand in the acquaintance relation, and it is not clear what these subjects are. In my view a (non-Russellian) dualist should probably endorse primitive subjects, holding that fundamental phenomenal properties are instantiated by fundamental entities (a kind of entity dualism or substance dualism). Frankish says that substance dualist views are unattractive and widely rejected, but he does not give specific objections, and is not clear that the obejctions here are worse than those applying to dualism (especially interactionist dualism) generally. A Russellian monist should probably endorse complex subjects, perhaps being somewhat deflationary about subjects. Frankish's objection to the Russellian view is that it "buries consciousness away in an ontological limbo, where it shadows causal processes without making a causal difference". But the Russellian view certainly gives a causal role for consciousness, and the phenomenal powers version of the view integrates consciousness quite centrally into the causal order. So I think the appeal to acquaintance can resist Frankish's critique.

Levine is sympathetic with acquaintance, holding that there are psychophysical laws ensuring that physical/computational representation of objects and properties yields phenomenal acquaintance with those objects and properties. He raises a somewhat different meta-problem-style worry for the view: it's hard to see how acquaintance could be the source of our knowledge of acquaintance. This would require that acquaintance is itself an an input to our cognitive mechanisms, and it is not clear how this would work.

In response, one could make the standard Russellian appeal to acquaintance with acquaintance. Perhaps we know about acquaintance because we are acquainted with it. At the cognitive level, we have some representation of acquaintance that gives rise to phenomenal acquaintance with acquaintance. It is not clear why things must go different for acquaintance with acquaintance as opposed to acquaintance with red. Still, it will be worrying for a realist if our judgments about acquaintance can be explained quite independently of acquaintance itself. One will have a more attractive model if acquaintance plays a key causal and explanatory role. Levine speculates that this might work if there is a necessary tie (perhaps of a Russellian monist variety) between being a certain sort of cognitive system and having appropriate experiences.

I think the phenomenal powers view and/or teleological laws can help here. If one accepts acquaintance, it is plausible that phenomenal acquaintance with objects and properties rational-

izes judgments that one is aware of those objects and properties. If we accept a tie between what experiences rationalize and what they cause, this suggests that acquaintance with objects and properties tend to bring about corresponding judgments. If one accepts that we are acquainted with acquaintance, this will rationalize and tend to cause judgments about acquaintance. In this way, acquaintance will have a central role in explaining our judgments about acquaintance.

Balog is also sympathetic with a key role for acquaintance, and uses it to resist debunking arguments in a slightly different way. She holds that acquaintance plays an ineliminable role in explaining our judgments about consciousness. As a result premise 1 of the justification-based argument, "There is a correct explanation of our beliefs about consciousness that is independent of consciousness", is false.

Balog allows that one could give a physical/functional explanation at phenomenal beliefs when they are conceived of topic-neutrally from the third-person point of view. This explanation need never mention acquaintance. But in this case she holds that the key version of premise 2 is false: the fact that phenomenal beliefs conceived this way can be explained topic-neutrally does not entail that phenomenal beliefs conceived the ordinary way (in terms of consciousness, from the first-person point of view) can be so explained, and crucially it does not entail that ordinary phenomenal beliefs are unjustified. This is in effect a response to premise 1 that appeals to constitutive role for consciousness in constituting ordinary beliefs about consciousness. Balog extends this to a constitutive role for acquaintance in explaining the process that leads to these ordinary beliefs (at least as ordinarily conceived).

Of course it would be nice to know how acquaintance plays its causal role on Balog's story, and how it can ultimately be a physical relation, as Balog's type-B materialism requires. Balog will presumably not appeal to fundamental phenomenal powers or teleological laws. Perhaps acquaintance on her view will be identical with an underlying functional property, such as a sort of access consciousness that disposes subjects to form certain functional states that are themselves identical to phenomenal beliefs. Frankish's worry about the naturalizability of acquaintance remains. I have argued (Chalmers 2007) that the type-B materialist cannot give a good explanation of acquaintance. In response, Balog (20xx) has argued this is to be expected: on the type-B materialist story, there will be an explanatory gap between physical processes and (not just consciousness but) acquaintance. We agree about the gap but disagree about how serious a problem it is for materialism. Still, it would be interesting to see the physical property that acquaintance is identical to, in order to assess the seriousness of the explanatory gap in context.

Balog says earlier on that she thinks the meta-problem is unsolvable because ordinary phe-

nomenal judgments cannot be explained topic-neutrally, i.e. without an appeal to acquaintance on consciousness. But the meta-problem as I state it requires only that phenomenal intuitions can be so explained, where phenomenal intuitions are the topic-neutral counterparts of phenomenal judgments. Balog appears to initially deny that even this meta-problem can be solved ("The problem intuitions, even if they can be given topic-neutral formulations, could not be fully explained topic neutrally"), but later in the article she appears to concede that it can be solved ("It is fairly uncontroversial that it is possible, at least in principle, to fully explain the problem intuitions, or their neural correlates so stated in physical/functional terms.").

Balog goes on to argue that this explanation of phenomenal intuitions will not yield an explanation of full quasi-phenomenal judgments in the ordinary sense, which undercuts the meta-problem program in spirit but not in letter. For my purposes it is enough that she endorses the letter. Of course it is a further question whether this thesis leads to further claims such as illusionism and the undermining of phenomenal beliefs. My view, like Balog's, is that it does not.

Illusionism

Many of the commentators touch on issues about illusionism, though relatively few of them focus on it at length. Surprisingly few endorse strong illusionism, on which phenomenal consciousness does not exist: perhaps only Dennett, Frankish, and Kammerer. Many more endorse weak illusionism. It is interesting to see that so many back off from holding strong illusionism. Even Dennett and Frankish seem to wobble on this sometimes, saying that they are not so much denying that there is something it is like to be us as denying certain theoretical conceptions of what this involves. One explanation is that they feel the force of the Moorean argument against strong illusionism. Another explanation is that others find strong illusionism unbelievable, and there is a strong force to put one's view into a form others can believe.

I will touch on the Moorean argument in the final section. For now, I will consider another sort of argument against illusionism. This is a regress argument, versions of which are given by Giovanni Merlo and Michelle Liu.

Michelle Liu focuses on Pereboom's version of illusionism, which holds that we misrepresent experiences as having certain what-it-is-like properties when they do not have them. (Here I will follow Liu in using "experience" so that experiences need not have wil-properties.)

Liu first asks whether the introspective misrepresentations in question are beliefs or experiences, and raises problems for both options. I think the illusionist should hold that the illusion in

question involves states more primitive than beliefs, so I will take the second horn. Then when an experience e_1 seems to have wil-properties, it is misrepresented by an introspective experience e_2. The introspective state seems to be phenomenal too, so we need a higher-order introspective representation e_3 to misrepresent it. And so on, with an infinite regress.

I think the illusionist should reject the regress at the second step. When I introspect an ordinary experience e_1, it seems to have wil-properties, but by and large the introspective experience e_2 does not. Perhaps very occasionally it does, when I engage in higher-order introspection. In those cases there can be an appeal to higher-order representations like e_3. But in the ordinary case where there is no higher-order introsopection, e_3 is not needed to do any explanatory work. So Liu's regress is avoided. At best there is a potential regress where arbitrary higher-order states in the sequence *can* exist, and not an actual regress where they do exist.

Merlo raises a related regress for an illusionist view, with a hierarchy of pain-related states required to account for illusions of pain and to make the illusions justified. The details are complex, but I think that as with Liu's regress, I think the illusionist should reply by saying that Merlo's regress is potential but not actual.

Merlo also gives another argument against illusionism, from two principles: (1) If S is having an illusion that p, then S is in a mental state that provides him/her with immediate justification to believe that p. (2) If S is in a mental state that provides him/her with immediate justification to form a certain phenomenal belief, then the content of that belief is true.

Here I think illusionists might happily reject either premise, though which to reject is a substantive matter. For someone who thinks consciousness grounds the justification of our phenomenal beliefs, it may be natural to reject (1) where illusions of consciousness are concerned. If not, then one can reject (2) by treating phenomenal illusions like perceptual illusions, which give immediate justifiction for false perceptual beliefs. In the worst case, if (1) and (2) are non-negotiable, the ilusionist may simply deny that strictly speaking the states in question are illusions; perhaps they are merely quasi-illusions. Not much else of importance changes.

Conclusion

In my view, the coincidence argument is the strongest argument for illusionism, and the Moorean argument is the strongest argument against it. Here we have explored some ways that the coincidence argument can be defanged, but the Moorean argument remains. One might put is as follows. If illusionism were true, we would be phenomenologically blank. But we are not phenomenologi-

cally blank. So illusionism is false.

Of course the Moorean argument is flatfooted and illusionists will find it easy to reject key premises. Still, there is little doubt that something like the Moorean argument is the reason that most people reject illusionism and many find it crazy. So I think illusionists need to do more to defang the intuitive force of the Moorean argument in order for the view to take hold.

We can put things in terms of the dual challenges for the illusionist and the realist to rescue their programs from the threat of absurdity.

> For the illusionist, what is needed is an explanation of how having a mind without phenomenal consciousness could be like this, even though it is not at all the way that it seems. What would be ideal is something that does more than explaining our reactions and judgments (which seems to simply miss the phenomenon), without going so far as explaining the conscious experience itself (which an illusionist cannot do).
>
> For the realist, what is needed is an explanation that shows how consciousness and meta-problem processes are inextricably intertwined. What would be ideal is an explanation of why the metaproblem processes are by their nature grounded in consciousness, even if it is metaphysically possible for them to occur without consciousness.

In the current symposium, Mørch and Saad try to meet the challenge for the realist. Mørch's phenomenal powers view and Saad's teleological laws have the potential to show how the processes producing judgments about consciousness are best explained by consciousness even though it is possible that they occur without consciousness. That is progress. It does not mean that the views are correct, and of course these views require serious metaphysical investments that many will resist. Still, they open up promising territory for exploration.

In the current symposium, no one takes up the challenge for the illusionist. Frankish is the only one to address it, giving the familiar response that if "like this" involves a phenomenal feel there is nothing to explain, and if "like this" involves a reactive state illusionists can explain it. In the passage above I am looking for something more: an explanation showing our states could be "like this" in an intermediate sense that does not require phenomenal feels, but involves than just our reactions and judgments.

One might say that what we need is a non-phenomenal explanation of our phenomenological non-blankness—and one that goes beyond a cheap explanation in terms of reactions and judgments

while still being an explanation in broadly functional terms. I don't know how this can go, and the basic worry that any functional story will be consistent with phenomenological blankness is hard to avoid. But if something like this can be done, it might open the way for illusionism to become far more widely accepted.

References

Balog, K. 2020. Disillusioned. *Journal of Consciousness Studies*.

Balog, K. 2012. In Defense of the Phenomenal Concept Strategy. *Philosophy and Phenomenological Research* 84:1-23.

Bayne, T. 2001. Chalmers on the Justification of Phenomenal Judgments. *Philosophical and Phenomenological Research* 62:407-419.

Block, N. 1980. Are absent qualia impossible? *Philosophical Review* 89:257-74.

Chalmers 1996. *The Conscious Mind*. Oxdord University Press.

Chalmers, D.J. 2003. The Content and Epistemology of Phenomenal Belief. In (Q. Smith & A. Jokic, eds.) *Consciousness: New Philosophical Perspectives*. Oxford University Press.

Chalmers, D.J. 2007. Phenomenal Concepts and the Explanatory Gap. In (T. Alter & S. Walter, eds) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford University Press.

Chalmers, D.J. 2018. The Meta-Problem of Consciousness. *Journal of Consciousness Studies* 25:6-61.

Chalmers, D.J. & Jackson, F. 2001. Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110:315-61.

Clarke-Doane, Justin 2019. Undermining Belief in Consciousness. *Journal of Consciousness Studies* 26 (9-10):34-47.

Field, H. 1989. Introduction to *Realism, Mathematics, and Modality*. Blackwell.

Frankish, K. 2019. The Meta-Problem is The Problem of Consciousness. *Journal of Consciousness Studies* 26 (9-10):83-94.

Joyce, R. 2006. *The Evolution of Morality*. MIT Press.

Korman, D.Z. 2019. Debunking Arguments. *Philosophy Compass* 14 (12).

Levine, J. 2019. On the Meta-Problem. *Journal of Consciousness Studies* 26 (9-10):148-159.

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

Mrch, H.H. 2020. The phenomenal powers view and the meta-problem of consciousness. *Journal of Consciousness Studies*.

Pautz, A. 2020. Consciousness and coincidence: Comments on Chalmers. *Journal of Consciousness Studies*.

Saad, B. 2019. A Teleological Strategy for Solving the Meta-Problem of Consciousness. *Journal of Consciousness Studies* 26 (9-10):205-216.

Street, S. 2006. A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies* 127:109-166.