David J. Chalmers

The Meta-Problem of Consciousness

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 meta-theory is a theory about theories, 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 meta-problem is the problem of explaining why we think consciousness poses a hard problem, or in other terms, the problem of explaining why we think consciousness is hard to explain.

The relevant sort of consciousness here is phenomenal consciousness, or subjective experience.[2] A system is phenomenally conscious if there is something it is like to be that system, from the first-person point of view. It is widely (although not universally) accepted that most human beings are phenomenally conscious (at least when they are awake), that bats, cats, and other non-human animals may well be

Correspondence:
Email: chalmers@nyu.edu

[1] For comments, thanks to Richard Brown, Catarina Dutilh Novaes, Keith Frankish, Nick Humphrey, François Kammerer, Jackson Kernion, Uriah Kriegel, Tom McClelland, Luke McGowan, Kelvin McQueen, Luke Muehlhauser, Josh Weisberg, and David Yates. Thanks also to audiences at ANU, Kansas, Marist College, Medellin, Melbourne, Rutgers, Tiradentes, and the ASSC, Barcelona, and Tucson consciousness conferences.

[2] I use 'experience', 'conscious experience', and 'subjective experience' more or less interchangeably as synonyms for phenomenal consciousness (when used as mass nouns) or for phenomenal states (when used as count nouns). Phenomenal states are 'what-it-is-like' states: states individuated by what it is like to be in them (e.g. what it is like to see a certain shade of red or to feel a specific sort of pain). Phenomenal properties are 'what-it-is-like' properties: properties individuated by what it is like to have them.

Journal of Consciousness Studies, **25**, No. 9–10, 2018, pp. 6–61

phenomenally conscious, and that trees and rocks are not. A mental state is phenomenally conscious when there is something it is like to be in that state. It is widely accepted that seeing a bright red square and feeling pain are phenomenally conscious, and that one's ordinary background beliefs (my belief that Paris is in France, say, when I am not thinking about the matter) are not.

The hard problem of explaining phenomenal consciousness is one of the most puzzling in all of science and philosophy, and at the present time there are no solutions that command any sort of consensus. The hard problem contrasts with the easy problems of explaining various objective behavioural or cognitive 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 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.

On this analysis, the hard problem is hard because explaining consciousness requires more than explaining objective behavioural or cognitive functions. Even after we have explained all the objective 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? Because of this further question, the standard methods in the cognitive sciences have difficulty in gaining purchase on the hard problem.

There is nevertheless one behavioural function that has an especially close tie to the hard problem. This behavioural function involves *phenomenal reports*: the things we say about consciousness (that is, about phenomenal consciousness). More specifically, 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', 'Explaining behaviour does not explain consciousness', 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 behaviour. 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 behaviour 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 deeply connected 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 non-physical. 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* (1781/1999), 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 (1968a) 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 (2016) has called *illusionism* about consciousness: 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 typically hold that a solution to the meta-problem will itself solve or dissolve the hard problem. For example, if we have a physical explanation of why it seems to us that we have special non-physical properties, then those properties can be dismissed as an illusion, and any problem in explaining them can be dismissed as resting on an illusion. Views like this have been held or explored in recent years by philosophers such as Daniel Dennett, Frankish, and Derk Pereboom, and scientists such as Michael Graziano and Nicholas Humphrey.[3]

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) are so-called debunking arguments that rest on there being a solution to the meta-problem that explains our beliefs about consciousness without invoking consciousness. 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. I think that, if anything, illusionism has been under-explored in recent years. I take the view seriously, and I have more sympathy with it than with most materialist views.[4]

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.

I have long thought that solving the meta-problem might be a key to solving the problem of consciousness.[5] Even a non-illusionist can

[3] See Dennett (2016), Frankish (2016), Graziano (2013), Humphrey (2011), and Pereboom (2011) (although only the first two explicitly endorse illusionism). Some other recent illusionists may include: (philosophers) Clark (2000), Kammerer (2016), Rey (1996), Schwarz (forthcoming); (others) Argonov (2014), Blackmore (2002), Drescher (2006), Muehlhauser (2017).

[4] 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 programme 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 article within the framework of old battles, one might think of it as lending opponents a friendly helping hand.

[5] 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, ‘Consciousness and Cognition’ (Chalmers, 1990), proposed a test for theories of consciousness along lines of the meta-problem challenge discussed later in this article, holding that these theories should be able to explain reports about consciousness in a way that coheres with their explanation of consciousness itself. That article also proposed a solution to the meta-problem, which I developed further in *The Conscious Mind* (1996, pp. 184–8 and pp. 289–92) and which I used to motivate the information-based theory of consciousness that I favoured at that time.

reasonably hope both that there be an explanation of our judgments about consciousness and that this solution will give us insight into consciousness itself. Presumably there is at least a very close tie between consciousness and the mechanisms that generate reports about it. 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.

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 the community as a whole to understand the problem as the meta-problem and focus on solving it.[6] Theorists can then draw their own conclusions about what follows.

The meta-problem opens up a large and exciting empirical and philosophical research programme. 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 programme that combines at least (i) work in experimental philosophy, experimental psychology, linguistics, and anthropology studying subjects' judgments 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

[6] 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 (2001) and François Kammerer (2018a). Clark's 'meta-hard problem' is the problem of whether there is a hard problem of consciousness. Kammerer's 'illusion meta-problem' 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.

assessment of potential mechanisms, including how well they match up with and explain philosophical judgments about consciousness.

The meta-problem is a problem for scientists and philosophers alike, reductionists and non-reductionists alike, dualists and physicalists alike, illusionists and non-illusionists alike. For the most part, this paper will take an ecumenical approach that I hope will be interesting to people on all sides of these divides. I am most interested to explore the meta-problem as a problem in its own right. In the first half of the paper, after clarifying what the meta-problem involves, I will present and evaluate a number of potential solutions to the meta-problem. In the second half of the paper, I will explore how the meta-problem may impact scientific and philosophical theories of consciousness, including both realist and illusionist theories.

1. The Meta-Problem Research Programme

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.

1.1. What needs to be explained?

What exactly are the problem data that need explaining? They 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 really what we want to explain. 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*. I will also use the term to apply to the problem reports and judgments themselves. There may be more to the states ordinarily called intuitions than this, but it is plausible that they at least involve these dispositions and judgments. As I am using the term, problem intuitions can result from inferences, so that judgments that result from philosophical arguments will count as problem intuitions. Still, it is plausible that, in solving the meta-problem, the most important problem intuitions will be non-inferential judgments that arise prior to philosophical argument, and I will focus especially on judgments of this sort.

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 behavioural functions does not suffice to explain consciousness. Closely related are *metaphysical intuitions*, including dualist intuitions holding that consciousness is non-physical, 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 common intuition that robots are not conscious is a distribution intuition. There are *self intuitions* concerning the self or the subject of experience. There are *quality intuitions* concerning the special qualities (colours and the like) that are presented in experience, and *presentation* intuitions concerning the direct way they 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 children's 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, Gray and Wegner, 2007; Knobe and Prinz, 2008; Sytsma 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,[7] but for the most part the intuitions in question have not been the core problem intuitions.

[7] For example, Sytsma and Machery (2010) observe 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 this result is predicted by Chalmers (1996, p. 18), which observes that ordinary mental terms like this have both a functional reading and a phenomenal reading, with sensational terms such as 'pain' more likely to suggest a phenomenal reading than perceptual terms such as 'see'. I think many or most subjects have concepts of specific phenomenal states such as feeling pain or experiencing colour, but I am neutral on whether they also have a unifying concept of phenomenal consciousness. Certainly the most common problem intuitions (e.g. colour inversion intuitions) seem to involve these specific phenomenal concepts. Other relevant experimental work includes Huebner (2010), Talbot (2012), and Peressini (2014). Relatedly, Wierzbicka (2010) uses cross-linguistic analysis to

What about experimental research on the core problem intuitions? In principle there is room for experimental work on conceivability 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 (2018) 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.[8]

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 many cases of rejection, there is an underlying intuition that is psychologically outweighed by other forces (for example, an inclination towards dualism 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.[9]

argue that an Anglophone background has distorted analytic philosophers' discussion of experience and related concepts.

[8] One related empirical study is the PhilPapers Survey of professional philosophers (Bourget and Chalmers, 2014) — although this study is not really experimental, and most questions concern considered judgments rather than immediate intuitions. The survey found that 36% of the target group judge that zombies are conceivable but not metaphysically possible, 16% judge that they are inconceivable, and 23% judge that they are metaphysically impossible (with 25% agnostic or giving other answers). 56% endorsed physicalism about the mind while 27% endorsed non-physicalism about the mind.

[9] For what it's worth, I predict that knowledge intuitions will be somewhat more widespread than conceivability intuitions, and that explanatory intuitions will be somewhat more widespread than metaphysical intuitions. Intuitions about Mary-style cases and inversion case ('your green could be my red') may be particularly robust. But, as always, a great deal will depend on the way that key claims are formulated (this may be particularly difficult where conceivability arguments are concerned). Furthermore, the fact that someone denies a key claim (say, that consciousness is non-physical) is consistent with their having an underlying intuition that is outweighed.

As a first approximation, I will work under the assumption that these intuitions are widely shared, or at least that they have a widely shared basis. This is an empirically defeasible assumption, and I would be delighted to see empirical research (including cross-cultural, developmental, and historical research) that tests it.[10] Human intuitions and reports about the mind are plausibly produced by a combination of near-universal factors (e.g. mental states and introspective mechanisms that most humans share) and more variable factors (e.g. cultural, linguistic, and theoretical background, and other factors that vary with historical period and individual psychology). Variable factors will yield a great deal of variation in reports and intuitions, and may sometimes overwhelm the contribution of more universal factors. Still, my working assumption is that there are also near-universal factors that play a central underlying role in explaining problem intuitions where they are present.

Even if the assumption is false, the more limited task of explaining the intuitions in people who have them (presumably in terms of variable factors) 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.

1.2. What counts as an explanation?[11]

What sort of explanation counts as an explanation of problem intuitions, 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 have certain problematic features, and we notice that? In some contexts that may count as a reasonable explanation. Still, this sort of explanation is no longer neutral between realism and illusionism, and it threatens to turn the easy problem of explaining phenomenal intuitions back into a hard problem, because we will need to explain consciousness to explain the intuitions.

[10] In the related domain of intuitions about the transfer of selves between bodies, Chudek *et al.* (2013) give cross-cultural evidence against the hypothesis of 'culturally acquired dualism' and in favour of a more universal 'intuitive dualism'.

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

For our purposes, it is useful to put more constraints on a solution to the meta-problem. Earlier I suggested that a solution would involve a physical or functional explanation (roughly, one in terms of neural or computational mechanisms), but it is useful to impose a more general constraint.

I understand the meta-problem as the problem of explaining phenomenal reports 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, including representational, rational, historical, and structural explanations.

Representational explanation allows us to explain problem intuitions in terms of internal states or models that represent the subject or the world as having certain properties. Rational explanations explain processes as doing certain things because they are rational. It may be 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 a topic-neutral explanation.

Historical explanations are especially important. 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 evolution of phenomenal intuitions will be more satisfactory than one that does not. In any case, a complete solution to the meta-problem should say something about these historical and teleological questions.

Structural explanations allow the meta-problem to generalize to views where not all behaviour can be explained in physical terms. For example, consider an interactionist dualist view on which consciousness is non-physical and interacts with the brain. Descartes held that the non-physical mind drives brain processes via the pineal gland, while some contemporary interactionists hold that non-physical consciousness drives physical processes by collapsing a quantum wave function. On these views, one can still ask what explains problem reports and the like. Presumably the interactionist will hold that physical processes do not explain these reports, and that non-physical consciousness plays a central role in the explanation. But many interactionists will also be able to give structural explanations that do not mention consciousness.

For example, suppose that non-physical 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 behaviour in computational terms without ever mentioning consciousness. Or suppose that non-physical 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 behaviour in structural terms by saying that there is something that collapses the wave function in those circumstances, without ever saying that what does the collapsing is consciousness. In principle this would yield a topic-neutral explanation of problem reports.

This brings out an important point: a topic-neutral solution to the meta-problem does not require that consciousness is causally or explanatorily irrelevant. On the interactionist views just discussed, consciousness will play a causal role in generating behaviour (including problem reports), and a truly complete explanation of human behaviour will mention consciousness. Nevertheless, it will be possible to give a good (if not truly complete) explanation of human behaviour 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 behaviour, but it is nevertheless possible to give a computational explanation of human behaviour 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.

Structural explanations also apply to various forms of panpsychism and Russellian monism. On these views, consciousness or protoconsciousness at the fundamental level serves as the basis of the microphysical roles specified in physics. On these views, consciousness plays a causal role in generating human behaviour. It will nevertheless 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.

Some theorists may nevertheless hold that there is no good explanation of problem intuitions in topic-neutral terms. For example, there may be anomalous dualist views on which consciousness plays a completely unpredictable role, with effects that somehow depend on the intrinsic non-structural 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 explanation will always be available. Likewise, some anomalous monists and others might also argue against there being good physical explanations of behaviour, even though physicalism is true. To accommodate views like this, we can understand the meta-problem somewhat more generally as the problem of explaining problem intuitions in topic-neutral terms, or explaining why no such explanation is possible.

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 something 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 these 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 non-physical', '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 judgment involves an element of consciousness, or that all meaningful language is grounded in consciousness, so that it is impossible to explain genuinely meaningful reports or judgments in topic-neutral terms. If one holds this view, one should understand problem intuitions as involving something less than full-scale reports and judgments. Perhaps they are mere propensities to make certain noises and inscriptions (naturally interpretable as concerning consciousness). That will at least leave open the possibility that problem intuitions can be explained in topic-neutral terms.

For some purposes realists may want to relax the topic-neutrality constraint. To fully develop a realist view of consciousness, it may well be crucial to focus on full-scale phenomenal beliefs and to explain these things in terms of consciousness. Still, for my current purposes of understanding the meta-problem as an easy problem, I will mainly adopt the topic-neutral approach in what follows.

2. Potential Solutions to the Meta-Problem

It is time to face up to the meta-problem: explain our problem intuitions in topic-neutral terms, or explain why this is impossible.

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 programme in the area.[12]

The proposals I consider are not mutually exclusive, and many of them work well together. The first seven or so proposals are ideas that I find especially promising. I think all of these ideas 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 challenge that each of these ideas face, in order to

[12] Two other authors who in effect consider and criticize multiple solutions to the meta-problem include Bogardus (2013), who argues that proposals (discussed in what follows) due to Fiala, Nagel, Loar, and Papineau, and others concerning the source of our dualist intuitions do not succeed as defeaters for arguments for dualism, and Kammerer (2018a), who argues that proposals due to Graziano and Pereboom cannot solve his 'illusion meta-problem' discussed earlier.

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 colour and belief can help us to determine whether a proposed mechanism explains what is distinctive about the phenomenal case.

1. *Introspective models*. 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.[13]

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 it suggested that internal self-models will inevitably produce problem intuitions, but this seems 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

[13] Many attempts at solving the meta-problem give a role to introspective models. Introspective models are especially central in Graziano's (2013) '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 beliefs about phenomenal properties. Hofstadter (2007) also develops a sort of illusionism based in self-models and self-reference, primarily directed at illusions about the self.

problem intuitions as do experiences. Some people claim to have a non-sensory 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*. 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.[14]

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 the 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

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

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 really 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 reason 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 do not explain the problem intuitions. When we 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, and these two forms of imagination are independent of each other. This approach has been taken more generally by Hill and McLaughlin (1999) 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 certainly 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.* 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 non-physical. In the *locus classicus* of this approach, David Armstrong (1968a) 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'.[15]

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 non-physical. I might have no idea how the processes on my computer are implemented, but they do not seem non-physical in the way that consciousness does. Likewise, when I introspect my beliefs, they certainly do not seem physical, but they also do not seem non-physical in the way that consciousness does. Something special is going on in the consciousness case: in so far as consciousness seems non-physical, 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 non-physical.

5. *Immediate knowledge*. A related idea is that a cognitive system will treat its being in certain states as something it has immediate

[15] 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). An historical precursor is Thomas Hobbes in *De Corpore* (1655/1981, section 3.4): '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.' Note that the relevant topic-neutral claim is that physical mechanisms are opaque to introspection, not that consciousness itself is opaque to introspection. A non-materialist who endorses the revelation thesis discussed below may accept the first opacity thesis while denying the second.

knowledge of, 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 simply in that state. It is not given access to the mechanisms that bring the state about. If the system is a reasoner that requires reasons for its claims, it will be natural for it to hold that it has basic or immediate evidence that it is in this state. In effect, being in these states will at least seem to play a foundational role for the system, serving both as evidence for introspective judgments (I am in this state) and for further judgments about the world. Given these special states that seem to be immediately known, it is natural to think that these will then be represented by the system as primitive states that it finds itself in.[16]

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 problem intuitions nearly as strongly 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.

[16] My own proposed solution to the meta-problem in Chalmers (1990; 1996) gave immediate knowledge a central role. On the story there, introspective opacity naturally leads to a sense of immediate knowledge, which naturally leads to primitive quality attribution as below. Clark's (2000) analysis of the meta-problem builds on this approach, focusing on immediate access to the sensory modality involved in acts of detection. Sturgeon (1994) develops a version of the phenomenal concept strategy based on the idea that phenomenal concepts are concepts of states that serve as their own evidence. Schwarz (forthcoming) uses introspective opacity to motivate the introduction of illusory representations of sensory states to play a foundational role in Bayes-style belief updating.

Perhaps the best response to the belief problem appeals to an idea closely connected to the immediate knowledge idea: we have a sense that we are directly *acquainted* with conscious experiences and with their objects. It is arguable that this sense of acquaintance is missing in our immediate knowledge of our beliefs. If so, one could argue that it is the sense of acquaintance specifically, rather than that of immediate knowledge more generally, that plays the central role in generating our puzzlement about consciousness.

There are a couple of distinct elements to the sense of acquaintance. One is the sense of *presentation*: that we are somehow immediately presented with our experiences. Another is the sense of *revelation*: that the full nature of consciousness, and of various phenomenal properties, is fully revealed to us in introspection.[17] Both of these may well play a role in generating problem intuitions. Presentation (which might be understood as acquaintance with a property instance) is especially relevant to explaining our certainty that we are having experiences, and revelation is especially relevant to explaining our sense that they are irreducible. It remains to provide an explanation of why we have the sense of presentation and of revelation.

6. *Primitive quality attribution*. A promising proposal due to Derk Pereboom (2011) and others picks up on an analogy with the meta-problem of colour: roughly, why do colours seem to be irreducible qualitative properties? It is common to observe (e.g. Chalmers, 2006) that vision presents colours as special qualities of objects that are 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 colour 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. The perceptual system knows little about underlying physical properties, so it would be hard to represent colours 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 (colours), and when two objects are

[17] Numerous anti-materialists, such as Goff (2017), have identified a revelation thesis as the source of their anti-materialism, while numerous materialists, such as Lewis (1995), have identified a revelation thesis as the central intuition they need to reject.

similar with respect to how they affect the relevant parts of visual system, represent them as having the same qualitative property.[18]

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 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 colours 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 colour 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 colour. The 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 non-relational qualities, or 'qualia'. Such a view might have the resources for explaining why our problem intuitions differ from our intuitions about belief: sensory

[18] 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 means of representing our own more complex states can be found in Chalmers (1990; 1996), Dennett (2015), and in Schwarz (forthcoming). Hall (2007) introduces dummy properties to account for illusions about colours, but does not apply it to phenomenal properties.

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 colour. 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 colours and phenomenal properties. It is typically easy for people to accept that colours 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 (2018a) 'illusion meta-problem', which I will call the *resistance* problem to avoid confusion: explain why there is so much intuitive resistance to illusionism.[19] Any primitive quality attribution account will need further ideas to explain the disanalogy.

7. *Primitive relation attribution*. 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

[19] 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.

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 representing a red square (without being sure whether the square is present), the system may model 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 relations.

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 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' (2013). 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.[20] Still, his main focus is using a simple mental relation

[20] Early in his book (Graziano, 2013, 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 non-

(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 colour 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 modelled. 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 counter-intuitive). 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 us as counter-intuitive). 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 to 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, which happens in the perceptual case but not the belief case. Perhaps the most promising response is to appeal again to the sense of acquaintance: the primitive relation attributed in the perceptual case seems to acquaint us directly with the quality it attributes, and also itself to be an object of acquaintance, 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 extant ideas about the meta-problem that I am less inclined to endorse.

8. *Introjection and the phenomenological fallacy*. U.T. Place (1956) diagnoses resistance to materialism as lying in the phenomenological

physical. It should also be noted that Graziano (2016) resists describing his view as illusionist.

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's experiences represent a red square that may not exist, or a naïve 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 Richard Avenarius (1891) 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.[21]

9. *The user illusion*. The centrepiece 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 provide 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.[22]

What is Dennett's account of problem intuitions? An overarching account is hard to find. Instead, he has appealed to a collection of ideas over the years, most of which are discussed elsewhere on this list.[23] One distinctive thesis is that the user illusion arose to facilitate communication, 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 in his account of how we project apparently simple properties like sweetness and redness onto 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.

[21] 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. In a somewhat related strategy, Byrne (2006) suggests that we have mislocated the source of the problem: there is a hard problem of colour but no hard problem of consciousness (he deflates the latter by arguing that the transparency of experience suggests that we have no access to the nature of experience).

[22] Tor Norretranders' (1991) 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.

[23] For useful recent summaries, see 'Why and How Does Consciousness Seem the Way it Seems?' (Dennett, 2015), and chapter 14 of *From Bacteria to Bach and Back* (Dennett, 2017).

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 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.[24]

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 in critiquing Papineau's discussion of the fallacy (which he calls the 'antipathetic fallacy'), the strategy also over-generates 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 favourite state'.[25] So I am inclined to set this strategy aside as one of the least promising explanations.

11. *Underestimating the physical*. It is sometimes suggested that we are only impressed by the mind–body problem because we underestimate the body. That is, we have an overly thin conception of the physical, and that with a more adequate conception the mental–physical gap may disappear. One version of this view (e.g. Stoljar, 2001; Strawson, 2006) holds that we conceive of the physical in structural terms (perhaps in terms of the equations of physics), ignoring its intrinsic nature, which may have a close tie to consciousness. This path often leads to panpsychism and Russellian monism.

I agree that a structural conception of the physical is partly responsible for problem intuitions involving a gap between physical processes and consciousness. That said, there are many problem intuitions that do not take this form. For example, the key intuition in setting up the hard problem is that explaining functions does not explain consciousness. These intuitions concern the phenomenal

[24] 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.

[25] Papineau (2011) responds to Sundstrom, and Kammerer (2018b) responds in turn to Papineau. Kammerer (ms) gives a general critique of the idea that the explanatory gap arises from a fallacy of reasoning or a cognitive illusion.

rather than the physical and are not removed by enriching our conception of the physical. Explaining these intuitions at the heart of the meta-problem will require some other strategy.

12. *Historical and cultural explanations*. In principle, any synchronic explanation of the processes responsible for problem intuitions will be enriched by a historical explanation of how and why the processes came to be that way. Among historical explanations, I have so far mostly focused on evolutionary design explanations, where a mechanism connected to problem intuitions (e.g. introspective models or primitive quality attribution) follows from some sensible design choice in a cognitive system. I have also touched on Dennett's evolutionary explanations that tie consciousness to communication. Others have suggested that dualist intuitions arise from underlying psychological drives, such as the fear of death and the desire for survival. Nicholas Humphrey (2006) inverts this sort of explanation, suggesting that mind–body dualism and our sense that we are conscious give us a conviction that we have something special to preserve in ourselves and our children, thereby enhancing the drive for survival and reproduction.

Historical explanations in terms of culture 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, or misled by certain English-language constructions. My own working assumption is that, while cultural factors and psychological drives certainly play a role in shaping phenomenal intuitions, the underlying basis of these intuitions runs deeper. I think that evolutionary design explanations are potentially the most powerful historical explanations here, but a wide variety of historical and cultural explanations are worth considering.

13. *Other proposals*. There are a number of further potential solutions to the meta-problem that I have not discussed. Humphrey (2011) proposes (in addition to his historical explanation mentioned above) that self-sustaining re-entrant feedback loops involving internal states in a high-dimensional space give rise to the 'illusion of extraordinary otherworldly properties'. Fiala, Arico and Nichols (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 sceptical about each of these proposed solutions for fairly predictable reasons, but I will not discuss the reasons here.

Summary. 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. Our perceptual models perceptually attribute primitive perceptual qualities to the world, and our introspective models attribute primitive mental relations to those qualities. We seem to have immediate knowledge that we stand in these primitive mental relations to primitive qualities, and we have the sense of being acquainted with them.

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. The characterization is neutral between realism and illusionism. A realist can take it as a picture of how genuine consciousness seems to us introspectively. An illusionist can take it as a picture of the illusion of consciousness. Either way, this approach might 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. In some respects it is closer to a fleshed-out explanandum (what needs to be explained) than an explanans (a substantial explanation). Still, an analysis 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 programme 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, summarized 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 programme. 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.[26]

3. The Meta-Problem Challenge for Theories of Consciousness

The hard problem of consciousness and the meta-problem are closely connected. If we had a solution to the hard problem, we would expect it to shed some light on the meta-problem. In reverse, if we had a solution to the meta-problem, we should expect it to shed some light on the hard problem.

Illusionism provides one way in which the meta-problem could shed light on the hard problem. Illusionists think that a solution to the meta-problem will dissolve the hard problem. I will discuss this sort of illusionist strategy in the next section. For now, I will focus on consequences of the meta-problem for non-illusionist theories of consciousness according to which consciousness is real.

A plausible connection between the hard problem and the meta-problem for non-illusionist theories runs as follows. Whatever explains consciousness (the hard problem) should also play a central role in explaining our judgments about consciousness (the meta-

[26] Computational models such as these may bear on the occasionally-discussed idea of using phenomenal reports as 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 system mirrors at least some aspects of our phenomenal reports, while being so simple that most people would deny that it is 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.

problem). That is, suppose there is a correct non-illusionist theory of consciousness according to which mechanism M is the basis of consciousness. Then we can reasonably expect that mechanism M plays a central role in explaining our judgments about consciousness.

The reasoning here is simple: consciousness and judgments about consciousness are closely connected. For a realist, judgments about consciousness systematically reflect the character of consciousness. It would be extremely strange if the mechanisms responsible for consciousness played no role, or only a minimal role, in generating judgments about consciousness. These mechanisms may not *wholly* explain our judgments. It is plausible that some further introspective processes beyond consciousness itself are required to generate the judgment. Still, a realist should expect that our judgments about consciousness are the way they are *because* consciousness is the way it is, or at least because the basis of consciousness is the way it is. So the mechanisms responsible for consciousness should play a central role in generating judgments about consciousness.[27]

This provides what we might call the *meta-problem challenge* for theories of consciousness. If a theory says that mechanism M is the basis of consciousness, then it needs to explain how mechanism M plays a central role in bringing about judgments about consciousness.[28]

We can apply this challenge to existing theories. For example, according to integrated information theory (Tononi, 2007), integrated information is the basis of consciousness. Our principle then suggests that integrated information should play a central role in explaining judgments about consciousness. The meta-problem challenge asks:

[27] Why not say more straightforwardly: consciousness should play a central role in generating judgments about consciousness? I do this in part because many of these theories leave the metaphysical relationship between consciousness and the underlying mechanisms somewhat unclear: for example, there are versions of IIT and GWT that are consistent with materialism, dualism, and epiphenomenalism. These issues about the metaphysical status and causal role of consciousness are addressed in the following sections. In this section I am focusing on a more metaphysically neutral issue concerning mechanisms that arises even if consciousness is distinct from the underlying mechanisms.

[28] The meta-problem challenge is closely related to the 'coherence test' for theories of consciousness in Chalmers (1990): 'A completed theory of mind must provide (C1) An account of why we are conscious. (C2) An account of why we claim to be conscious; why we think we are conscious. Further (C3), accounts (C1) and (C2) must cohere with each other.'

how does integrated information explain judgments about consciousness?

The answer is far from obvious. On the face of it there is no obvious link between integration of information and these judgments. In fact, according to IIT, for every system with high integrated information there will be a computationally isomorphic simulated system with zero integrated information. The simulated system will be behaviourally identical to the original systems, making the same reports as we understand them, and more generally will be functionally identical, having the same intuitions. It is arguable that there is a computational explanation of the intuitions that applies equally to both the high-phi system and the zero-phi system. It is also not clear why high-phi systems (e.g. those with strongly integrated sensory modalities) should be more likely to make the reports than intermediate-phi systems (e.g. those with weakly integrated sensory modalities). If so, it looks as if high integrated information plays no essential role in explaining phenomenal intuitions. If that is right, then it provides at least some reason to doubt whether integrated information is the basis of consciousness.

Of course this is not a knockdown argument against IIT, and there are various ways in which an IIT advocate might respond. But it provides at least a *prima facie* challenge for IIT to address. A similar challenge arises for biological theories of consciousness (e.g. Block, 2009) that allow that there could be functional duplicates of conscious beings (made of silicon, for example) that are not conscious. As with IIT, these duplicates will share our intuitions, and it looks as if specifically biological mechanisms will not play an essential role in explaining the intuitions.[29]

One could apply the same challenge to global workspace theories (Baars, 1988), according to which the basis of consciousness is a global workspace that makes information available to other systems in the brain. Here the challenge asks: how does the global workspace help to explain our judgments about consciousness? A version of the same challenge applies to many first-order representationalist theories

[29] One response for the advocate of these theories is to advocate a version of the realizationism discussed in the next section: mechanism M is not necessary for the reports but, at least in systems with M, M causes the reports. Perhaps this will work for biological theories, but in the case of IIT it is far from obvious how high phi even causes the reports. These theories will also have to say something about the problems for realizationism discussed below.

of consciousness (e.g. Dretske, 1995), which typically hold that conscious states are those representational states that are globally available or that are poised for the control of action and reasoning in some way.

Global workspace and first-order representational theories can at least begin to answer the challenge. For example, it is plausible that information is only reported at all if it is globally available, giving an initial connection between the workspace and reports. Still, it is not obvious how the workspace explains problem reports involving a sense that consciousness is puzzling. So there is at least more for a workspace theorist or a first-order representationalist to say here.

Higher-order thought theories (Rosenthal, 2002) say that conscious states are those that are the objects of higher-order thoughts. Here the meta-problem challenge asks: how do higher-order thoughts explain problem reports? Now, of course higher-order thoughts are involved in any introspective process, so they play a central role in explaining how we explain any mental states. But as Alvin Goldman (2000) pointed out, it is not clear how mere higher-order thoughts explain why we report mental states as being *conscious* (*prima facie*, that requires having higher-order thoughts about higher-order thoughts), and likewise it is not obvious how higher-order thoughts explain why we report conscious states as puzzling. So again there is a challenge to answer here.[30]

We can apply a similar challenge to all sorts of theories. Consider quantum theories (Hameroff and Penrose, 1996; Stapp, 1993) that say there is a strong tie between wave-function collapse and consciousness. Does wave-function collapse play a central role in explaining reports of consciousness? One might worry that the answer is no, since wave-function collapse only selects one of multiple branches of the wave function. If a subject says 'I am conscious' in the selected branch, it is arguable that the subject also says 'I am conscious' in

[30] An interesting variant of higher-order thought theory that is driven by the meta-problem is *primitive higher-order thought* theory, which says that conscious states are the objects of *primitive* higher-order thoughts. A primitive higher-order thought is one that attributes primitive qualities and/or relations to a lower-order state, as in the primitive quality/relation accounts discussed earlier. Another version is *dispositional primitive higher-order thought theory*, on which a mental state is conscious if it is disposed to bring about primitive higher-order thoughts. These views are closely related to the versions of weak illusionism discussed later.

many unselected branches. If so it looks as if there may be an explanation of the reports which is prior to wave-function collapse.

The challenge even applies to panpsychist theories (e.g. Brüntrup and Jaskolla, 2017). These theories typically hold that human consciousness is some sort of combination of micro-consciousnesses in fundamental entities. The combination problem for panpsychism is to explain how micro-consciousnesses can combine to yield our consciousness. In light of the meta-problem challenge, there is a further aspect of the combination problem: explain how these combination states play a central role in bringing about reports of consciousness.

One philosophical theory that has an especially hard time meeting the meta-problem challenge is analytic functionalism (e.g. Armstrong, 1968b), according to which all of our mental concepts, including our concepts of consciousness, are functional concepts (concepts playing certain functional roles, such as discrimination, integration, report, and the like). Nagel (1970) pointed out that Armstrong's analytic functionalism 'leaves it a complete mystery why [the problem of other minds] has ever bothered anyone'. If analytic functionalism is right, the problem of other minds is conceptually equivalent to the problem of knowing that others have the relevant functional states, which does not bother us in the same way. Nagel's point applies equally to the mind–body problem. If analytic functionalism were true, the hard problem would be conceptually equivalent to the easy problems, and *prima facie* we would not be any more bothered by the hard problem than by the easy problems. As a result, analytic functionalism seems to be psychologically inadequate in addressing the meta-problem.

What sort of realist theory of consciousness does especially well where the meta-problem challenge is concerned? My own tentative view is that the most promising solution to the meta-problem lies in primitive relation attribution and the sense of acquaintance: our experiences seem to primitively acquaint us with qualities in the environment, and these experiences are themselves objects of acquaintance. I favour a realist theory of consciousness where consciousness does in fact involve acquaintance in this way. This line tends to suggest a combination of a first-order representational view of consciousness (consciousness involves immediate awareness of worldly properties) with a self-representational view of consciousness (consciousness involves immediate awareness of itself).

I do not think this sort of awareness is reducible to brain mechanisms, but one might expect some sort of corresponding structure at the level of brain mechanisms. To speculate: one might expect a corresponding structure of first-order representational states that are themselves objects of a simple sort of representation, or that are at least disposed to be objects of representation. Such a view could be developed in a higher-order way, perhaps bringing in a version of the dispositional primitive higher-order thought view mentioned earlier. My own inclination would be to develop this view in a first-order way, by finding a first-order property F that grounds the higher-order representation. First-order states with F will be disposed to be objects of primitive higher-order representation (in the presence of appropriate introspective systems), but can also exist without any higher-order representation (in the absence of appropriate introspective systems). If there is such a first-order property F, it would play a key role in solving the meta-problem and would be a strong candidate to be the physical basis of consciousness itself. The residual question is then to find out just what property this might be.[31]

4. Six Reactions to the Meta-Problem

I now step back from specific theories of consciousness and specific solutions to the meta-problem to think about the issues at a greater level of generality, looking at general metaphysical options that arise in light of the meta-problem.

Let us 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 higher-order models with introspective concepts that attribute special mental states (such as primitive relations to primitive qualities) to ourselves when our brains are in certain lower-order cognitive states (such as perception, attention, or access consciousness, characterized topic-neutrally). I will call these lower-order 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

[31] In 'Consciousness and Cognition' (1990) and *The Conscious Mind* (1996), I argued in effect that any first-order informational state could play this role in grounding primitive higher-order representations (in the presence of appropriate introspective systems). That is, any informational state might seem to an introspecting subject to be a primitive qualitative state. If so, the property of being informational would be the key to the meta-problem and a potential basis of consciousness. An obvious objection arises from the case of belief, however, which suggests that more constraints may be required.

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 non-reductionist reactions, which need not involve 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 non-reductionists may embrace *meta-problem nihilism*: there is no solution to the meta-problem. Alternatively, if we understand the meta-problem more broadly as 'Explain our problem intuitions in topic-neutral terms, or explain why this is impossible', then the meta-problem nihilist says that any solution must take the second horn. As discussed earlier, a version of this view might be taken by anomalous dualists and anomalous materialists for whom behaviour and/or the causal role of consciousness cannot be systematized in topic-neutral terms. It is far from clear how this would work, but there is at least room to investigate the possibility.

2. *Consciousness correlates with the meta-problem processes*. A second non-reductionist 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 non-phenomenal 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, to be discussed in more depth shortly, 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 non-reductionists 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, not least 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 the original target of the meta-problem processes. One might justify the view this way: (1) phenomenal consciousness is what our introspective models are modelling, (2) these introspective models are really modelling 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 non-physical) 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.

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

5. Debunking Arguments for Illusionism

The meta-problem can be leveraged into an argument for illusionism via a *debunking argument*. A debunking argument is roughly an argument that starts from a premise about how beliefs about a domain are formed and concludes that beliefs in that domain are not justified or reliable. For example, Richard Joyce (2006) argues from the claim that moral beliefs are formed because of evolutionary pressures to the conclusion that moral beliefs are not justified. Sometimes a debunking argument argues against certain theories in a domain by making a case that *if* those theories are true, beliefs in that domain are not justified or reliable. For example, Sharon Street (2006) uses an evolutionary analysis of moral beliefs to argue against moral realism, on the grounds that if moral realism were true, our moral beliefs would be unreliable.[32]

In the case of consciousness, one could make a debunking argument against non-reductionist views of consciousness, as follows: if there is a broadly reductionist explanation of our non-reductionist beliefs

[32] As far as I know, there has not been much explicit formulation and discussion of general debunking arguments regarding consciousness. Many illusionists endorse both the premises and conclusions of debunking arguments, but they have not focused on arguments from the premises to the conclusions. Some theorists have tried to debunk specific theoretical beliefs about consciousness (e.g. materialists debunking beliefs in dualism) without debunking beliefs about consciousness in general. One relevant general argument is formulated by Sydney Shoemaker in 'Functionalism and Qualia' (1975). Shoemaker argues for functionalism by arguing that if non-conscious functional duplicates of us with the same beliefs were possible, we could not know we are conscious. Shoemaker's premises concern possibility (e.g. the possibility of absent qualia) or about causation (consciousness plays no relevant causal role) rather than explanation, but there is at least a flavour of debunking. In *The Conscious Mind* (1996, pp. 192–3) and in 'The Content and Epistemology of Phenomenal Belief' (2003), I formulate and reply to related debunking arguments that bring in an underlying premise about the explanatory irrelevance of consciousness. I am interested to hear of other sources, both recent and historical.

about consciousness (as the meta-problem may suggest), non-reductionist beliefs will not be justified. In effect, the reductionist explanation of non-reductionist beliefs debunks our reasons to think that non-reductionist beliefs are correct.

One can also make a debunking argument about beliefs about phenomenal consciousness in general, perhaps with some variety of non-reductionism operating as a background assumption. There are various ways to lay out such an argument, but perhaps the most straightforward 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.

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 the two. 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. (Of course much depends on what 'independent' means; I will return to that matter shortly.) 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. 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 non-reductionist 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, regarding premise 1: if there is a solution to the meta-problem, it follows that there is a topic-neutral explanation of phenomenal intuitions, but it does not follow that there is a topic-neutral 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, regarding premise 2: this arguably 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, regarding both premises: one can argue that there is a sense of 'independent' in which premise 1 is true and a sense in which premise 2 is true, but these are different senses. The sense in which premise 1 follows simply from a solution to the meta-problem is what we might call *descriptive independence*: the explanation doesn't mention consciousness. If one endorses the possibility of zombies (a further non-reductionist commitment), we also have *modal independence*: the elements of the explanation could obtain without consciousness. However, descriptive and modal independence do not make premise 2 plausible. There are brain-based explanations of our beliefs about tables that do not mention tables, and that could occur in the absence of tables, but as long as tables in fact cause the beliefs, the beliefs seem well-enough grounded in tables to avoid debunking. Similarly, there are physics-based explanations of our beliefs about tables that do not mention tables, but as long as the physical processes in question *constitute* tables, this descriptive independence does not debunk table beliefs. To be plausible, premise 2 requires something more: perhaps a combination of *causal independence* (consciousness plays no causal role in the processes invoked in the explanation) and *constitutive independence* (consciousness does not constitute and is not constituted by elements of the explanation).

A solution to the meta-problem does not guarantee an explanation of phenomenal beliefs that is causally and constitutively independent of consciousness. In the first response above I have already mentioned one way in which consciousness may play a constitutive role in the explanation: consciousness may constitute phenomenal beliefs themselves. Furthermore, on the realizationist view discussed in the last section, consciousness will play a causal role in meta-problem processes. If so, then we do not have any sense of 'independent' on which premise 1 follows from a solution to the meta-problem (and/or non-reductionism) and in which premise 2 is plausible, so the argument does not go through.

I think any of these replies can block the debunking argument. Still, there unquestionably remains some discomfort in the vicinity. 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 related *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 (at least if independence is understood as descriptive independence), and premises 2 and 3 have some *prima facie* plausibility. As before, though, much depends on what is meant by 'independent', and a version of the third objection may apply.

As before, a solution to the meta-problem guarantees only descriptive independence and perhaps modal independence in premise 1, while the case of tables suggests that these varieties of independence are not enough to justify premise 2. As long as we have causal or constitutive dependence of phenomenal intuitions on consciousness, the sense of coincidence may be removed. As before a realizationist might argue that consciousness plays a causal role in explaining phenomenal intuitions, so their truth is not a coincidence. Even more weakly, a correlationist may argue that *nomic* dependence is enough to avoid coincidence: psychophysical laws connecting processes and consciousness explain the correctness of our intuitions.

Even for these views, however, it is hard to avoid a sense of coincidence entirely. As long as we have modal independence, so that the meta-problem processes *could* have come apart from consciousness, it can seem lucky that they have not. Where psychophysical laws are concerned, it seems lucky that the laws are as they are. Only this luck ensures that we are not in a zombie world with physical processes and phenomenal intuitions but no consciousness, or in an inverted world where these processes yield pleasure when we feel pain. Where realization is concerned, it seems lucky that the meta-problem processes are in fact realized by consciousness rather than by something else. As before it is not obvious that modal independence is always objectionable in this way: in the case of tables, the mere possibility that our table-beliefs could occur without tables need not make the truth of these beliefs coincidental. And it is arguable that there is always some luck in beliefs governed by laws of nature. Nevertheless, in the case of consciousness one has a very strong sort of modal independence, in that there seems to be a *near-complete* structural explanation of the intuitions that could obtain without consciousness. It is easy to get the sense that what really explains the intuitions is the structure of cognitive processes, and the fact that consciousness is connected to that structure is something of a fortunate and optional extra. More needs to be said to remove any sense of fortunate coincidence.[33]

One way to go further is to develop a view where *only* consciousness could realize the relevant meta-problem processes, perhaps given certain constraints. For example, on the ‘phenomenal powers’ view put forward by Mørch (2018), phenomenal states are causal powers as part of their nature. On a strong version of this view, certain causal powers are essentially phenomenal powers, and the relevant causal

[33] For coincidence arguments based on modal independence, see Latham (1998) and Yudkowsky (2008). See also Shaffer (1964) for a coincidence argument against parallelism.

roles could not be played without consciousness. On a view like this, it need not be a coincidence that the relevant judgments are brought about by consciousness.

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 non-reductionists 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.

6. 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 non-physical, or non-representational, or primitive, or ineffable, or non-functional, but it is not.

In practice, there are more weak illusionists than strong illusionism, since strong illusionism is widely regarded as very implausible. A number of illusionists (e.g. Graziano, 2016; Humphrey, 2016; forthcoming) have explicitly rejected strong illusionism in favour 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 non-physical, or non-representational, 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 non-physicality, non-representationality, primitiveness, ineffability, and so on. Of course if the appearance that consciousness is non-physical 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, non-physicality, 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 sort of weak illusionism allows that there is phenomenal consciousness, but only in the sense where phenomenal consciousness is understood 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 behaviours and other cognitive consequences. This is what generates the gap between explaining behavioural 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.

At one level, 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 in other work (e.g. Chalmers, 2007), and I will not repeat them 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. In so far 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.

7. 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 (see e.g. Blackmore, 2002). 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 we feel pain in a non-phenomenal way or non-experiential or non-conscious 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 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 as central a version of the hard problem as any, and it can be posed using the introspectively obvious datum alone. 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 unbelievable, 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.
>
> Illusionist: 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, reflecting on all the data, I think that the realist's side is the right one.

8. 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 non-reductionist realism about consciousness in light of the meta-problem (focusing especially on epiphenomenal property dualism):

> Based on my limited experience, the Zombie Argument may be a candidate for the most deranged idea in all of philosophy… I do not see any way to evade the charge that, on Chalmers's own theory, this separable outer Chalmers is deranged. This is the part of Chalmers that is the same in this world, or the Zombie World; and in either world it writes philosophy papers on consciousness for no valid reason. Chalmers's philosophy papers are not output by that inner core of awareness and

> belief-in-awareness, they are output by the mere physics of the internal narrative that makes Chalmers's fingers strike the keys of his computer. And yet this deranged outer Chalmers is writing philosophy papers that just happen to be perfectly right, by a separate and additional miracle.

Of course there is middle ground between denying consciousness and saying that consciousness is epiphenomenal, but the middle ground tends to lead back to Scylla or Charybdis. On the deflationary side, there are certainly forms of weak illusionism, but these do not help much 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 realist side, there are non-epiphenomenalist 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**,pp. 51–70.

Armstrong, D.M. (1968a) The headless woman illusion and the defence of materialism, *Analysis*, **29**, pp. 48–49.

Armstrong, D.M. (1968b) *A Materialist Theory of the Mind*, London: Routledge and Kegan Paul.

Avenarius, R. (1891) *Der menschliche Weltbegri*, Leipzig: Reisland.

Baars, B. (1988) *A Cognitive Theory of Consciousness*, Cambridge: Cambridge University Press.

Balog, K. (2009) Phenomenal concepts, in McLaughlin, B., Beckermann, A. & Walter, S. (eds.) *Oxford Handbook in the Philosophy of Mind*, Oxford: Oxford University Press.

Blackmore, S.J. (2002) The grand illusion: Why consciousness exists only when you look for it, *New Scientist*, **174** (2348), pp. 26–29.

Block, N. (2009) Comparing the major theories of consciousness, in Gazzaniga, M. (ed.) *The Cognitive Neurosciences IV*, Cambridge, MA: MIT Press.

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

Bogardus, T. (2013) Undefeated dualism, *Philosophical Studies*, **165**, pp. 445–466.

Bourget, D. & Chalmers, D.J. (2014) What do philosophers believe?, *Philosophical Studies*, **170**, pp. 465–500.

Brüntrup, G. & Jaskolla, L. (2017) *Panpsychism: Contemporary Perspectives*, Oxford: Oxford University Press.

Byrne, A. (2006) Color and the mind–body problem, *Dialectica*, **60**, pp. 223–244.

Chalmers, D.J. (1987) Consciousness: The first-person and third-person views, [Online], http://consc.net/papers/oxford1.pdf.

Chalmers, D.J. (1990) Consciousness and cognition, [Online], http://consc.net/papers/c-and-c.html.

Chalmers, D.J. (1996) *The Conscious Mind*, New York: Oxford University Press.

Chalmers, D.J. (2003) The content and epistemology of phenomenal belief, in Smith Q. & Jokic, A. (eds.) *Consciousness: New Philosophical Perspectives*, Oxford: Oxford University Press.

Chalmers, D.J. (2006) Perception and the fall from Eden, in Gendler T. & Hawthorne, J. (eds.) *Perceptual Experience*, Oxford: Oxford University Press.

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

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), pp. 30–37.

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

Dennett, D.C. (1978) Why you can't make a computer that feels pain, *Brainstorms*, Cambridge, MA: MIT Press.

Dennett, D.C. (1992) *Consciousness Explained*, Boston, MA: Little-Brown.

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

Dennett, D.C. (2016) Illusionism as the obvious default theory of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 65–72. Reprinted in Frankish,

K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

Dennett, D.C. (2017) *From Bacteria to Bach and Back: The Evolution of Minds*, New York: W.W. Norton.

Drescher, G.L. (2006) *Good and Real: Demystifying Paradoxes From Physics to Ethics*, Cambridge, MA: Bradford Books.

Dretske, F. (1995) *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Fiala, B., Arico, A. & Nichols, S. (2011) On the psychological origins of dualism: Dual-process cognition and the explanatory gap, in Slingerland, E. & Collard, M. (eds.) *Creating Consilience: Integrating the Sciences and the Humanities*, Oxford: Oxford University Press.

Frankish, K. (2012) Quining diet qualia, *Consciousness and Cognition*, **21** (2), pp. 667–676.

Frankish, K. (2016) Illusionism as a theory of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 11–39. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

Goff, P. (2017) *Consciousness and Fundamental Reality*, Oxford: Oxford University Press.

Goldman, A. (2000) Can science know when you are conscious?, *Journal of Consciousness Studies*, **7** (5), pp. 3–22.

Gottlieb, S. & Lombrozo, T. (2018) Can science explain the human mind? Intuitive judgments about the limits of science, *Psychological Science*, **29**, pp. 121–130.

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

Graziano, M. (2013) *Consciousness and the Social Brain*, New York: Oxford University Press.

Graziano, M. (2016) Consciousness engineered, *Journal of Consciousness Studies*, **23** (11–12), pp. 116–123. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

Hall, R.J. (2007) Phenomenal properties as dummy properties, *Philosophical Studies*, **135**, pp. 199–223.

Hameroff, S.R. & Penrose, R. (1996) Conscious events as orchestrated space-time selections, *Journal of Consciousness Studies*, **3** (1), pp. 36–53.

Hill, C. & McLaughlin, B. (1999) There are fewer things in reality than are dreamt of in Chalmers's philosophy, *Philosophy and Phenomenological Research*, **59**, pp. 445–454.

Hobbes, T. (1655/1981) *De Corpore*, in Martinich, A.P. (trans.) *Part I of De Corpore*, Norwalk, CT: Abaris Books.

Hofstadter, D.R. (2007) *I am a Strange Loop*, New York: Basic Books.

Huebner, B. (2010) Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies?, *Phenomenology and the Cognitive Sciences*, **9**, pp. 133–155.

Humphrey, N. (2006) *Seeing Red*, Cambridge, MA: Harvard University Press.

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

Humphrey, N. (2016) Redder than red: Illusionism or phenomenal surrealism?, *Journal of Consciousness Studies*, **23** (11–12), pp. 116–123. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

Humphrey, N. (forthcoming) The invention of consciousness, *Topoi*.

Ismael, J. (1999) Science and the phenomenal, *Philosophy of Science*, **66**, pp. 351–369.

Jackson, F. (2003) Mind and illusion, *Royal Institute of Philosophy Supplements*, **53**, pp. 251–271.

Joyce, R. (2006) *The Evolution of Morality*, Cambridge, MA: MIT Press.

Kammerer, F. (2016) The hardest aspect of the illusion problem — and how to solve it, *Journal of Consciousness Studies*, **23** (11–12), pp. 124–139. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

Kammerer, F. (2018a) Can you believe it? Illusionism and the illusion meta-problem, *Philosophical Psychology*, **31**, pp. 44–67.

Kammerer, F. (2018b) Is the antipathetic fallacy responsible for the intuition that consciousness is distinct from the physical?, *Croatian Journal of Philosophy*, **18**, pp. 59–73.

Kammerer, F. (ms) Does the explanatory gap rest on a fallacy?

Kant, I. (1781/1999) *The Critique of Pure Reason*, Guyer, P. & Wood, A. (trans.), Cambridge: Cambridge University Press.

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

Latham, N. (1998) Chalmers on the addition of consciousness to the physical world, *Philosophical Studies*, **98**, pp. 71–97.

Lewis, D. (1995) Should a materialist believe in qualia?, *Australasian Journal of Philosophy*, **73**, pp. 140–144.

Loar, B. (1990) Phenomenal states, *Philosophical Perspectives*, **4**, pp. 81–108.

Lycan, W. (1987) *Consciousness*, Cambridge, MA: MIT Press.

Metzinger, T. (2003) *Being No One*, Cambridge, MA: MIT Press.

Molyneux, B. (2012) How the problem of consciousness could emerge in robots, *Minds and Machines*, **22**, pp. 277–297.

Mørch, H.H. (2018) The evolutionary argument for phenomenal powers, *Philosophical Perspectives*, **32**, pp. 293–316.

Muehlhauser, L. (2017) A software agent illustrating some features of an illusionist account of consciousness, *OpenPhilanthropy*, [Online], https://www.openphilanthropy.org/software-agent-illustrating-some-features-illusionist-account-consciousness.

Nagel, T. (1970) Armstrong on the mind, *Philosophical Review*, **79**, pp. 394–403.

Nagel, T. (1974) What is it like to be a bat?, *Philosophical Review*, **83**, pp. 435–450.

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

Papineau, D. (2002) *Thinking about Consciousness*, Oxford: Oxford University Press.

Papineau, D. (2007) Phenomenal and perceptual concepts, in Alter, T. & Walter, S. (eds.) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, Oxford: Oxford University Press.

Papineau, D. (2011) What exactly is the explanatory gap?, *Philosophia*, **39**, pp. 5–19.

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

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

Perry, J. (2001) *Knowledge, Possibility, and Consciousness*, Cambridge, MA: MIT Press.

Place, U.T. (1956) Is consciousness a brain process?, *British Journal of Psychology*, **47**, pp. 44–50.

Rey, G. (1996) Towards a projectivist account of conscious experience, in Metzinger, T. (ed.) *Conscious Experience*, Paderborn: Ferdinand-Schoeningh-Verlag.

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

Rosenthal, D. (1999) Sensory qualities and the relocation story, *Philosophical Topics*, **26**, pp. 321–350.

Rosenthal, D. (2002) Explaining consciousness, in Chalmers, D.J. (ed.) *Philosophy of Mind: Classical and Contemporary Readings*, New York: Oxford University Press.

Schneider, S. & Turner, E. (2017) Is anyone home? A way to find out if AI has become self-aware, *Scientific American*, blog, [Online], https://blogs.scientificamerican.com/observations/is-anyone-home-a-way-to-find-out-if-ai-has-become-self-aware/.

Schwarz, W. (forthcoming) Imaginary foundations, *Ergo*, [Online], https://www.umsu.de/papers/imaginary.pdf.

Shaffer, J. (1964) *Philosophy of Mind*, Upper Saddle River, NJ: Prentice-Hall.

Shoemaker, S. (1975) Functionalism and qualia, *Philosophical Studies*, **27**, pp. 291–315.

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 Intelligence*, [Online], https://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai-consciousness.pdf.

Stapp, H. (1993) *Mind, Matter, and Quantum Mechanics*, Berlin: Springer Verlag.

Stoljar, D. (2001) Two conceptions of the physical, *Philosophy and Phenomenological Research*, **62**, pp. 253–281.

Strawson, G. (2006) Realistic monism: Why physicalism entails panpsychism, *Journal of Consciousness Studies*, **13** (10–11), pp. 3–31.

Strawson, G. (2017) One hundred years of consciousness, *Isaiah Berlin Lecture*, Wolfson College, Oxford, 25 May 2017.

Street, S. (2006) A Darwinian dilemma for realist theories of value, *Philosophical Studies*, **127**, pp. 109–166.

Sturgeon, S. (1994) The epistemic basis of subjectivity, *Journal of Philosophy*, **91**, pp. 221–235.

Sundstrom, P. (2008) Is the mystery an illusion? Papineau on the problem of consciousness, *Synthese*, **163**, pp. 133–143.

Sytsma, J. & Machery, E. (2010) Two conceptions of subjective experience, *Philosophical Studies*, **151**, pp. 299–327.

Talbot, B. (2012) The irrelevance of folk intuitions to the 'hard problem' of consciousness, *Consciousness and Cognition*, **21**, pp. 644–650.

Tononi, G. (2007) Integrated information theory, in Velmans, M. & Schneider, S. (eds.) *The Blackwell Companion to Consciousness*, Oxford: Blackwell.

Tye, M. (1999) Phenomenal consciousness: The explanatory gap as a cognitive illusion, *Mind*, **108**, pp. 705–725.

Wierzbicka, A. (2010) *Experience, Evidence, and Sense: The Hidden Cultural Legacy of English*, Oxford: Oxford University Press.

Yudkowsky, E. (2008) Zombies! Zombies?, *Rationality: From AI to Zombies*, [Online], https://www.lesswrong.com/posts/fdEWWr8St59bXLbQr/zombies-zombies.

Call for Papers

Please find a call for papers for commentaries on this article here: https://philevents.org/event/show/64626.