The Meta-Problem of Consciousness

David J. Chalmers

The meta-problem of consciousness is (to a first approximation) the problem of explaining why we think that there is a problem of consciousness.[1]

Just as metacognition is cognition about cognition, and a metatheory is a theory about a theory, the meta-problem is a problem about a problem. The initial problem is the hard problem of consciousness: why and how do physical processes in the brain give rise to conscious experience? The relevant sort of consciousness here is phenomenal consciousness. A system is phenomenally conscious if there is something it is like to be that system, from the first-person point of view. The meta-problem is roughly the problem of explaining why we think phenomenal consciousness poses a hard problem, or in other terms, the problem of explaining why we think consciousness is hard to explain.

The hard problem of consciousness is one of the most puzzling in all of science and philosophy, and there are currently no solutions that command any sort of consensus. The hard problem contrasts with the easy problems of explaining various behavioral functions such as learning, memory, perceptual integration, and verbal report. The easy problems are easy because we have a standard paradigm for explaining them. To explain a behavioral function, we just need to find an appropriate neural or computational mechanism that performs that function. We know how to do this at least in principle. In practice, the cognitive sciences have been making steady progress on the easy problems.

The hard problem is hard because explaining consciousness seems to require more than explaining behavioral functions. Even after we have explained all the behavioral functions that we like, there may still remain a further question: why is all this functioning accompanied by conscious experience? When a system is set up to perform those functions, from the objective point of view, why is there something it is like to be the system, from the subjective point of view?

[1]This is an early draft. For comments, thanks to Richard Brown, Dan Dennett, Francois Kammerer, Uriah Kriegel, Luke Muehlhauser, Claudia Passos, and Josh Weisberg.

Because of this further question, the standard methods in the cognitive sciences have difficulty in gaining purchase on the hard problem.

However, there is one behavioral function that has an especially close tie to the hard problem. This behavioral function involves *phenomenal reports*: the things we say about consciousness (that is, about phenomenal consciousness). In particular, many people make *problem reports* expressing our sense that consciousness poses a hard problem. I say things like "There is a hard problem of consciousness", "It is hard to see how consciousness could be physical", "After explaining behavioral functions, there remains a further question", and so on. So do many others. It is easy to get ordinary people to express puzzlement about how consciousness could be explained in terms of brain processes, and there is a significant body of psychological data on the "intuitive dualist" judgments of both children and adults.

The meta-problem of consciousness is (to a second approximation) the problem of explaining these problem reports. Problem reports are a fact of human behavior. Because of this, the meta-problem of explaining them is strictly speaking one of the easy problems of consciousness. At least if we accept that all human behavior can be explained in physical and functional terms, then we should accept that problem reports can be explained in physical and functional terms. For example, they might be explained in terms of neural or computational mechanisms that generate the reports.

Although the meta-problem is strictly speaking an easy problem, it is closely tied to the hard problem. We can reasonably hope that a solution to the meta-problem will shed significant light on the hard problem. A particularly strong line holds that a solution to the meta-problem will solve or dissolve the hard problem. A weaker line holds that it will not remove the hard problem, but it will constrain the form of a solution.

Like the hard problem, the meta-problem has a long history. One distinguished tradition involves materialists, who hold that the mind is wholly physical, trying to undermine dualist opponents by explaining away our intuitive judgment that the mind is nonphysical. One can find versions of this strategy in historical philosophers such as Hobbes, Hume, Spinoza, and Kant. For example, in the first paralogism in the *Critique of Pure Reason*, Kant argues that a "transcendental illusion" is responsible for our intuition that the self is a simple substance. More recently, U.T. Place (1956) diagnoses dualist intuitions about consciousness as resting on a "phenomenological fallacy", David Armstrong (1968) diagnoses them as resting on a "headless woman illusion", and Daniel Dennett (1992) diagnoses them as resting on a "user illusion".

This strategy typically involves what Keith Frankish has called *illusionism* about conscious-

ness: the view that consciousness is or involves a sort of introspective illusion. Frankish calls the problem of explaining the illusion of consciousness the *illusion problem*. The illusion problem is a close relative of the meta-problem: it is the version of the meta-problem that arises if one adds the thesis that consciousness is an illusion. Illusionists (who include philosophers such as Daniel Dennett, Frankish, and Derk Pereboom, and scientists such as Michael Graziano and Nicholas Humphrey) typically hold that a solution to the meta-problem will itself solve or dissolve the hard problem.[2] On this approach, if we have a physical explanation of why it seems to us that we have special nonphysical properties, then those properties can be dismissed as an illusion, and any problem in explaining them can be dismissed as resting on an illusion.

As a result, the meta-problem is especially important for illusionists. The best arguments for illusionism (as I will discuss toward the end of this article) rest on there being a solution to the meta-problem in physical or functional terms. If a consensus solution of this sort ever develops, then support for illusionism may swell. Even without a consensus solution to the meta-problem, thinking hard about the meta-problem may well make illusionism more appealing to more people. Speaking for myself: I have said before (e.g. Chalmers 1996, p. 189) that if I were a materialist, I would be an illusionist.[3] I think that if anything, illusionism has been underexplored in recent years. I take the view seriously, and I have more sympathy with it than with most materialist views.

That said, I am not an illusionist. On my view, consciousness is real, and explaining our judgments about consciousness does not suffice to solve or dissolve the problem of consciousness. But the meta-problem is not just a problem for illusionists. It is a problem for everybody. Even a non-illusionist can reasonably hope both that there be an explanation of our judgments about consciousness and that this solution will help us with the hard problem. Presumably there is at least a very close tie between the mechanisms that generate phenomenal reports and consciousness itself. Perhaps consciousness itself plays a key role in the mechanisms, or perhaps those mechanisms serve somehow as the basis of consciousness. Either way, understanding the mechanisms may well take us some distance in understanding consciousness.

[2]See Dennett 2016, Frankish 2016, Graziano 2013, Humphrey 2011, and Pereboom 2011. Some other recent illusionists may include: (philosophers) Clark 2000, Jackson 2003, Kammerer 2016, Rey 1996, Schwarz 2017; (others) Argonov 2014, Blackmore 2002, Drescher 2006, Hofstadter 2007, Muehlhauser 2017.

[3]Upon hearing about this article, some people have wondered whether I am converting to illusionism, while others have suspected that I am trying to subvert the illusionist program for opposing purposes. Neither reaction is quite correct. I am really interested in the meta-problem as a problem in its own right. But if one wants to place the paper within the framework of old battles, one might think of it as lending opponents a friendly helping hand.

I have long thought that solving the meta-problem might be a key to solving the problem of consciousness. My first serious article on consciousness (Chalmers 1987) argued that almost any intelligent machine would say that it is conscious and would be puzzled about consciousness, and argued from here that any convincing theory of consciousness must grant consciousness to machines. A subsequent article (Chalmers 1990) proposed a "coherence test" for theories of consciousness, holding that the explanation of reports about consciousness must cohere with the explanation of consciousness itself. That article also proposed a solution to the meta-problem, which was developed further in my book *The Conscious Mind* (1996, pp. 184-8 and pp. 289-92).

In effect, the meta-problem subsumes the illusion problem while being more general and more neutral. The meta-problem is neutral on the existence and nature of consciousness, while the illusion problem presupposes an extremely strong view about the existence and nature of consciousness. Since illusionism is held only by a small minority of theorists, it makes sense for community as a whole to understand the problem as the meta-problem and focus on solving it.[4] Theorists can then draw their own conclusions about what follows.

The meta-problem is a problem for reductionists and nonreductionists alike, dualists and physicalists alike, illusionists and non-illusionists alike. For the most part, this paper will stay reasonably neutral on those questions. I am most interested to explore the meta-problem as a problem in its own right. Toward the end of the paper, I will explore how the meta-problem may impact philosophical theories of consciousness, focusing especially on the prospects for illusionism and related views.

The meta-problem opens up a large and exciting empirical and philsophical research program. The question of what mechanisms bring about our problem reports is in principle an empirical one. We can bring philosophical methods to bear on assessing solutions, but as with the other "easy problems", the methods of psychology, neuroscience, and other cognitive sciences will play a crucial role.

In practice, one can already see the glimmer of a research program that combines at least (i) work in experimental philosophy and experimental psychology studying subjects' judgments

[4] I suggested the name "illusion problem" to Frankish, who had previously been calling the illusionist version of the problem the "magic problem" (a name with its own limitations). Mea culpa. I should also note that related "meta" problems have been suggested by Andy Clark and Francois Kammerer. Clark's "meta-hard problem" is the problem of whether there is a hard problem of consciousness. Kammerer's "illusion meta-problem" (2017) is the problem of why illusionism about consciousness is so hard to accept. These problems are distinct from what I am calling the meta-problem, but they are certainly related to it.

about consciousness, (ii) work in psychology and neuroscience on the mechanisms that underlie our self-models and bring about problem reports and other phenomenal reports, (iii) work in artificial intelligence and computational cognitive science on computational models of phenomenal reports, yielding computational systems that produce reports like ours, and (iv) philosophical assessment of potential mechanisms, including how well they match up with and explain philosophical judgments about consciousness.

In what follows, I will first clarify just what the meta-problem involves. Next, I will present and evaluate a number of possible solutions to the meta-problem that have been offered in the existing literature, and try to narrow things down to the solutions that I think are the most promising. Finally, I will discuss how the meta-problem bears on debates about theories of consciousness (especially illusionism), and how a solution to the meta-problem might shed light on the problem of consciousness more generally.

1 What is the Meta-Problem?

I introduced the meta-problem as the problem of explaining why we think there is a problem of consciousness. I elaborated it as the problem of explaining our problem reports, where these are our reports about consciousness that reflect our sense that consciousness poses a special problem. It is time to be a bit more specific about what this comes to: in particular, what needs to be explained, and what sort of explanation counts.

What needs to be explained? The data that need explaining can be construed as verbal reports (my saying "Consciousness is hard to explain"), as judgments (my forming the judgment that consciousness is hard to explain), or as dispositions to make these reports and judgments. Verbal reports are perhaps the most objective data here, but they are also a relatively superficial expression of an underlying state that is the real target of investigation. So I will generally focus on dispositions to make verbal reports and judgments as what we want to explain.

I will call dispositions to make specific problem reports and judgments *problem intuitions*. There may be more to the states ordinarily called intuitions than this, but it is plausible that they at least involve these dispositions. As I am using the term, problem intuitions can result from inferences, so that even judgments that result from philosophical arguments will count as problem intuitions. At the same time, problem intuitions that result from inferences typically rest on more fundamental problem intuitions that do not. So I will focus especially on non-inferential problem intuitions that arise prior to philosophical argument.

Next, which intuitions need to be explained to solve the meta-problem? In principle phenomenal reports include any reports about consciousness, including mundane reports such as 'I am feeling pain now'. The problem of explaining the corresponding intuitions is certainly an interesting problem. The meta-problem proper, however, is the problem of explaining problem intuitions: intuitions that reflect our sense that there is some sort of special problem involving consciousness, and especially some sort of gap between physical processes and consciousness. For example, 'I can't see how consciousness could be physical' is a problem report, and the disposition to judge and report this is a problem intuition.

Problem intuitions divide into a number of categories. Perhaps the core intuitions for the meta-problem as defined are *explanatory intuitions* holding that consciousness is hard to explain. These include gap intuitions holding that there is an explanatory gap between physical processes and consciousness, and anti-functionalist intuitions holding that explaining behavioral functions does not suffice to explain consciousness. Closely related are *metaphysical intuitions*, including dualist intuitions holding that consciousness is nonphysical, and fundamentality intuitions holding that consciousness is somehow fundamental or simple. There are also *knowledge intuitions*: these include both first-person knowledge intuitions holding that consciousness provides special knowledge from the first-person perspective (like Mary's knowledge of what it is like to see red on leaving the black and white room), and third-person ignorance intuitions, such as the intuition that it is hard to know the consciousness of other people or other organisms (such as what it is like to be a bat). There are *modal intuitions* about what is possible or conceivable, including the "zombie" intuition that a physical or functional duplicate of us might lack consciousness and "inversion" intuitions, such as that someone else might be experiencing red when I experience green.

I will take these four classes (explanatory, metaphysical, knowledge, and modal intuitions) to be the central cases of problem intuitions, with the first two being the most central. There are also some nearby intuitions that are closely related. For example, there are *value intuitions*, holding that consciousness has special value: perhaps that it makes life worth living, for example. There are *distribution intuitions*, concerning which systems do and don't have consciousness: for example, the intuition that robots or plants are not conscious. There are *self intuitions* concerning the self or the subject of experience. There are *quality intuitions* concerning the special qualities (colors and the like) that are presented in experience. There are *presentation* intuitions concerning the direct way these qualities are presented to us. The list goes on. I will not attempt to draw up a full list here.

The range of these intuitions is an empirical question. I could perhaps be accused of focusing

on the intuitions of philosophers, and of a subclass of philosophers at that. But I think the central intuitions are widely shared well beyond philosophy. It is highly plausible that versions of many of these intuitions can be teased out of ordinary subjects, but it is an empirical matter just how widespread they are.

There is a large body of research in experimental psychology and experimental philosophy on people's intuitions about the mind, but surprisingly little of it to date has concerned core intuitions about the problem of consciousness. Perhaps the largest body of research concerns childrens' intuitions about belief: for example, does a three-year old have the concept of false belief? Another large body concerns intuitions about the self and personal identity: for example, do people think that the self goes with the body or the brain in a brain transplant case? Where consciousness is concerned, the largest body of research concerns the distribution of consciousness (e.g. Gray et al 2007; Knobe and Prinz 2008; Systema and Machery 2010): for example, do people think that machines or corporations can feel pain? Some attempts have been made to connect this research to the hard problem of consciousness,[5] but for the most part the intuitions in question have not been the core problem intuitions.

What about experimental research on the core problem intuitions? In principle there is room for experimental work on modal intuitions (e.g. the conceivability of zombies) or knowledge intuitions (e.g. Mary's knowledge in and out of her black and white room), but I do not know of any work along these lines to date. Where metaphysical intuitions are concerned, there is a non-negligible body of literature on "intuitive dualism" (e.g. Bloom 2004; Chudek et al 2013; Richert and Harris 2008), but the main body of this research largely focuses on intuitions about the self (e.g. could a self move between bodies or survive bodily death?) rather than about consciousness per se. There is a small body of relevant work on explanatory intuitions. For example, Gottlieb and Lombrozo (forthcoming) elicit judgments about when various phenomena are hard for science to explain, and find that people judge that phenomena tied to subjective experience and to privileged access are relatively hard to explain.[6]

[5]For example, Systma and Machery (2010) find that ordinary subjects are much more likely to say that a robot can see red than that it can feel pain, and they conclude that ordinary subjects do not have a unified category of phenomenal consciousness, subsuming seeing red and feeling pain, that generates the hard problem. In fact I predicted a version of their finding in Chalmers 1996 (p. 18), which observes that ordinary mental terms like this have both a functional reading and a phenomenal readings, with sensational terms such as "pain" more likely to suggest a phenomenal reading than perceptual terms such as "see". Other relevant work includes Huebner 2010, Talbot 2012, and Peressini 2014.

[6]I'm happy to told about other relevant work on problem intuitions! One related empirical study is the PhilPapers Survey of professional philosophers (Bourget and Chalmers 2014)—although this is not really experimental, and most

As a result, it is hard to know how widely shared the problem intuitions are. It is clear that they are not universal, at least at the level of reflective judgment. All of them are rejected by some people. In some cases of rejection, there may be an underlying intuition that is outweighed by other forces (for example, a dualist intuition might be outweighed by reasons to accept physicalism), but it is not obvious that there is always such an underlying intuition. A fully adequate solution to the meta-problem should be able to explain not only why these intuitions are widely shared, if they are, but also why they are not universal, if indeed they are not.

As a first approximation, I will work under the assumption that these intuitions are strong, robust, and widely shared. Of course this is an empirically defeasible assumption, and I would be delighted to see experimental work that tests it.[7] Even if the assumption is false, the more limited task of explaining the intuitions in people who have them will still be of considerable interest. For example, it will still be crucial for illusionists to explain those intuitions, in order to make the case that they are illusory. Solving the meta-problem will remain an important project either way.

What counts as an explanation?[8]

What sort of explanation counts as an explanation of the problem reports, for the purposes of the meta-problem? For example, does it count as an explanation to say that we judge that consciousness poses a problem because consciousness does indeed pose a problem, and we notice that? Perhaps in some contexts that would count as an explanation, but it is not the sort of explanation we are concerned with here.

Earlier, I motivated the meta-problem as follows: "if we accept that all human behavior can be explained in physical and functional terms, then we should accept that phenomenal reports can be explained in physical and functional terms." To a first approximation, then the meta-problem asks for an explanation of problem reports in physical or functional terms. Ideally, we would like to specify neural or computational mechanisms that are responsible for phenomenal reports.

---

questions concern considered judgments rather than immediate intuitions. The survey found that 36% of the target group judge that zombies are conceivable but not possible, 23% judge that they are inconceivable, and 16inconceivable (with 25% agnostic or giving other answers). 56% endorsed physicalism about the mind while 27% endorsed non-physicalism about the mind.

[7]For what it's worth, I predict that knowledge intuitions will be somewhat more widespread than modal intuitions, and that explanatory intuitions will be somewhat more widespread than metaphysical intuitions. But as always, a great deal will depend on the way that key claims are formulated (this may be particularly difficult where modal intuitions are concerned), and the fact that someone denies a key claim (say, that consciousness is nonphysical) is consistent with their having an underlying intuition that is outweighed.

[8]This section goes into a bit more philosophical detail than other sections and can easily be skipped by readers without much background in philosophy.

To a second approximation, we can require an explanation of problem intuitions in *topic-neutral* terms: roughly, terms that do not mention consciousness (or cognate notions such as qualia, awareness, subjectivity and so on). Physical and functional explanations will be topic-neutral explanations, but so will some other explanations.

An advantage of doing things this way is that the meta-problem then arises even for views where behavior cannot be explained in physical terms. These include especially interactionist dualist views, on which consciousness is nonphysical and interacts with the brain. On Descartes' view, for example, an explanation of human behavior will appeal to nonphysical consciousness that drives brain processes via the pineal gland. Descartes' view is no longer popular, but there are contemporary views that share its spirit. For example, some theorists hold that nonphysical consciousness drives physical processes by collapsing a quantum wave function in the brain, in which case a full explanation of human behavior must appeal to nonphysical consciousness.

One might think that an interactionist view evades the meta-problem, but in fact a version of it still arises. For example, suppose that nonphysical consciousness is arranged in such a way that it carries out a specific computation (in ectoplasm, say), and its causal role always goes through the outcome of such a computation. Then we could explain human behavior in computational terms without ever mentioning consciousness. Or suppose that nonphysical consciousness always collapses the quantum wave function in certain specifiable circumstances according to the standard probabilities (given by the Born rule). Then in principle we could explain human behavior by saying that there is something that collapses the wave function in those circumstances, without ever saying that what does the collapsing is consciousness.

Nothing here entails that consciousness is causally irrelevant. On these interactionist views, consciousness will play a causal role in generating behavior, and a truly complete explanation of human behavior will mention consciousness. Nevertheless, it will be possible to give a good explanation of human behavior in topic-neutral terms that do not mention consciousness. This is roughly analogous to the way that on a standard physicalist view, neurons play a crucial causal role in generating behavior, but it is nevertheless possible to give a computational explanation of human behavior that does not mention neurons. In effect, the topic-neutral explanation specifies a structure, and neurons (or consciousness) play their role by undergirding or realizing that structure.

Something similar applies to panpsychist views, on which consciousness plays a causal role at the fundamental level in physics, by serving as the underlying basis of the microphysical roles specified in physics. On these views, consciousness plays a causal role in generating human behavior. Nevertheless, assuming that consciousness does not violate the laws of physics, it will

be possible to explain physical processes in topic-neutral mathematical terms that do not mention consciousness. Again, there may be something incomplete about this topic-neutral explanation, but it will still be an explanation. In principle, panpsychism is no obstacle to there being a solution to the meta-problem in topic-neutral terms.

We can then understand the meta-problem as "Explain problem intuitions in topic-neutral terms, or explain why such an explanation is impossible". The second horn allows that there may be views on which there is no good topic-neutral explanation. For example, there may be anomalous dualist views on which consciousness plays a completely unpredictable role, with effects that somehow depend on the intrinsic nonstructural features of consciousness itself. One could try to turn this structure into a topic-neutral explanation, but it is not clear that an adequate topic-neutral explanatiom will always be available. Another possibility: perhaps some anomalous monists and others might also argue against there being good physical explanations of behavior, even though physicalism is true. So it is at least open to respond to the meta-problem by making the case that there can be no adequate topic-neutral explanation of problem intuitions.

The move to topic-neutral explanation also opens up the possibility of further forms of explanation. For examples, it allows us to invoke *representational* explanation, perhaps in terms of models that represent the subject or the world as having certain properties. It is desirable that such an explanation can eventually be cashed out as a physical/functional explanation, but as long as it does not directly mention consciousness or cognates, it will count as topic-neutral. We can also invoke *rational* explanation, characterizing processes as doing certain things because they are rational.

An especially important form of explanation for the meta-problem is historical or teleological explanation. We do not just want to know (synchronically) how problem intuitions are produced. We want to know how problem-intuition-producing systems came to exist in the first place. Why were phenomenal intuitions a good idea? What evolutionary function did they serve, if any? A solution that gives a well-motivated story about the function of phenomenal intuitions will be more satisfactory than one that does not. Any complete solution to the meta-problem should say something about these historical and teleological questions.

A subtlety of the move to topic-neutral terms is that we have to reconstrue what we are explaining—problem intuitions—in topic-neutral terms. As initially described, problem intuitions concern *consciousness*, so that explaining them requires saying somthing specific about consciousness. Some problem intuitions may even concern specific phenomenal qualities such as the quality of pain. It is far from clear that the fact that our intuitions concern phenomenal properties can itself

be explained in topic-neutral terms. Many theorists (including me) hold that phenomenal beliefs turn on the existence of consciousness itself, so they cannot be fully explained in topic-neutral terms. To handle this, we need to reconstrue problem intuitions themselves in topic-neutral terms.

There are a couple of ways to do this. One could put phenomenal intuitions in an existential form, such as "We have special properties that are hard to explain" or "that are nonphysical", "that provide special first-person knowledge", "that could be missing in robots", and so on. Alternatively, one could simply require that phenomenal intuitions be explained up to but not including the fact that they are specifically about consciousness. Once we have explained judgments of the form "We have special first-person knowledge of X which is hard to explain in physical terms", and so on, we have done enough to solve the meta-problem. In the language of Chalmers (2007), we can call these quasi-phenomenal judgments. Quasi-phenomenal judgments do not so obviously depend on consciousness, and might even be shared by zombies.

A related issue is that some people think that all meaning is grounded in consciousness, so that it is impossible to explain genuinely meaningful reports or judgments in topic-neutral terms. On a view like this, one might nevertheless be able to explain our propensity to make certain noises and inscriptions (those we make when we make phenomenal reports) in topic-neutral terms, so one could trying construing these as the target for the meta-problem. Alternatively, one could use this view to argue that no topic-neutral explanation can be given.

To simplify, in what follows I will stipulate that problem intuitions are individuated as functional states. To a first approximation, one can think of them as dispositions to make quasi-phenomenal reports, where reports are understood as outputs that even a non-conscious being could make. These dispositions may be watered-down states compared to full-blown phenomenal beliefs, but they will still be interesting enough to pose an interesting meta-problem.

The meta-problem then becomes: Explain our problem intuitions in topic-neutral terms. For many purposes, especially when more exotic philosophical issues are set aside, it may suffice to think of the problem roughly as stated earlier: Explain in physical/functional terms why we think there is a problem of consciousness.

2 Potential solutions to the meta-problem

In what follows I will examine a number of candidate solutions to the meta-problem, involving topic-neutral explanations of our problem intuitions, focusing on their strengths and limitations. Many of these ideas have been put forward in the literature, often more than once. It is typical of

proposals about the meta-problem that they are made in isolation from other proposals, often without acknowledging any other work on the subject. I hope that bringing these proposals together will contribute to a more integrated research program in the area.

The first seven or so proposals are ideas that I find especially promising and that I think may form part of a correct account. After these, I will also discuss some ideas from others that I am less inclined to endorse, but which are nevertheless useful or instructive in thinking about the meta-problem. My overall aim is constructive: I would like to build a framework that may lead to a solution to the meta-problem. At the same time, I will be pointing out limitations and challenges that each of these ideas face, in order to clarify some of the further work that needs to be done for a convincing solution to the meta-problem.

I will often approach the meta-problem from the design stance. It may help to think of building a robot which perceives the world, acts on the world, and communicates. It may be that certain mechanisms that are helpful for the robot, for example in monitoring its own states, might also generate something like problem intuitions. At the same time I will keep one eye on what is distinctive about phenomenal intuitions in the human case. Contrasting these intuitions with our related intuitions about phenomena such as color and belief can help us to determine whether a proposed mechanism explains what is distinctive about the phenomenal case.

1. *Introspective models*:[9] An obvious place to start is that any intelligent system will need representations of its own internal states. If a system visually represents a certain image, it will be helpful for it to represent the fact that it represents that image. If a system judges that it is in danger, it will be helpful for it to represent the fact that it judges this. If a system has a certain goal, it will be helpful for it to represent the fact that it judges this. In general, we should expect any intelligent system to have an internal model of its own cognitive states. It is natural to hold that our phenomenal intuitions in general and our problem intuitions more specifically arise from such an internal model.

While this claim may be a key element of any solution to the meta-problem, it does not itself constitute anything close to a solution. For it to yield a solution, one would need an explanation of why and how our internal self-models produce problem intuitions. I have occasionally heard the suggestion that internal self-models will inevitably produce problem intuitions, but this seem

[9]Many attempts at solving the meta-problem give a role to introspective models. Introspective models are especially central in Graziano's "attention schema" theory of consciousness, which explains our sense of consciousness as a model of attention. Metzinger (2003) focuses on "phenomenal self-models" that appeal to phenomenal properties to explain certain illusory beliefs about the self, rather than explaining beliefs about phenomenal properties.

clearly false. We represent our own beliefs (such as my belief that Canberra is in Australia), but these representations do not typically go along with problem intuitions or anything like them. While there are interesting philosophical issues about explaining beliefs, they do not seem to raise the same acute problem intuitions as do experiences. Some people claim to have a nonsensory experience of thinking, but these intuitions are much less universal and also less striking than those in the case of sensory experience. Even if there are such experiences, it is not clear that introspecting one's beliefs (e.g. that Paris is the capital of France) always involves them. So more is needed to explain why the distinctive intuitions are generated in the phenomenal case.

2. *Phenomenal concepts*:[10] Another obvious starting point focuses on our concepts of consciousness, or phenomenal concepts. These function as special concepts to represent our phenomenal states, especially when we detect those states by introspection. The well-known phenomenal concept strategy tries to explain many of our problem intuitions in terms of features of our phenomenal concepts. If this works, and if the relevant features can then be explained in topic-neutral terms, we will then have a solution to the meta-problem.

I have criticized the phenomenal concept strategy elsewhere (Chalmers 2007), arguing that there are no features of phenomenal concepts that can both be explained in physical terms and that can explain our epistemic situation when it comes to consciousness. In that paper I construed the phenomenal concept strategy as a version of "type-B" materialism, which accepts a robust understanding of our epistemic situation on which many of our problem intuitions (e.g. knowledge and conceivability intuitions) are correct. To solve the meta-problem, however, we need only explain the fact that we have the problem intuitions; we do not also need to explain their correctness. There is an illusionist (or "type-A") version of the phenomenal concept strategy which holds that our problem intuitions are incorrect and our epistemic situation is not as we think it is (e.g. Mary does not gain new knowledge on seeing red for the first time), but on which features of phenomenal concepts explain why we have these intuitions in the first place. This use of phenomenal concepts is explicitly set aside in my earlier paper and is not threatened by the critique there.

Still, everything depends on what the account says about phenomenal concepts. In the earlier paper, I argued on the most common accounts where the features of phenomenal concepts can be physically explained, the concepts are too "thin" to explain our problem intuitions. For example, the suggestion that phenomenal concepts are indexical concepts such as "this state" does not re-

[10]The locus classicus of the phenomenal concept approach is Loar's (1990) appeal to recognitional concepts. Also relevant are the appeal to indexical concepts by Ismael (1999) and Perry (2001), the appeal to quotational concepts by Balog (2009) and Papineau (2007), and others.

ally explain our knowledge intuitions and others: when we pick out a state indexically as "this state", we are silent on its nature and there is no obvious reasons why it should generate problem intuitions. Similarly, the suggestions that phenomenal concepts are recognitional concepts akin to our concepts of a certain sort of cactus also does not explain the problem intuitions: when recognize a cactus, we do not have problem intuitions anything like those we have in the phenomenal case. Something similar goes for many extant suggestions. It may be that some other feature of phenomenal concepts can both explain our problem intuitions and be explained in physical terms, but if so it is this feature that will be doing the explanatory work.

3. *Independent roles:* It is sometimes suggested that many of our problem intuitions can be explained by the fact that physical and phenomenal concepts have independent conceptual roles, without strong inferential connections from one to the other. For example, Nagel (1974) suggests that our conceivability intuitions might be explained by the fact that physical concepts are tied to perceptual imagination and phenomenal concepts are tied to sympathetic imagination, where these two forms of imagination are independent of each other. This approach has been taken more generally by Hill and McLaughlin (1996) and others who hold that the fact that phenomenal concepts and physical concepts have independent roles can explain our explanatory intuitions and knowledge intuitions as well as conceivability intuitions.

There is certain something to this view, but it suffers from a familiar problem: our concepts of belief also seem independent from our physical concepts, but they do not generate the same problem intuitions. Phenomenal states seem problematic in large part because they seem to have a specific qualitative nature that is hard to explain in physical terms (where beliefs do not), and this seeming is not explained simply by the independence of phenomenal concepts. Ultimately, we need to explain why these qualitative properties seem to populate our minds, which requires an account of why we have introspective concepts that attribute these qualitative properties. Merely pointing to the independence of introspective concepts does not explain this.

4. *Introspective opacity*.[11] A central element of many attempts to address the meta-problem turns on the fact that the physical mechanisms underlying our mental states are opaque to introspection. We do not represent our states as physical, so we represent them as nonphysical. In the locus classicus of this approach, David Armstrong (1968) makes an analogy with the "headless woman" illusion. A sheet covers a woman's head, so we do not see her head. As a result, she seems to be headless. Armstrong suggests that we somehow move from "I do not perceive that the woman has a head" to "I perceive that the woman has no head". Likewise, in the case of consciousness, we move from "I do not introspect that consciousness is a brain process" to "I

introspect that consciousness is not a brain process".

An obvious problem is that the move is far from automatic. There are many cases where one perceives someone's body but not their head (perhaps their head is obscured by someone else's), but one does not typically perceive them as headless. Something special is going on in the headless woman case: rather than simply failing to perceive her head, one perceives her as headless, and this seeming itself needs to be explained. Likewise, there are any number of cases where one does not perceive that some phenomenon is physical, without perceiving that it is nonphysical. I might have no idea how the processes on my computer are implemented, but they do not seem nonphysical in the way that consciousness does. Likewise, when I introspect my beliefs, they certainly do not seem physical, but they also do not seem nonphysical in the way that consciousness does. Something special is going on in the consciousness case: insofar as consciousness seems nonphysical, this seeming itself needs to be explained. Perhaps introspective opacity can play a role in explaining this, but more work is needed to explain the transition from not seeming physical to seeming nonphysical.

5. *Direct access*:[12] A related idea, stressed in my own earlier work on the meta-problem, is that when a cognitive system introspects its own state, it will at least seem to have a sort of direct access to that state, not inferred from or mediated by any other knowledge. For example, if a computer system with both perceptual and introspective representations says that a green object is present, and one asks for its reasons, it might naturally answer that it is representing the presence of a green object. But if one asks for its reasons for saying that it is representing the presence of a green object, it may well have no further reasons. The system is thrust into that state by its introspective mechanisms, and is not given access to the mechanisms that bring the state about. It simply represents itself as representing greenness, without further reasons for this claim. In effect, introspective representations will at least seem to play a foundational role for the system. It is natural to think that these will then be represented by the system as primitive states that it finds itself in.

[11] Versions of the introspective opacity move can be found in Dennett's appeal to user illusions, Drescher's appeal to "gensyms", Graziano's appeal to attention schemas, Tegmark's appeal to substrate-independence, as well as my own appeal to information in Chalmers (1990; 1996). A historical precursor is Thomas Hobbes: "The gross errors of certain metaphysicians take their origin from this; for from the fact that it is possible to consider thinking without considering body, they infer that there is no need for a thinking body" (*De Corpore*, 3,4).

[12] My own proposed solution to the meta-problem in Chalmers (1990; 1996) used introspective opacity to motivate the direct access idea, and suggested that these phenomena would naturally lead to primitive quality attribution as below. Clark's (2000) analysis of the meta-problem builds on this approach, focusing on direct access to the sensory

Here the familiar problem strikes again: Everything I have said about the case of perception also applies to the case of belief. When a system introspects its own beliefs, it will typically do so directly, without access to further reasons for thinking it has those beliefs. Nevertheless, our beliefs do not generate nearly as strong problem intuitions as our phenomenal experiences do. So more is needed to diagnose what is special about the phenomenal case. At this point Clark (2000) appeals to the fact that we have direct access to the sensory modality involved in an experience (seeing rather than smelling, say), suggesting that this access entails that the subject will represent an experience as qualitative. However, in the case of belief we also have access to an attitude (believing rather than desiring, say), and it is not really clear why access to a modality as opposed to an attitude should make such a striking difference.

6. *Primitive quality attribution*:[13] A promising proposal picks up on an analogy with the meta-problem of color: roughly, why do colors seem to be irreducible qualitative properties? It is common to observe that vision presents colors as special qualities of objects, irreducible to their physical properties. It is also common to observe that this is an illusion, and that objects do not really have those special qualities. Why, then, do we represent them that way? A natural suggestion is that it is useful to do so, to mark similarities and differences between objects in a particularly straightforward way.[14] The perceptual system knows little about underlying physical properties, so it would be hard to represent colors in those terms. Perhaps it could just represent similarities and differences between objects without representing specific qualities, but this would be inefficient. Instead, evolution hit on a natural solution: introduce representations of a novel set of primitive qualities (colors), and when two objects are similar with respect to how they affect the relevant parts of visual system, represent them as having the same qualitative property.

Nothing here requires that the qualitative properties be instantiated in the actual world. In fact, nothing really requires that such properties exist even as universals or as categories. What matters is there seem to be such qualities, and we represent objects as having those qualities. (In

modality involved in acts of detection. Schwarz (2017) uses introspective opacity to motivate the introduction of illusory representations of sensory states to play a foundational role in Bayes-style belief updating.

[14]Derk Pereboom's "qualitative inaccuracy" thesis (2011) is roughly the idea that we misrepresent experiences (like external objects) as having primitive qualitative properties that they do not have. Versions of the idea that a cognitive system would naturally represent primitive qualities as a natural means of representing our own more complex states can be found in Chalmers (1990; 1996) and in Schwarz (2017). Hall (2007) introduces dummy properties to account for illusions about colors, but does not apply it to phenomenal properties.

[14]Check which color irrealists actualy say things along these lines: Averill? Hardin? Maund? Boghossian and Velleman? Pautz? Chalmers?

philosophers' language, we could represent the qualities de dicto rather than de re: that is, we could represent that objects have primitive qualities, even if there are no primitive qualities such that we represent objects as having them). In the words of Richard Hall (2007), experienced colors may be *dummy properties*, introduced to make the work of perception more straightforward. It is easy enough to come up with a computational system of color representation that works just this way, introducing a representational system that encodes qualities along an R-G axis, a B-Y axis, and a brightness axis. Because these axes are represented independently of other physical dimensions such as spatial dimensions, the corresponding qualities seem irreducible to physical qualities.

Something like this is plausibly at least part of the solution to the meta-problem of color. We represent primitive colors as a useful model of complex physical properties (such as reflectance properties) in our environment. Even if no such primitive colors are instantiated in our environment, the mere representation of apparently primitive properties suffices to explain their apparent irreducibility.

This idea can be extended to the meta-problem of consciousness by saying that introspection attributes primitive qualities to mental states for similar reasons. It needs to keep track of similarities and differences in mental states, but doing so directly would be inefficient, and it does not have access to underlying physical states. So it introduces a novel representational system that encodes mental states as having special qualities. Because these qualities are represented independently of other physical dimensions such as spatial dimensions, the corresponding qualities seem irreducible to physical properties.

This proposal works especially well on a view where phenomenal properties are (or seem to be) simple "qualia". Such a view might have the resources for explaining why our problem intuitions differ from our intuitions about belief: sensory states are represented as simple qualities, while beliefs are represented as relations to complex contents (the cat is on the mat) that do not require a novel space of qualities. However, the qualia view is widely rejected these days, even as an account of how experiences seem to us introspectively. It is much more common to hold that experiences are (or seem to be) representational or relational states. For example, the experience of greenness does not involve a simple "green" quality, but instead seems to involve awareness of greenness, the color. Here greenness is the same quality already used to represent external objects in perceptual representations, and awareness is a mental relation, understood as some sort of representation (on a representationalist view) or some sort of perception (on a relational view). On a view like this, it is unclear how a novel space of primitive qualities attributed in introspection will enter the picture.

This account also needs to address a crucial disanalogy between the representation of colors and phenomenal properties. It is typically easy for people to accept that colors are illusions and are not really instantiated in the external world, but it is much harder for people to accept that phenomenal properties are illusions and are not really instantiated in our minds. This worry is an instance of Kammerer's "illusion meta-problem", which I will call the *resistance* problem to avoid confusion: explain why there is so much intuitive resistance to illusionism.[15] Any primitive quality attribution account will need further ideas to explain the disanalogy.

7. *Primitive relation attribution*.[16] A closely related idea is that our introspective models attribute primitive *relations* to qualities and contents. Here we can think of a robot that visually senses the world around it, attends to certain objects, and has introspective representations of its own states. In fact the robot will stand in highly complex relations to objects and properties in the external world—a complex causal relation of seeing, an equally complex functional relation of visual representation, a complex functional relation of attending, and so on. The robot may not have access to all that complexity, and there may be little need to model all the details. So it is not unnatural to suppose that such a system's introspective models will introduce primitive relations of seeing, attending, and so on instead. When the robot sees a red square, instead of representing a complex causal relation to the red square, the system will model itself as standing in the primitive relation of seeing to the red square. Likewise, when it is in the complex state of visually represents a red square (without being sure whether the square is present), the system may itself more simply as standing in a primitive relation of visual experience or awareness to red squares. The same may go for attention and other complex cognitive relation.

We can then suppose that this sort of primitive relation attribution is present in our own introspective models. Perhaps this picture could be combined with primitive quality attribution grounded in perceptual models. Then our introspective models represent ourselves as standing in primitive relations (such as awareness) to primitive qualities (such as primitive greenness), when our physical states actually involve complex causal relations to complex physical qualities in the

[15] Kammerer's own proposal to solve the resistance problem is that we understand an illusion of X as a state in which we are affected the same way as in a correct perception of X, but without the underlying reality. Under those constraints illusionism about the phenomenal is incoherent, since one of the ways we are affected in a correct perception of conscious experiences is having conscious experiences. I think this diagnosis can account for some resistance to illusionism (people do often argue that illusionism is incoherent in this way), although I think it is easy to formulate understandings of illusions that avoid the problem. In any case, because this diagnosis turns on our concept of "illusion", it cannot account for the fact that we are equally resistant to nearby views that do not use that concept: for example, the view that we do not have any conscious experiences. To explain our resistance to views like this, we need to go deeper.

environment. This model might help to explain a number of our problem intuitions: experiencing a red object will seem relatively simple and primitive, when the underlying physical reality is complex.

This idea bears a structural resemblance to the key slogan of Graziano's attention schema model, according to which "awareness is a model of attention". I take Graziano to mean something like: "our model of awareness is in fact a simplified model of attention". That is, our introspective models represent a simple relation of awareness as a stand-in for the complex relation of attention that is present in the brain. Graziano does not speak of primitiveness here, and he says surprisingly little to apply his attention schema model to the meta-problem.[17] Still, his main focus is using a simple mental relation (awareness) to model a complex one (attention). One could easily adapt his slogan to the current context with primitive relation attribution by saying: our model of awareness is a primitive model of complex relations of attention. (Compare: our model of color qualities is a primitive model of complex physical reflectance properties.)

Another difference with Graziano is that I am not sure that attention is the right choice for the complex relation that is being modeled. On the face of it, perceptual awareness presents itself as a model of all perception, whether attended or unattended (which is why the idea of unconscious perception initially strikes us as counterintuitive). So I would be inclined to say: our model of perceptual consciousness is a primitive model of complex relations of perception. More generally, consciousness presents itself as a model of all mental representation (which is why the idea of unconscious representation initially strikes as counterintuitive). So I would say: our model of consciousness is a primitive model of complex relations of representation; or in Graziano's simpler terms, consciousness is a model of representation.

The familiar problem of belief still arises. On the face of it, it would make as much sense to represent a complex belief relation as primitive too, but we do not find the same problem intuitions. One response would be argue that awareness is represented as primitive but belief is not, perhaps because the functional nature of belief is easier to represent. Another would be to argue that our strongest problem intuitions arise from combining primitive relations with primitive qualities,

[17] Early in his book (p. 17) Graziano indicates that he intends the attention schema model as a way of dissolving the hard problem. "The answer may be that there is no hard problem. The properties of conscious experience—[...] the feeling, the vividness, the raw experienceness , and the ethereal nature of it [...]—these properties may be explainable as components of a descriptive model." After introducing the model, Graziano occasionally says that the model describes awareness as "ethereal", but he does not really explain why the model should represent awareness as ethereal or nonphysical. It should also be noted that Graziano (2017) resists describing his view as illusionist.

which happens in the perceptual case but not the belief case. A third response is to suggest that the primitive relation attributed in the perceptual case has some special properties: for example, it seems to be *presentational*, acquainting us directly with the quality it attributes, whereas the primitive relation attributed in the case of belief does not. Kammerer's resistance problem also arises here: a story needs to be told about why we find it much harder to deny that primitive mental relations are instantiated than that primitive external qualities are instantiated.

I move now to ideas about the meta-problem drawn from others with which I am less sympathetic.

8. Introjection and the phenomenological fallacy. U.T. Place (1956) diagnoses resistance to materialism as lying in the phenomenological fallacy: "the mistake of supposing that when the subject describes his experience, when he describes how things look, sound, smell, taste, or feel to him, he is describing the literal properties of objects and events on a peculiar sort of internal cinema or television screen". The phenomenological fallacy is closely related to the traditional sense-datum fallacy: the idea that when we have an experience of a red square, there must be some sort of internal red square sense-datum of which we are aware. If there were such sense-data, they would be hard to physically explain, so the fallacy (if we commit it) provides a potential explanation of problem intuitions.

An obvious objection is that many people explicitly reject the sense-datum fallacy, but their problem intuitions remain as strong as ever. On the face of it, an experience as of a red square raises the hard problem whether or not anything is red or square. Even if one is a representationalist who holds that one experiences represent a red square that may not exist, or a naive realist who holds that the experience is a direct perception of a red square in the external world, the hard problem seems as hard as ever. Why should the physical processes associated with perception and representation yield any experience at all? Perhaps Place could argue those who ask this question are still in the grip of the fallacy despite explicitly rejecting it. But I think it is more plausible that he has misdiagnosed the roots of our problem intuitions.

The phenomenological fallacy is an instance of what Avenarius called "introjection": roughly, perceiving something outside the head as being inside the head. Introjection has been used in various other ways to deflate problem intuitions. Frank Jackson (2003) suggests that we mistake intensional properties (e.g. experiences' representing redness) for instantiated properties (e.g. experiences' being red). Instantiated phenomenal properties would give rise to a hard problem, but mere representations of them do not. David Rosenthal (1999) suggests that when we "relocate" perceived qualities in the mind, we falsely infer that these qualities must always be conscious.

These moves are perhaps most promising for deflating the explanatory gap tied to qualities such as redness: if these qualities are merely represented or can occur unconsciously, they pose less of a gap. As before, however, the core of the hard problem is posed not by the qualities themselves but by our *experience* of these qualities: roughly, the distinctive phenomenal way in which we represent the qualities or are conscious of them. Recognizing the introjective fallacy for qualities does little to deflate the problem of explaining our experience of them.[18]

9. The user illusion. The centerpiece of Daniel Dennett's illusionism in recent years has been the claim that consciousness is a "user illusion", analogous to illusions generated when the user of a computer interacts with icons on a computer screen. The rough idea is that the icons provides a convenient way of representing the computer screen that greatly oversimplifies or falsifies the underlying reality: for example, there is not literally a folder anywhere in the computer. This is a nice statement of introspective-model illusionism, but as it stands it does not provide much guidance on the specific mechanisms of how the illusion of consciousness is generated.[19]

What is Dennett's account of problem intuitions? A general account is hard to find. One point that is clear in his recent book *From Bacteria to Bach and Back* is that he thinks the user illusion arose to facilitate communication (pp. 341-2), which he thinks is the most important use for self-monitoring. There are elements of introspective opacity in Dennett's repeated stress on the idea that we lack access to the details that underpin our representations. There are elements of primitive quality attribution about perception in his account of how we project apparently simple properties like sweetness and red stripes into the world. There are elements of the phenomenological-fallacy idea in his account of why we take it that if there is no red stripe in the world, there must be a red stripe in our mind. We have seen that all of these have limitations as accounts of problem intuitions, and Dennett's account is subject to the same limitations.

10. *The use-mention fallacy*: Advocates of the phenomenal concept strategy sometimes suggest that because thinking about consciousness is so different from thinking about brain states, we illegitimately infer that consciousness cannot be a brain state. This is a sort of use-mention error, since it involves mistaking a difference in our representations of an object for a difference in the

[18]Jackson and Rosenthal go on to try to explain conscious experience of these qualities in terms of distinctive functional manners of representation (Jackson) and higher-order thoughts (Rosenthal). These moves give rise to familiar explanatory gaps which need further tools to diagnose or deflate.

[19]Tor Norretranders' book *The User Illusion: Cutting Consciousness Down to Size*, which Dennett credits for the phrase, does not help much. Norretranders is mostly concerned with illusions about the self and about free will, and in particular the illusion that the conscious self is in control of our actions. He does not really try to argue that consciousness in general is an illusion.

object. In developing this idea, Loar (1990), Tye (1999), and Papineau (2002) all stress the fact that deploying a phenomenal concept itself has a sensory or imagistic phenomenology which is not involved in deploying a physical concept.[20]

This strategy requires a serious lack of charity concerning philosophers who have problem intuitions, as philosophers usually avoid use-mention errors like this quite easily. As Sundstrom (2008) points out, the strategy overgenerates to falsely suggest that we should not accept all sorts of identities that we in fact accept, such as my "pain is my brother's least favorite state". So I am inclined to set this strategy aside as one of the least promising explanations.

11. *Historical explanations*. I have also not said very much about historical explanations of the problem intuitions. In this area I have mostly focused on (evolutionary) design explanations, where the existence of problem intuitions follows from some sensible design choice in a cognitive system. I have touched briefly on evolutionary explanations from Dennett (in terms of communication). Nicholas Humphrey offers a different sort of evolutionary explanation: the illusion of consciousness makes life more worth living and so enhances the drive for survival. Non-evolutionary historical explanations are also available. Some might give genealogical accounts of problem intuitions in terms of accidents of cultural history. Perhaps we have all been over-influenced by Descartes, for example. Others might give psychoanalytic explanations, perhaps in terms of fear of death, or our yearning to be special. I am skeptical that explanations of this sort go deep enough, and I think that a design explanation is likely to be the most compelling, but a wide range of historical explanations are worth considering.

There are a number of further solutions to the meta-problem that I have not discussed. Humphrey (2014) proposes (in addition to his historical explanation mentioned above) that self-sustaining reentrant feedback loops involving internal representations in a high-dimensional space gives rise to the "illusion of extraordinary otherworldly properties". Fiala et al (2011) suggest that problem intuitions may arise from conflicting verdicts about consciousness from our fast (automatic) system and our slow (controlled) system in a dual-systems model. Molyneux (2012) argues that a robot would inevitably have problem intuitions due to a regress in making subjective-objective identifications. Drescher (2006) suggests (in addition to his "gensym" explanation of qualia intuitions mentioned above) that a "Cartesian camcorder" higher-order monitoring system explains why consciousness seems like an intrinsic property of mental events. I am skeptical about each of these proposed solutions for fairly predictable reasons, but I will not discuss the reasons here.

[20]The "stereoscopic fallacy" of Lycan (1987, p. 76) is a perceptual variant on this idea: seeing the brain of someone seeing red is not like seeing red, so seeing red is not a brain state.

To sum up what I see as the most promising approach: we have introspective models deploying introspective concepts of our internal states that are largely independent of our physical concepts. These concepts are introspectively opaque, not revealing any of the underlying physical or computational mechanisms. We simply find ourselves in certain internal states without having any more basic evidence for this. Our perceptual models perceptually attribute primitive perceptual qualities to the world, and our introspective models attribute primitive mental relations to those qualities. These models produce the sense of acquaintance both with those qualities and with our awareness of those qualities.

I hope that something like this is simultaneously (i) a reasonably plausible picture of how consciousness seems to us introspectively, and (ii) a reasonably well-motivated picture of how a well-designed cognitive system might represent its own states to itself. If so, then this approach might at least take us some distance toward a solution to the meta-problem of explaining our problem phenomenal intuitions in topic-neutral terms. Certainly this account is not a complete account, and I have indicated various challenges that still need to be answered, but an account like this is at least a start.

Ideally, potential solutions to the meta-problem can be tested both experimentally and with computational models. Experimentally, we can investigate human problem intuitions to see how well they conform to what a given proposal predicts. As discussed in the previous section, there is a small body of relevant experimental work as things stand. There is certainly room for much more work here, in principle yielding a serious research program in experimental philosophy and experimental psychology.

Computationally, we can build computational models that build in versions of the proposed mechanisms, and we can see whether these models reproduce something along the lines of human phenomenal reports. The only work along these lines that I know of is by Luke Muehlhauser and Buck Shlegeris, summarised by Muehlhauser (2017). They build a simple software agent using a theorem prover, based on principles that they attribute to Chalmers (1990; 1996) and Kammerer (2016). The system produces some simple reports that are structurally analogous to human phenomenal reports in certain respects. This is a very simple system with obvious limitations, but it suggests a research program. Using principled underlying mechanisms, we can attempt to build increasingly sophisticated system that exhibit human-like phenomenal reports with increasing scope and accuracy. If it is possible to build a reasonably accurate system of this sort, the mechanisms it uses will provide a candidate solution to the meta-problem.[21]

[21]Computational models such as these may bear on the occasionally-discussed idea of using phenomenal reports as

3 Philosophical Consequences

Suppose we have a solution to the meta-problem: a correct explanation of our problem intuitions in topic-neutral terms. What follows? In particular, what follows for scientific and philosophical theories of consciousness? Of course illusionism is one possible reaction, but there are many others.

In what follows, I will say that the *meta-problem processes* are the topic-neutrally characterized processes that explain phenomenal intuitions. For concreteness it may help to think of the meta-problem processes along the lines above: they involve introspective models with introspective concepts that attribute primitive mental states (such as primitive relations to primitive qualities) to ourselves when our brains are in complex cognitive states (such as perception, attention, or access consciousness, characterized topic-neutrally). I will call these cognitive states that drive the meta-problem *lower-order meta-problem states* (Frankish 2016 calls them "quasi-phenomenal" states, as these are the states that are misrepresented as phenomenal), and I will call the introspective states that attribute primitive properties *higher-order meta-problem states*.

We can then ask: if there is a solution to the meta-problem, involving meta-problem processes, what is the relationship between consciousness (that is, phenomenal consciousness, or subjective experience) and the meta-problem processes. A number of views are possible. I will discuss three broadly nonreductionist reactions, without any element of illusionism, and three broadly reductionist reactions, each with an element of illusionism.

1. *There is no solution to the meta-problem.* Some nonreductionists may embrace *meta-problem nihilism*: there is no solution to the meta-problem, or at least any solution will take the second horn, according to which there is no correct topic-neutral explanation of our problem intuitions. Something like this view might be taken on anomalous dualist views discussed earlier where consciousness plays a causal role that cannot be systematized in topic-neutral terms, or perhaps even by some anomalous materialist views where not all behavior can be explained in topic-neutral terms. It is far from clear how this would work, but there is at least room to investigate

a test for machine consciousness: roughly, if a machine behaves as if it is puzzled about consciousness, that is reason to think it is conscious. Versions of this idea include Sloman's "demanding new Turing test for robot philosophers" (2007), Argonov's "non-Turing test" for machine consciousness (2014), and Schneider and Turner's "artificial consciousness test" (2017). The Muehlhauser/Shlegeris model mirrors at least some aspects of our phenomenal reports, while being so simple that most people would deny that they are conscious. If this pattern continues with more developed computational approaches to the meta-problem, then we should probably be cautious about this sort of test for machine consciousness.

the possibility.

2. *Consciousness correlates with the meta-problem processes.* A second nonreductionist view is *meta-problem correlationism*, on which consciousness plays no causal role in the meta-problem processes, but it correlates with those processes. At least typically, when there is a phenomenal intuition generated by a first-order nonphenomenal state, there is a corresponding phenomenal state that renders the phenomenal intuition largely correct. On one version of the view, the phenomenal state will be present only when the phenomenal intuition (or some meta-problem process) is present, while on another view (preferable, I think) the phenomenal state will correlate with first-order states whether or not the meta-problem process is present.

An obvious problem for this view is that it seems to make our phenomenal intuitions correct as a matter of luck. If consciousness plays no role in generating the intuitions, it seems to be at best a coincidence that they are correct at all. A proponent of this view might respond by finding a deep underlying principle connecting first-order states to phenomenal states that makes the connection more than a coincidence. But as usual, there is work to be done.

3. *Consciousness realizes the meta-problem processes*. A third view available to nonreductionists is *meta-problem realizationism*, on which consciousness plays a role in realizing meta-problem processes. We saw earlier that theorists may hold that a topic-neutral explanation of phenomenal beliefs is correct but not complete, because consciousness realizes some of those processes, thereby playing a causal role with respect to their outcome. Perhaps panpsychist consciousness plays a role in physical dynamics. Perhaps interactionist consciousness plays a role in high-level dynamics. Meta-problem realizationism is also available to some reductionists. For example, some biological materialists may hold that consciousness is essentially biological and realizes computational processes that generate phenomenal intuitions. Likewise, some quantum-mechanical materialists may hold that consciousness is a quantum process that realizes the meta-problem processes.

On one version of the realizationist view to which I have some attraction, phenomenal consciousness realizes access consciousness. That is, wherever there is access consciousness functionally characterized, it is actually phenomenal consciousness that does the underlying causal work, either via the interactionist model or via the pan(proto)psychist model. This way phenomenal consciousness would serve as the basic cause of the processes that generate our phenomenal intuitions. At the same time, pictures of this sort have many challenges. It is by no means straightforward to see how consciousness could play precisely the role required, either on a panpsychist or an interactionist picture (or even on a biological or quantum picture). But there is at least room

to investigate this sort of possibility.

If realizationism is true, consciousness will not be causally irrelevant to our problem intuitions. Rather, consciousness will be a primary cause of those intuitions. More deeply, consciousness may be causally responsible for some key meta-problem processes. For example, our introspective models representing primitive properties may themselves be causally grounded in the presence of primitive properties of consciousness. These models of consciousness may also usefully serve as a simplified model of more complex physical processes of perception, attention, and representation, but consciousness itself will play a key role in the models. One may still worry about whether it plays a central enough role (perhaps because the structure of the processes may seem to explain our intuitions even without consciousness) but this view gives at least a promising start in integrating consciousness with the meta-problem processes.

4. *Phenomenal consciousness does not exist.* The first reductionist view is *strong illusionism*, which holds that consciousness itself is an illusion and does not exist. On the most obvious version of this view, consciousness is identified with the special primitive properties that are (or seem to be) attributed by our introspective models. No such special primitive properties are instantiated in our brains, so phenomenal consciousness does not exist. Our sense of being phenomenally conscious is an illusion.

5. *Consciousness is a lower-order meta-problem state*. On this view, phenomenal consciousness is identified with the cognitive states such as perception, attention, and access consciousness that serve as original target of the meta-problem processes. One might justify the view this way: (1) Phenomenal consciousness is what our introspective models are modeling, (2) These introspective models are really modeling access consciousness (albeit imperfectly), so (3) Phenomenal consciousness is really access consciousness (or perception, attention, or whatever).

This view will probably be a form of *weak illusionism*, on which phenomenal consciousness exists but some of our intuitions about it are illusions. For example, dualist and primitivist intuitions (consciousness is primitive and nonphysical) will be incorrect on this model, as will explanatory intuitions (consciousness cannot be physically explained). Depending on how the view is developed, the same may or may not be true for knowledge and conceivability intuitions.

6. *Consciousness is a higher-order meta-problem state.* On this view, consciousness is identified with certain meta-problem processes that attribute special states to ourselves. On this view, only creatures with certain introspective models will be phenomenally conscious. One might justify the view this way: (1) Phenomenal consciousness is the sense of being in special states, (2) This sense is identical to certain meta-problem states, so (3) Phenomenal consciousness is certain

meta-problem states.

This view will also lead to a sort of weak illusionism where at least our metaphysical and explanatory intuitions are false. It shares something in common with higher-order theories of consciousness, in that consciousness will involve certain higher-order representations of lower-order states. One obvious problem with this view is that it seems to involve a level confusion: on the face of it, consciousness is what our introspective models described earlier are *about*. But perhaps there is room for some terminological revisionism here.

4 Two arguments for illusionism[22]

Which of these six options is best? I will explore this by first discussing two ways of leveraging the meta-problem into an argument for illusionism: a debunking argument and a coincidence argument. This both clarifies the case for illusionism and also clarifies the best views for a nonreductionist to take in response to the meta-problem. In the following section, I will discuss the best views for an illusionist. To telegraph my conclusions, I think the most important views here are realizationism (for the nonreductionist) and strong illusionism (for the reductionist).

A simple way to leverage the meta-problem into an argument for illusionism is via a *debunking argument*. To put the general idea simply: if there is a broadly reductionist explanation of our nonreductionist beliefs about consciousness, nonreductionist beliefs will not be justified. In effect, the reductionist explanation of nonreductionist beliefs debunks our reasons to think that nonreductionist beliefs are correct. Something similar may go for our beliefs about phenomenal consciousness generally. There are various ways to lay out the argument more carefully, but one straightforward way is as follows:

> 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.
>
> ———
>
> 3. Our beliefs about consciousness are not justified.

[22] I may or may not end up splitting off the next few sections into a separate paper on illusionism.

Premise 1 is close to the claim that there is a topic-neutral explanation of our phenomenal intuitions, although there is a little daylight between them. Premise 2 is an instance of a general debunking principle: if there is an explanation of our beliefs about X that is independent of X, those beliefs are not justified. The conclusion is not exactly a statement of illusionism, but once it is accepted, illusionism is a natural consequence.

The argument is roughly analogous to debunking arguments that have been offered about god and morality. For example, debunking arguments about god argue that there is an explanation for beliefs in god that is independent of any gods, and use this to argue that our beliefs in god are unjustified. Debunking arguments about morality argue that there is an explanation for our moral beliefs that is independent of any objective moral truths, and use this to argue that our beliefs in objective moral truths are unjustified. Some principle like this is at work in the debunking arguments about god and morality mentioned above. One backing idea is that if the explanation of our beliefs about X is independent of X, then our beliefs about X will themselves be independent of X. If so, it will be entirely a matter of luck whether those beliefs are correct, so that the beliefs are not justified. Of course there is much to say about arguments of this form, about the underlying debunking principles, and about the precise sense of "independent" that might make the premises true.

What can a nonreductionist say in response? Some may reject premise 1 on the grounds that there is no solution to the meta-problem. I am not inclined to reject the argument on these grounds, but there are at least three other ways in which the argument can be rejected.

First: even if there is a solution to the meta-problem, premise 1 does not follow. A solution entails that there is a topic-neutral explanation of phenomenal intuitions, but it does not entail that there is such an explanation of phenomenal beliefs. I argued in Chalmers (2003) that consciousness plays a constitutive role in phenomenal beliefs (which are the objects of justification), so the explanation of those beliefs is not independent of consciousness. And it is phenomenal beliefs, not intuitions, that are objects of justification.

Second: premise 2 requires something like a causal account of justification, which is far from obvious where consciousness is concerned. On the view that I developed in *The Conscious Mind*, beliefs about consciousness are justified by our immediate acquaintance with consciousness, not by any causal background. As long as meta-problem processes do not undermine that acquaintance, they do not undermine our justification. So even if there is a causal explanation of our beliefs about consciousness in which consciousness plays no role, those beliefs may still be justified.

Third: there may be a sense of "independent" in which premise 1 is true (the explanation does

not mention consciousness), and a sense in which premise 2 is true (if consciousness plays no role in causing or constituting the beliefs, they are unjustified), but these are different senses. By analogy, there may be a brain-based or physics-based explanation of table-beliefs that does not mention tables, but as long as tables play a role in causing or constituting the beliefs, the beliefs may still be justified. So for the argument form to have a chance of being sound, both premise 1 and premise 2 must understand "independent of X" as something like "causally and constitutively independent of X" rather than as "does not mention X". However, a solution to the meta-problem does not guarantee an explanation of phenomenal beliefs that is causally and constitutively independent of consciousness. I have already discussed a way in which consciousness may play a constitutive role in the explanation. Furthermore, on the realizationist view discussed in the last section, consciousness will play a causal role in meta-problem processes. On this view, premise 2 (appropriately interpreted) will be false, and beliefs about consciousness may be justified.

I think any of these replies can block the debunking argument, but there remains unquestionably some discomfort in each of them. On all these views, it seems that at least an uncomfortably large part of the formation of our phenomenal beliefs can be explained without any role for consciousness, yielding a strange coincidence between our phenomenal intuitions and consciousness itself. One might use this discomfort to mount a *coincidence argument* for illusionism.

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. The correctness of phenomenal intuitions is not a coincidence.

———

4. Our phenomenal intuitions are not correct.

Because this argument concerns phenomenal intuitions (rather than beliefs) and concerns coincidence (rather than justification), the first and second objections to the previous argument do not really get a grip here. Premise 1 now just says that there is a solution to the meta-problem, and premises 2 and 3 have some prima facie plausibility.

Perhaps the most vulnerable premise is the second. There are a number of ways to try to reduce the sense of coincidence. A correlationist might argue that appropriate psychophysical

laws connecting processes and consciousness explain the apparent coincidence, and so remove any problematic coincidence. Still, it is hard to avoid the sense that on this view, it is lucky that the laws are as they are. For example, it can seem lucky that we are not in a zombie world with physical processes and phenomenal intuitions but no consciousness.[23] Likewise, it is arguably lucky that we are not in an inverted world where these physical processes yield quite different states of consciousness, such as pleasure where we feel pain. Perhaps there is always some luck in beliefs governed by laws of nature, but under correlationism, it seems that a very large amount of luck is required in order to ensure that there are just the states of consciousness to make our phenomenal intuitions correct.

More promisingly, a realizationist might argue that consciousness plays a causal role in explaining phenomenal intuitions, so their truth is not a coincidence. Again there may still be a sense of luck: if the meta-problem processes *might* have been realized without consciousness, it is perhaps lucky that they have been realized by consciousness. It is not clear that this weak sort of luck is objectionable, however. One finds something like it with most ordinary beliefs: for example, my belief that there is a table in front of me might have been caused by something other than a table. Still, more needs to be said to remove any sense of fortunate coincidence. Alternatively, perhaps one can develop a view where *only* consciousness could realize the relevant meta-problem processes, at least within certain constraints.[24]

All this brings out the strong pressure for any non-illusionist view of consciousness to integrate consciousness and meta-problem processes as closely as we can. I think the most promising view for reductionists and nonreductionists alike is realizationism. The research project for the realizationist is to spell out a satisfactory version of the view showing how consciousness realizes meta-problem processes in a way that removes the worries about debunking and about coincidence.

[23] Yudkowsky (2008) mounts a version of the argument from coincidence against this sort of view: "And yet this deranged outer [physical] Chalmers is writing philosophy papers that just happen to be perfectly right, by a separate and additional miracle. Not a logically necessary miracle (then the Zombie World would not be logically possible). A physically contingent miracle, that happens to be true in what we think is our universe, even though science can never distinguish our universe from the Zombie World.".

[24] For example, on Mørch's "phenomenal powers" view, phenomenal states are causal powers as part of their nature. On a strong version of this view, certain causal powers can essentially phenomenal powers, and the relevant causal roles could not be played without consciousness.

5 What sort of illusionism?

Strong illusionists deny that consciousness exists. *Weak illusionists* allow that consciousness exists, but say that it does not have certain crucial properties that it seems to have. For example, weak illusionists may hold that consciousness seems to be intrinsic, or nonphysical, or nonrepresentational, or primitive, or ineffable, or nonfunctional, but it is not.

In practice, there are more weak illusionists than strong illusionism, since strong illusionism is widely regarded as very implausible. Apparently paradigmatic illusionists such as Dennett, Graziano, and Humphrey have all tended to reject strong illusionism in favor of some sort of weak illusionism in recent years.

Still, Frankish (2012; 2016) has argued that illusionists who want to use illusionism to dissolve the hard problem of consciousness should be strong illusionists. I think he is correct about this, although my reasons are somewhat different. The basic reason, as I see it, is that the hard problem does not turn on the claim that consciousness is intrinsic, or nonphysical, or nonrepresentational, or primitive, and so on. For example, we can be agnostic about whether consciousness is intrinsic, or hold that it is extrinsic, and the hard problem arises as strongly as ever: why is it that when certain brain processes occur, there is something it is like to be us? The same goes for nonphysicality, nonrepresentationality, primitiveness, ineffability, and so on. Of course if the appearance that consciousness is nonphysical is an illusion, then consciousness is physical, and the letter of materialism is saved. But this does little to address the hard problem: we still have no explanation of why there is something it is like to be us.

To generate the hard problem of consciousness, all we need is the basic fact that there is something it is like to be us. We do not need further claims about intrinsicness, nonphysicality, and so on. So if an illusionist wants to reject this route to the hard problem, they need to deny that there is anything it is like to be us, or perhaps to hold that the whole idea of there being something it is like to be us is incoherent. But to do this is to deny that we are phenomenally conscious, or to hold that the whole idea of phenomenal consciousness is incoherent. And to do this is to be a strong illusionist. So to dissolve the hard problem of consciousness, illusionists need to be strong illusionists.

There is one sort of weak illusionism that may seem to escape this critique. This view allows that there is phenomenal consciousness, but only in the sense where phenomenal consciousness is understood as functionally: for example, perhaps phenomenal consciousness might be understood as whatever brings about our reports about consciousness. The hard problem turns crucially on

the claim that the concept of phenomenal consciousness is not a functional concept: that is, it is not a concept of bringing about certain behaviors and other cognitive consequences. This is what generates the gap between explaining behavioral functions and explaining consciousness. If phenomenal consciousness is a functional concept, the gap disappears.

I think this view is an important one, but it should be understood as a form of strong illusionism. The reason is that any plausible form of illusionism should allow that our *ordinary* concept of phenomenal consciousness is not a functional concept. Our ordinary concepts of phenomenal consciousness are *phenomenal concepts*, which are the central introspective concepts deployed in the meta-problem processes. The thesis that these concepts are not functional concepts is crucial to solving the meta-problem. If our ordinary concepts of consciousness were functional concepts, then there would be no hard problem of consciousness, or at least the problem would be much easier to dismiss. So any view that says there is phenomenal consciousness only in a sense where this is understood functionally is in effect a view where our ordinary (non-functionally-defined) concept of phenomenal consciousness does not refer. And that is a form of strong illusionism.

Something like this analysis can be applied to the three varieties of illusionism distinguished earlier. The first view, a form of strong illusionism, identifies consciousness with the primitive properties represented by meta-problem processes, and denies that they exist. The latter two views, which are forms of weak illusionism, identify consciousness with either lower-order or higher-order meta-problem processes, allowing that consciousness exists but it is not as it seems to be.

In a way, the choice between these options is verbal. All three views can allow that the primitive properties do not exist while allowing that higher-order and lower-order meta-problem processes exist. The three views just differ in which of these three they call "phenomenal consciousness".

At the same time, there is a natural constraint on what to call "phenomenal consciousness". As we have seen, phenomenal consciousness is what is picked out by phenomenal concepts, which are the central introspective concepts involved in meta-problem processes. These concepts purport to pick out primitive properties, and on the illusionist view no such primitive properties are instantiated. So it makes sense for illusionists to be strong illusionists, holding that phenomenal consciousness is not instantiated.

From this perspective, the version of weak illusionism where consciousness is identified with higher-order meta-problem states is especially unmotivated. It is extremely implausible that phenomenal concepts pick out these higher-order states. There is perhaps somewhat more motivation for the alternative version of weak illusionism where consciousness is identified with lower-order

meta-problem states, such as physical/functional states of perception, attention, or representation. Some illusionists may hold that although phenomenal concepts purport to pick out primitive properties, they in fact pick out lower-order meta-problem states, perhaps on the ground that these states are what phenomenal concepts are tracking in the actual world.

This lower-order variety of weak illusionism is most naturally seen as a sort of type-B materialism about consciousness, on which our concept of phenomenal consciousness is a non-functional concept, so that there is an epistemic gap between the physical and the phenomenal, but on which this concept picks out physical/functional properties, so that there is no ontological gap between the physical and the phenomenal. I have given extensive arguments against views of this sort elsewhere (e.g. Chalmers 2007, 2009), and I will not repeat those arguments here. Type-B materialism is a familiar philosophical strategy for dealing with the problem of consciousness, with familiar benefits and problems.

The really distinctive illusionist approach to the mind–body problem is instead a version of type-A materialism, on which there is no epistemic gap. The illusionist should allow that there seems to be an epistemic gap—that is, there seem to be phenomenal truths that are not deducible from physical truths—but that in fact this apparent gap is an illusion. Given the very plausible claim that our phenomenal concepts are not functional concepts, so that there are no a priori connections between physical and phenomenal concepts, it is natural for the type-A illusionist to cash out their position by saying that there are no phenomenal truths. Phenomenal consciousness seems to exist, but it does not exist.

This analysis of specific responses to the meta-problem coheres with the general analysis above. Insofar as illusionism is to be a distinctive way of dissolving the hard problem, the best form of illusionism is strong illusionism.

This is not to say that weak illusionism is false. In fact, I think some version of it is almost certainly true. For example, I think that visual consciousness initially seems to be fully detailed through the visual field, but this is an introspective illusion. But weak illusionism of this sort does not do much to dissolve the hard problem. The same goes for other forms of weak illusionism that I have discussed above. To make the hard problem itself into a sort of illusion, strong illusionism is required.

6 An argument against illusionism

This makes for a simple argument against illusionism, at least as a strategy for dissolving the hard problem.

1. If illusionism can dissolve the hard problem, strong illusionism is true.

2. Strong illusionism is false.

———-

3. Illusionism cannot dissolve the hard problem.

I have defended the first premise above, so it remains to defend the second. Some philosophers think that strong illusionism is incoherent. They hold that illusions are automatically experiences (phenomenally conscious states), so that if consciousness is an illusion, the illusion is itself an experience, so that there is phenomenal consciousness after all. I do not think strong illusionism is incoherent. The strong illusionist can simply understand illusions non-experientially as judgments, intuitions, or dispositions to report.

A nice illustration is provided by the "grand illusion" on which visual consciousness seems to be detailed throughout the visual field. We have the illusion that we have detailed conscious experiences all the way through. This illusion need not correspond to an experience of its own. It is simply a false judgment that need not be a phenomenally conscious state. We can think of strong illusionism as simply extending this illusionism about *some* apparent conscious experiences to *all* conscious experiences. We judge that there are experiences, when in fact there are not. Perhaps this view is implausible, but it is not incoherent.

Strong illusionism is not incoherent, but I think it is empirically false. I think the best argument against it is a simple Moorean argument, reminiscent of Moore's pointing to his hands to demonstrate that there is an external world.

1. People sometimes feel pain.

2. If strong illusionism is true, no one feels pain.

————————

3. Strong illusionism is false.

Premise 1 seems obviously true. Premise 2 follows from the dual claims that feeling pain is a conscious experience, and that illusionism denies that there are any conscious experiences.

At this point, a non-full-blooded illusionist might say they do not intend to deny that we feel pain. For example, they might say that that we feel pain in a nonphenomenal way or nonexperiential or nonconscious way. But this claim is of dubious coherence. In the ordinary sense of the word "feel", to feel pain is to experience pain. And when one feels pain in this sense, there is something it is like to undergo the pain, almost by definition.

I think a strong illusionist should really deny premise 1 (as Dennett did in his 1978 paper "Why You Can't Make a Computer that Feels Pain"). That is, they should deny that people ever feel pain, at least in any sense that entails that they experience pain. The illusionist can allow that at best people *undergo* processes of pain, and register them, but they do not experience pain, and they do not feel pain.

Of course to deny that people feel pain is to deny something apparently obvious. But I think it is of the essence of strong illusionism about consciousness to deny something apparently obvious – something so initially obvious that it seems undeniable. If the strong illusionist tries to avoid this route, they will not do justice to the strength of the intuitions that underlie the hard problem. For example, sophisticated illusionists may suggest that we feel pain in a weak (functional) sense but not a strong (phenomenal) sense. But crucially, the sense in which it is introspectively obvious that we feel pain is the phenomenal sense. In particular, it is the sense involving phenomenal concepts, which are the key concepts in our introspective self-models.

Strong illusionists about consciousness are committed to denying the central apparently introspectively obvious data about consciousness, and should not try to avoid it. If they do so, they will inevitably fail to dissolve the hard problem in the same way that weak illusionists maneuvers failed in the previous section. The moment one acknowledges that people genuinely feel pain (in the introspectively obvious sense), one faces the hard problem: why are physical pain processes accompanied by the feeling of pain? This is a central version of the hard problem. To dissolve it in the illusionist way, an illusionist should hold that the feeling of pain is an illusion.

Certainly, if I were a strong illusionist, I would deny that anyone ever feels pain. I would say that the experience of pain is an introspective illusion. When we seem to be experiencing pain, our brains are simply registering and negatively evaluating some states of one's body, with associated dispositions to change these states where possible. There is no experience of pain, and no feeling of pain. Experiences and feelings are simply states represented by misleading introspective models, and these states do not really exist.

That said, I think illusionism is obviously false, because it is obvious that people feel pain.

Around this point there is a familiar sort of dialogue:

Realist: People obviously feel pain, so illusionism is false.

Illusionist: You are begging the question against me, since I deny that people feel pain.

Realist: I am not begging the question. It is antecedently obvious that people feel pain, and the claim has support that does not depend on assuming any philosophical conclusions. In fact this claim is more obvious than any philosophical view, including those views that motivate illusionism.

Illusionist: I agree that it is obvious that people feel pain, but obvious claims can be false, and this is one of them. In fact, my illusionist view predicts that people will find it obvious that they feel pain, even though they do not.

Realist: I agree that illusionism predicts this. Nevertheless, the datum here is not that I find it obvious that people feel pain. The datum is that people feel pain. Your view denies this datum, so it is false.

Illusionist: My view predicts that you will find my view undeniable, so your denial simply confirms my view rather than opposing it.

Realist: I agree that my denial is not evidence against your view. The evidence against your view is that people feel pain.

Illusionist: I don't think that is genuine evidence.

Realist: If you were right, being me would be nothing like this. But it is something like this.

Illusionst: No. If "this" is how being you seems to be, then in fact being you is nothing like this. If "this" is how being you actually is, then being you is just like this, but it is unlike how being you seems to be.

And the dialogue goes on. Dialectically, the illusionist side is much more interesting than the realist side. Looking at the dialectic abstractly, it is easy to sympathize with the illusionist's debunking against the realist's foot-stamping. Still, as a conscious being reflecting on all the data, I think that the realist's side is the right one.

7 Conclusion

The meta-problem of consciousness is interesting not least because it is hard to avoid taking a position that others regard as crazy.

Here is Galen Strawson (2017) on strong illusionism, in a lecture entitled "One Hundred Years of Consciousness ('A Long Training in Absurdity'):

> There occurred in the twentieth century the most remarkable episode in the whole history of ideas—the whole history of human thought. A number of thinkers denied the existence of something we know with certainty to exist: consciousness, conscious experience.

Here is Eliezer Yudkowsky (2008) on nonreductionist realism about consciousness in light of the meta-problem (focusing especially in epiphenomenal property dualism):

> Based on my limited experience, the Zombie Argument may be a candidate for the most deranged idea in all of philosophy [...] According to Chalmers, the causally closed cognitive system of Chalmers's internal narrative is (mysteriously) malfunctioning in a way that, not by necessity, but just in our universe, miraculously happens to be correct. Furthermore, the internal narrative asserts "the internal narrative is mysteriously malfunctioning, but miraculously happens to be correctly echoing the justified thoughts of the epiphenomenal inner core", and again, in our universe, miraculously happens to be correct.

Of course there is middle ground between these views, but it tends to lead back to Scylla or Charybdis. There are certainly forms of weak illusionism, but these do not help a great deal with the hard problem. Versions that help with the hard problem need to deny the obvious, which is precisely what makes them seem absurd. On the other side, there are other forms of realism about consciousness, but most of these can be subjected to a weaker form of the same critique: once we can explain our conviction that consciousness exists without assuming that consciousness exists, the fact that the conviction is true seems somewhat miraculous.

I think that as things stand, neither illusionism nor realism has a truly satisfactory response to the charge of absurdity. Perhaps such a response can be found, but it will require major new ideas.

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 meta-problem processes are by their nature grounded in consciousness, even if it is metaphysically possible for them to occur without consciousness.

We do not have these explanations yet. If they can be developed, they might push us toward a satisfactory solution to the hard problem of consciousness. In the meantime, the meta-problem is a potentially tractable research project for everyone.

References

Argonov, V. 2014. Experimental Methods for Unraveling the Mind–body Problem: The Phenomenal Judgment Approach. *Journal of Mind and Behavior* 35:51-70.

Armstrong, D.M. 1968. The Headless Woman Illusion and the Defence of Materialism. *Analysis* 29:48-49.

Blackmore, S.J. 2002. The Grand Illusion: Why Consciousness Exists Only When You Look for It. *New Scientist* 174 (2348):26-29.

Balog, K. 2016. Illusionism's discontent. *Journal of Consciousness Studies* 23:40-51.

Balog, K. 2009. Phenomenal Concepts. In Brian McLaughlin, Ansgar Beckermann & Sven Walter (eds.), *Oxford Handbook in the Philosophy of Mind*. Oxford University Press. pp. 292–312.

Blackmore, S. 2016. Delusions of consciousness. *Journal of Consciousness Studies* 23:116-23.

Bloom, P. 2004. *Descartes' Baby: How the Science of Child Development Explains What Makes Us Human*. Basic Books.

Bourget, D. & Chalmers, D.J. 2014. What do philosophers believe?

Chalmers, D.J. 1987. Intelligent behavior and consciousness.

Chalmers, D.J. 1990. Consciousness and cognition.

Chalmers, D.J. 1996. *The Conscious Mind*. Oxford University Press/

Chalmers, D.J. 2003. The content and epistemology of phenomenal belief.

Chalmers, D.J. 2006. Perception and the fall from Eden.

Chalmers, D.J. 2007. Phenomenal concepts and the explanatory gap.

Chalmers, D.J. 2009. The two-dimensional argument against materialism.

Chalmers, D.J. 2013. Panpsychism and panprotopsychism.

Chudek, M., McNamara, R., Burch, S., Bloom, P., Henrich, J. 2013. Developmental and cross-cultural evidence for intuitive dualism. *Psychological Science* 20.

Clark, A. 2000. A case where access implies qualia? *Analysis* 60 (1):30-37.

Clark, A. 2001. Consciousness and the meta-hard problem. Appendix to *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford University Press.

Dennett, D.C. 1978. Why you can't make a computer that feels pain? In *Brainstorms*. MIT Press.

Dennett, D.C. 1992. *Consciousness Explained*. Little-Brown.

Dennett, D.C. 2015. Why and how does consciousness seem the way it seems? In (T. Metzinger & J. M. Windt, eds) *Open MIND*. Frankfurt am Main: MIND Group.

Dennett, D.C. 2016. Illusionism as the obvious default theory of consciousness. *Journal of Consciousness Studies*.

Drescher, G.L. 2006. *Good and Real: Demystifying Paradoxes From Physics to Ethics*. Bradford.

Fiala, B., Arica, A. & Nichols, S. 2011. On the Psychological Origins of Dualism: Dual-Process Cognition and the Explanatory Gap. In (E. Slingerland and M. Collard, eds.) *Creating Consilience: Integrating the Sciences and the Humanities*. Oxford University Press.

Frankish, K. 2012. Quining Diet Qualia. *Consciousness and Cognition* 21 (2):667-676.

Frankish, K. 2017. Illusionism as a theory of consciousness. *Journal of Consciousness Studies*.

Gottlieb, S. & Lombrozo, T. 2017. Can science explain the human mind? Intuitive judgments about the limits of science. *Psychological Science*.

Gray, H., Gray, K., & Wegner, D. 2007. Dimensions of mind perception. *Science* 315(5812): 619.

Graziano, M. *Consciousness and the Social Brain*.

Graziano, M. & Webb, T. 2014. A mechanistic theory of consciousness. *International Journal of Machine Consciousness* 6:163-176.

Graziano, M. 2016. Consciousness engineered, *Journal of Consciousness Studies* 23:116-23.

Hall, R.J. 2007. Phenomenal Properties as Dummy Properties. *Philosophical Studies* 135:199-223.

Hill, C. & McLaughlin, B. 1999. There Are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy *Philosophy and Phenomenological Research* 59:445-454.

Hofstadter, D.R. 2007. *I am a Strange Loop*. Basic Books.

Huebner, B. 2010. Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? *Phenomenology and the Cognitive Sciences* 9:133-55.

Humphrey, N. 2011. *Soul Dust: The Magic of Consciousness*. Princeton University Press.

Humphrey, N. 2016. Redder than Red Illusionism or Phenomenal Surrealism? *Journal of Consciousness Studies* 23:116-23.

Ismael, J. 1999. Science and the Phenomenal. *Philosophy of Science* 66:351-69.

Jackson, F. 2003. Mind and illusion. *Philosophy*.

Kammerer, F. 2016. The hardest aspect of the illusion problem - and how to solve it. *Journal of Consciousness Studies* 23 (11-12): 124-139.

Kammerer, F. forthcoming. Can You Believe It? Illusionism and the Illusion Meta-Problem. Philosophical Psychology:1-24.

Knobe, J. & Prinz, J. 2008. Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences* 7: 67-83.

Loar, Brian (1990/1997). Phenomenal states. Originally published in J. Tomberlin (ed.), Philosophical Perspectives, 4: Action Theory and Philosophy of Mind. Atascadero, CA: Ridgeview, 1990. Reprinted in revised form in N. Block et al. (eds.), The Nature of Consciousness: Philosophical Debates, Cambridge, Mass.: MIT Press, 1997.

Lycan, W. 1987. *Consciousness*. MIT Press.

Metzinger, T. 2003. *Being No One*. MIT Press.

Molyneux, B. 2012. How the Problem of Consciousness Could Emerge in Robots. *Minds and Machines* 22:277-97.

Mørch, H.H. forthcoming. Phenomenal powers panpsychism.

Muehlhauser, L. 2017. A Software Agent Illustrating Some Features of an Illusionist Account of Consciousness. OpenPhilanthropy [https://www.openphilanthropy.org/software-agent-illustrating-some-features-illusionist-account-consciousness]

Nagel, T. 1974. What is it like to be a bat? Philosophical Review, 83, 435-50.

Norretranders, T. 1991. *The User Illusion: Cutting Consciousness Down to Size*. Viking Penguin.

Papineau, D. 2007. Phenomenal and perceptual concepts. In T. Alter and S. Walter (eds.), Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism. Oxford: Oxford UP.

Papineau, D. 2002. *Thinking about Consciousness*. Oxford: Oxford UP.

Pereboom, D. 2011. *Consciousness and the Prospects of Physicalism*. Oxford University Press.

Peressini, A.F. 2014. Blurring two conceptions of subjective experience: Folk versus philosophical phenomenality. *Philosophical Psychology* 27:862-889.

Place, U.T. 1956. Is consciousness a brain process? *British Journal of Psychology* 47:44-50.

Rey, G. 1996. Towards a Projectivist Account of Conscious Experience. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand-Schoeningh-Verlag

Richert, R.A. & Harris, P.L. 2008. Dualism revisited: Body vs Mind vs Soul. *Journal of Cognition and Culture* 8:99-115.

Rosenthal, D. 1999. Sensory qualities and the relocation story. *Philosophical Topics*.

Schneider, S. & Turner, E. 2017. Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware. *Scientific American* blog. [https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/]

Schwarz, W. 2017. Imaginary foundations. [https://www.umsu.de/papers/imaginary.pdf]

Sloman, A. 2007. Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosophers. Association for Advancement of Artificial Inteligence. [https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai-consciousness.pdf]

Sundstrom, P. 2008. Is the mystery an illusion? Papineau on the problem of consciousness. *Synthese* 163; 133-43.

Smart, J.J.C. 2006. Metaphysical Illusions. *Australasian Journal of Philosophy* 3:167-175.

Systma, J. & Machery, E. 2010. Two conceptions of subjective experience. *Philosophical Studies* 151: 299-327.

Talbot, B. 2012. The Irrelevance of Folk Intuitions to the "Hard Problem" of Consciousness. *Consciousness and Cognition* 21: 644-650.

Tye, M. 1999. Phenomenal consciousness: the explanatory gap as a cognitive illusion. *Mind* 108: 705-25.

Yudkowsky, E. 2008. Zombies! Zombies? In *Rationality: From AI to Zombies*. [http://lesswrong.com/lw/p7/zombies_zombies/]