Explanation, Understanding, and the Methodological Problem in Consciousness Science

This paper is forthcoming in Synthese (accepted March 9th 2025). Please cite the published version.

Abstract

Philosophers of mind and philosophers of science have markedly different views on the relationship between explanation and understanding. Reflecting on these differences highlights two ways in which explaining consciousness might be uniquely difficult. First, scientific theories may fail to provide a psychologically satisfying sense of understanding—consciousness might still seem mysterious even after we develop a scientific theory of it. Second, our limited epistemic access to consciousness may make it difficult to adjudicate between competing theories. Of course, both challenges may apply. While the first has received extensive philosophical attention, in this paper I aim to draw greater attention to the second. In consciousness science, the two standard methods for advancing understanding—theory testing and refining measurement procedures through epistemic iteration—face serious challenges.

Keywords: Scientific understanding, explanatory gap, theory-testing, epistemic iteration, natural kind method

Introduction

Philosophers of mind and philosophers of science have markedly different views on the relationship between explanation and understanding. For philosophers of mind, explanation is closely tied to logical entailment and the psychologically satisfying feeling of understanding. Philosophers of science, by contrast, tend to view explanation and scientific understanding in terms of prediction and control: if a theory or model allows you to predict how a system will behave under novel conditions and fix it when it breaks, that is sufficient for explanation.

Recognizing these different conceptions of explanation reveals two ways in which explaining consciousness may be uniquely difficult. First, constructing theories that yield reliable prediction and control might fall short of logical entailment and fail to provide a psychologically satisfying sense of understanding—consciousness may still seem somewhat mysterious. Thus, scientific explanations of consciousness might not meet the standards for explanation set by philosophers of mind. Second, our limited epistemic access to consciousness may make it unusually difficult to adjudicate between competing theories. In other words, our ability to achieve a scientific understanding of consciousness—to construct models that enable reliable prediction and control—might be profoundly limited. The first of these issues has received significant philosophical attention. Philosophers of mind have debated at length whether scientific theories of consciousness can fully dispel the sense of mystery and what, if anything, follows if they cannot. In this paper, I suggest setting these questions aside and focusing instead on a methodological problem that may prevent us from developing a reliable theory of consciousness in the first place.

In Section 1, I elaborate on the different ways in which philosophers of mind and philosophers of science have approached the relationship between explanation and understanding. Section 2 leverages this difference to motivate a shift of attention away from the traditional hard problem of consciousness and toward the epistemological challenges currently facing consciousness science. In Section 3, I discuss the two main approaches to addressing these epistemological challenges—theory testing and epistemic iteration—and explain why neither approach is clearly capable of resolving these challenges. I conclude that consciousness science faces a serious methodological problem. If we fail to resolve this issue, our ability to achieve a scientific understanding of consciousness will be profoundly limited.

1. Differing views on explanation in philosophy of mind and philosophy of science

It is possible to distinguish two components of explanation. On the one hand, explanations have an ontic component. Good explanations locate a phenomenon within the causal structure of the world—they identify what makes a difference to the phenomenon, and what the phenomenon, in turn, makes a difference to (Hempel, 1965; Salmon, 1984; Woodward, 2003; Craver, 2014). On the other hand, explanation also has a psychological component. Successful explanations typically yield understanding (Friedman, 1974; de Regt, 2017). When things are going smoothly, these two components of explanation come together. Locating a phenomenon within the causal structure of the world helps us understand why it is the way it is and not some other way. But things do not always go smoothly. Sometimes these two components of explanation come apart.

When my friend's 4-year-old daughter asks me why snowflakes have 6-sides and I tell her that it is due to the structure of the water molecule—the two hydrogens are not directly opposite one another but offset to form an angle of roughly 120° with the oxygen—my explanation succeeds in citing relevant causal (or, if you prefer, constitutive) details. But it does not succeed in leaving my friend's daughter feeling as though she understands why snowflakes have six sides. Being 4, she does not know what it means for the hydrogens to form an angle of 120° with the oxygen, nor does she understand why that might be relevant to snowflakes having six sides.

Conversely, suppose I were to tell her that snowflakes are made by ice spren—little elf-like creatures that live in snowflake factories up in the clouds—and that ice spren really like the number 6. This might leave her feeling as though she understands why snowflakes have 6 sides—the ice spren make them that way—even though the details it cites are utterly fictitious.

The general point here is that the ontic and the psychological components of explanation can come apart. Explanations that get the causal structure of the world wrong can sometimes leave us feeling as though we understand why things are the way they are. And sometimes, locating a phenomenon within the causal structure of the world does not leave us with a psychologically satisfying sense of understanding.

Philosophers of science and philosophers of mind have strikingly different attitudes toward the psychological component of explanation. But diagnosing the difference is a little harder than it at first appears. It will be helpful to start with a natural, though mistaken, picture that one can get from a superficial reading of the literature.

Surveying the literature on explanation in philosophy of mind, it is not too hard to come across statements that place a lot of weight on the psychological component of explanation. For example, in a widely cited paper on the topic of explanation and its relation to metaphysical dependence, Jaegwon Kim wrote that:

> The idea of explaining something is inseparable from the idea of making it intelligible; to seek an explanation of something is to seek to understand it, to render it intelligible. These are simple conceptual points, and I take them to be untendentious and uncontroversial. (Kim, 1994, p. 54)

And Joseph Levine makes a similar claim in his extended discussion of the nature of scientific explanation when he insists that "Explaining a phenomenon should yield understanding of the phenomenon" (Levine, 2001, p. 72).

From statements like these, one might get the impression that philosophers of mind view the psychologically satisfying feeling of understanding to be particularly important for explanation, perhaps even essential to it.

On the other hand, there is a tradition within the philosophy of science that views explanations as things in the world—independent of any human cognisers. Carl Hempel for example, insisted that "proper scientific inquiry… [is] independent of idiosyncratic beliefs and attitudes on the part of the scientific investigators" (Hempel, 2001, p. 374). And J.D Trout has argued that "what makes an explanation good concerns a property that it has independent of the psychology of the explainers; it concerns features of external objects, independent of particular minds" (2002, p. 217). This is the "ontic view" of scientific explanation (Craver, 2014). It is often traced back to Wesley Salmon (1984; 2006), who in turn traces it to Coffa (1974). According to the ontic view of explanation:

> explanations… are fully objective and, where explanations of nonhuman facts are concerned, they exist whether or not anyone ever discovers or describes them. Explanations are not epistemically relativized, nor (outside of the realm of human psychology) do they have psychological components, nor do they have pragmatic dimensions. (Salmon, 2006, p.133)

Based on comments like these, it can be tempting to think that while philosophers of mind care about understanding philosophers of science do not. However, that diagnosis is a bit too quick. On the one hand, although the passages from Kim and Levine quoted above explicitly talk of understanding, for the most part, it's not really understanding that philosophers of mind are concerned with, but *understandability*. They grant that explanation and understanding can come apart in cases where individuals lack the requisite background knowledge or cognitive abilities—as when my friend's daughter failed to grasp the explanation for why snowflakes have six sides—but they insist that explanations should at least be understandable. At the very least, they should be understandable to ideally rational agents with unlimited cognitive capabilities.

On the other hand, while there is a tradition within philosophy of science that views scientific explanations as mind independent features of objective reality, this is by no means universally accepted among philosophers of science (Friedman, 1974). Indeed, in recent years a significant literature has emerged investigating the nature of scientific understanding (de Regt, et al., 2009; Lawler et al., 2022). And, on closer inspection, even those who do favour the ontic view of explanation do not eschew understanding entirely. J.D. Trout, for example, grants that genuine "understanding is important to theory construction" (2005, p.199). And Wesley Salmon, goes so far as to say that: "Perhaps the most important fruit of modern science is the *understanding* it provides of the world in which we live, and the phenomena that transpire within it" (Salmon, 1984, pp. 19–20 my italics). To grasp the difference between how philosophers of mind and philosophers of science view explanation and understanding we need to go a little deeper.

1.1. Pragmatic Understanding in Philosophy of Science.

To understand the view of explanation and understanding one finds in philosophy of science we need to distinguish the *feeling* of understanding—the psychologically satisfying ah-ha! phenomenology that often accompanies grasping an explanation—from understanding as a cognitive achievement, or what Henk de Regt calls "pragmatic understanding" (de Regt, 2017).

Start with the *feeling* of understanding. Philosophers of science typically don't place too much weight on this aspect of understanding. Some are particularly scathing of it. J.D. Trout, for instance, has argued that the *feeling* of understanding is largely a product of confirmation bias and hindsight bias, and is not only inessential for successful explanation, but also a poor guide that our explanations are on the right track.

> The fact is, our history is littered with inaccurate explanations we confidently thought were obviously true: the explanation for mental illness in terms of demonic possession, the humoral theory of illness, and so on. The sense of understanding would be epistemically idle phenomenology were it not so poisonous a combination of seduction and unreliability. It actually does harm, sometimes making us squeamish about accepting true claims that we don't personally understand, and more often operating in the opposite direction, causing us to overconfidently accept false claims because they have a

> kind of anecdotal or theoretical charm. (Trout, 2002, pp. 229–230; see also Gopnik, 1998)

Not all philosophers of science are as distrustful of the feeling of understanding as Trout. Grimm (2009) and Lipton (2009) for instance, each point out that while there are indeed cases where the *feeling* of understanding has led scientists astray, there are also many cases where it has proven to be a useful guide (2009). Grimm and Lipton argue that while we should not put too much trust in the *feeling* of understanding, discarding it outright would be going too far.

But even those like Grimm and Lipton, who are sympathetic to the idea that the *feeling* of understanding can be a useful guide that our explanations are on the right track, agree that the *feeling* of understanding is neither necessary nor sufficient for explanation (Grimm, 2009, de Regt, 2017). On this, philosophers of science are largely in agreement (see also Trout, 2002; Craver, 2014).

The notion of 'understanding' that philosophers of science do take to be important, is understanding as a cognitive achievement, or what Henk de Regt calls "pragmatic understanding" (de Regt, 2017). Exactly how best to understand the notion of understanding as a cognitive achievement is a lively area of debate (de Regt, 2017; Khalifa, 2017; Smith, 2014; Wilkenfeld, 2013). The general idea, however, is that it is not quite enough to simply have a model or theory that accurately locates the phenomenon within the causal structure of the world, you also need to know, or understand, how to use the model to reliably predict and manipulate the phenomenon of interest. This requirement that scientific models not only track the causal structure of the world but also be usable by limited being such as us explains why scientific explanations often involve constructs known to be idealizations and abstractions (Craver, 2019; Potochnik, 2017).

The general picture that we get in philosophy of science then is that while the feeling of understanding is inessential, perhaps even harmful, understanding as a cognitive achievement isn't. A good explanation of a phenomenon, one that yields a scientific understanding, should enable researchers to reliably predict and control that phenomenon.

1.2. Understanding and Understandability in Philosophy of Mind

How about philosophers of mind? In philosophy of mind the *feeling* of understanding appears to play a much more central role in thinking about explanation. This is particularly clear in the context of explanations of consciousness where the feeling that there is something profoundly mysterious about how consciousness fits into the natural world has been a central theme for the past 30 years and arguably much longer. Consider a few examples to illustrate:

> "It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises… It seems objectively unreasonable that it should, and yet it does." (Chalmers, 1995, p. 201)
>
> "We have at present no conception of how a single event or thing could have both physical and phenomenological aspect, or how if it did they might be related" (Nagel, 1986, p. 47)
>
> "At this stage in the relevant sciences we have no idea how the neural substrate of my pain can explain why my pain feels like this rather than some other way or no way at all." (Block, 1997, p. 175).
>
> "[W]e have no idea, I contend, how a physical object could constitute a subject of experience, enjoying, not merely instantiating, states with all sorts of qualitative character" (Levine, 2001, p. 76).

These passages, as I read them, are not questioning our ability to build models that allow us to reliably predict and control consciousness. They do not seem to be concerned with the scientific or pragmatic sense of understanding. Instead, they question whether having such a model would be psychologically satisfying. The worry is that consciousness will always seem a little mysterious.

As mentioned earlier, it's not really understanding that philosophers of mind are concerned with, but *understandability*. They grant that explanation, and the feeling of understanding can come apart in cases where individuals lack the requisite background knowledge or cognitive abilities, but they insist that explanations should at least be understandable. At the very least, an ideally rational being with unlimited cognitive capabilities should be able to understand them.

Among philosophers of mind, both notions—understanding and understandability—are typically taken to be intimately connected to deductive arguments and logical entailment. This is explicitly the case in Levine's work. Levine takes "explanation to essentially involve deduction" and argues that "we achieve understanding when we can see why, given the information cited in the explanans, the phenomenon cited in the explanandum *had to be*" (Levine, 2001 p. 75 my italics).

This link between explanation, understanding, and logical entailment is also implicit in Chalmers' influential view that logical supervenience is a necessary precondition for successful reductive explanation. According to Chalmers, reductive explanation can only be successful in cases where "once we have told the lower-level story in enough detail, any sense of fundamental mystery goes away" (1996, p. 42). And this, he suggests, requires demonstrating that all facts about the higher-level phenomenon are logically entailed by facts about the lower-level mechanisms.

The general idea in philosophy of mind, then, appears to be something like this: a phenomenon is understandable when it is, in principle at least, logically derivable from more fundamental facts about the world. We understand that phenomenon when we can, in fact, derive it.

Pulling these threads together, it's clear that philosophers of mind and philosophers of science are operating with quite different views on explanation, understanding, and the relationship between the two. Philosophers of mind tend to think that a good explanation should show how the phenomenon to be explained is a logical consequence of other, more fundamental phenomena, and in doing so, generate a psychologically satisfying feeling of understanding. Philosophers of science, by contrast, downplay the importance of the feeling of understanding, and instead think of explanation and scientific understanding in terms of the construction and use of models to predict and control the phenomenon to be explained.

2. The hard problem and the methodological problem

For present purposes, we do not need to take a stand on how best to think about understanding and its relationship to explanation. Recognizing the different ways in which philosophers of mind and philosophers of science view explanation and understanding is itself enough to distinguish two ways in which explaining consciousness might be particularly hard. On the one hand, one might think that explaining consciousness is hard because there are principled reasons why locating consciousness within the causal structure of the world will not render all the facts about consciousness logically derivable from facts about the brain, nor will it deliver a psychologically satisfying sense of understanding—consciousness will still seem a bit mysterious. On the other hand, one might think that explaining consciousness is hard because our limited epistemic access to consciousness might limit our ability to build models that enable us to predict and control states of subjective experience. Of course, one might think both of these things too.

Philosophers of mind have primarily focussed on the first of these issues. Thinking about explanation and understanding as inherently tied to the *feeling* of understanding and logical entailment has led philosophers of mind to focus on one aspect of consciousness that they take to be particularly hard to explain. According to the familiar narrative, explaining the various capacities associated with consciousness—perceptual discrimination, categorization, internal access, verbal report, and so on—are the "easy problems" of consciousness (Chalmers, 1995, 1996). In labelling these "easy problems" Chalmers is under no delusion that explaining how brains perform these functions will be straightforward. Rather, he takes them to be "easy"—at least comparatively so—because capacities can be defined functionally, and functions are precisely the sort of thing that can be logically entailed by the description of a mechanism. The "hard" problem, according to this familiar narrative, is to explain why these capacities are accompanied by subjective experience—to explain why it feels like something to be us. This is taken to be "hard" because facts about subjective experience do not seem to be logically entailed by facts about neurocognitive mechanisms. Here's how Chalmers puts it.

> What makes the hard problem hard and almost unique is that it goes beyond problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience—perceptual discrimination, categorization, internal access, verbal report—there may still remain a further unanswered question: Why is the performance of these functions accompanied by experience? A simple explanation of the functions leaves this question open…
>
> To explain experience, we need a new approach. The usual explanatory methods of cognitive science and neuroscience do not suffice. These methods have been developed precisely to explain the performance of cognitive functions, and they do a good job of it. But as these methods stand, they are only equipped to explain the performance of functions. When it comes to the hard problem, the standard approach has nothing to say (Chalmers 1995, pp. 5–6).

Given the differences between how philosophers of mind and philosophers of science think about explanation and understanding outlined in the previous section, one might begin to question just how "standard" the approach Chalmers appeals to really is. In fact, a number of authors have already pointed out that the notion of explanation and understanding that is at play in debates about the hard problem and the explanatory gap is quite removed from how scientific explanation and understanding actually operate in the areas of science most relevant to the mind. Taylor (2015), Mirrachi (2017), and Klein and Baron (2020), each point out that explaining consciousness is only "hard" in Chalmers sense if we assume that doing so requires demonstrating that facts about consciousness are logically entailed by facts about the brain. Once we realize that successful scientific explanations—especially those in biology and the mind sciences—often fall well short of this, and instead aim to cite systematic difference-making relationships of the sort that are particularly useful for prediction and control, then there appears to be no deep barrier to providing cognitive or neuroscientific explanations of consciousness. And in a slightly different vein, Wright has argued that once we accept that the psychologically satisfying sense of understanding is neither necessary nor sufficient for successful scientific explanation, the question of whether possessing a successful theory of consciousness will make all sense of mystery go away becomes “scientifically irrelevant” (2007, p. 301).

This point—that explaining consciousness is only uniquely hard if we buy into the view of explanation implicit within the philosophy of mind—is important. But it is equally important not to overstate what this achieves. It does not make the issues that philosophers of mind have been concerned with go away. Rather, it sidesteps them.

Although the hard problem of consciousness and the explanatory gap are often presented as ‘explanatory’ problems, at their core they really concern a question about logical supervenience. Even if we grant that explaining consciousness and achieving a scientific understanding of it, does not require demonstrating that consciousness logically supervenes on cognition and neurobiology, philosophers of mind are well within their rights to point out that it still seems to be the case that facts about the qualitative character of consciousness are not logically entailed by facts about our neurocognitive mechanisms. And that still seems a bit weird. Elsewere in science, high-level phenomena do seem to be logically entailed by underlying mechanisms, even if the explanations scientists actually provide tend to fall well short of valid deductive arguments.

Philosophers of mind are also well within their rights to continue to wonder about the implications this failure of logical supervenience may have for the metaphysics of consciousness. Does it imply that consciousness involves the instantiation of non-physical properties (Chalmers, 1996) or that physics needs to be expanded (Goff, 2017)? Can the failure of logical supervenience be explained away by appealing to the special features of phenomenal concepts (Carruthers, 2019; Papineau, 2002)? Will the apparent failure of logical supervenience gradually “fade away, eventually vanishing in a puff of metaphysical smoke” as we develop better theories that enable us to predict and control consciousness (Seth, 2021, p. 28; see also Flanagan, 1992; Klein & Barron, 2020)? Or, do we merely think there’s a failure of logical supervenience due to limitations in our cognitive abilities (McGinn, 1989; Stoljar, 2006)?

Pointing out that logical supervenience is not a precondition for successful scientific explanation and scientific understanding doesn’t touch any of this. It may provide a reason to stop framing the hard problem of consciousness as an explanatory problem, but it doesn’t make the hard problem of consciousness go away.

What it does do, however, is suggest a shift in focus. Recognising that scientific explanation and understanding do not require one to be able to logically deduce the explanandum from the explanans, suggests that those interested in the science of consciousness need not worry too much about the metaphysical issues on which philosophers of mind have primarily focused. Instead, they should direct their efforts towards a second reason why explaining consciousness might prove to be particularly hard: methodological difficulties might significantly limit our ability to develop theories that allow us to predict and control—and hence, achieve a scientific understanding of—consciousness.

Many of the central questions in consciousness science today revolve around measurement. Although everyone agrees, give or take, that introspective reports can be trusted in a wide range of easy cases, there is no consensus on how to proceed in cases where introspective reports are either suspect or unavailable. How rich is conscious perception outside the focus of attention (Block 2007; Kouider et al., 2010; Phillips, 2018)? Does blindsight involve truly unconscious perception, or degraded but nonetheless conscious perception that goes unreported due to a conservative response bias (Philips, 2021a; Michel & Lau, 2021)? Are patients in disordered states of consciousness such as coma and unresponsive wakefulness syndrome conscious, and if so, when (Owen et al., 2006; Shea & Bayne, 2010)? Are infants conscious (Bayne et al., 2023)? How widespread is consciousness across the animal kingdom (Birch, 2022; 2024; Gisnburg & Jablonka, 2019)? Can artificial systems with radically different cognitive architecture than our own be conscious (Schwitzgebel, 2015; Dung, 2023; Butlin et al., 2023; Bayne & Williams, 2023; Seth, 2024; Mckilliam, forthcoming)? Are they already? If we cannot answer these questions, then our ability to develop models that allow us to reliably predict and control consciousness—to achieve a scientific understanding—is fundamentally limited.

Addressing these epistemological questions does not require tackling the hard problem of consciousness front on. We don't need to take a stand on the question of whether facts about consciousness logically supervenes on facts about cognition or neural mechanisms in order to answer questions about the distribution of consciousness. Nor do we need to take a stand on what, if anything, follows if they don't. But these epistemological problems aren't easy problems either.

Alvin Goldman once wrote that "the epistemological dimensions of consciousness research are just as difficult and daunting as the metaphysical ones, on which most of the recent philosophical discussion has focused" (Goldman, 2004, p. 21).[1] I'm not sure I'd go quite so far as that. But

[1] Block even suggests that one aspect of it may be even harder (2002).

Goldman is right that this is not merely another easy problem. Part of what makes the easy problems of consciousness easy is that we already have strategies in place capable of delivering answers: they are "directly susceptible to the standard methods of cognitive science" as Chalmers' put it (Chalmers, 1995). This is arguably not the case for the epistemological dimensions of consciousness research. The epistemological dimensions of consciousness research isn't tricky because we don't yet have answers to questions about consciousness outside the focus of attention, in non-human animals, and artificial systems, the real difficulty is that we don't know, or at least don't agree, on how to find answers to these questions.

3. The methodological problem in consciousness science

Broadly speaking there are two strategies researchers are currently deploying to make progress on the epistemological challenges in consciousness science. One is theory-driven. The theory-driven strategy aims to start with a wide range of competing theories of consciousness and try to eke out a winner through theory testing and structured adversarial collaborations (Del-pin et al., 2020; Mudrik et al., 2023). The hope is that once a winner emerges, we can then appeal to our theory of consciousness to decide how to proceed in the tricky cases where introspection is either suspect or unavailable. The other approach focusses on measurement practices themselves. It draws inspiration from the refinement of measurement practices elsewhere in science and aims to move beyond a naïve reliance on introspection via epistemic iteration and natural kind reasoning (Shea & Bayne, 2010; Shea, 2012; Birch, 2022; Michel, 2022; Mckilliam, 2024; Mckilliam, forthcoming b). Both face challenges.

3.1. The Theory-Driven Strategy

One strategy for making progress on the epistemological dimensions of consciousness science is to take existing theories of consciousness, derive some predictions from them, and then conduct experiments to test those predictions. If, at the end of the day, one theory proves to be superior to all the others, then the hope is that we will be able to leverage that theory to answer questions about consciousness in cases where introspective reports are either suspect or unavailable.

There are two difficulties with this approach. First, in their current form at least, each existing theory of consciousness is silent on at least some of the epistemological questions in consciousness science. And second, it is not clear that theory testing in consciousness science will lead to one theory being deemed unambiguously superior to others.

Start with the first. Many existing theories have been developed with one of two questions in mind. Either they aim to specify what determines whether a *mental state* is conscious rather than unconscious, or they aim to specify what determines whether a *creature* is conscious rather than unconscious. In general, theories developed with one of these questions in mind have little to say about the other.

For example, the global workspace theory is, primarily, a theory of what makes a mental state conscious rather than not (Mashour et al., 2020). It tells us that neurally encoded information becomes conscious when it is mobilized into a workspace that renders it globally available for a wide range of cognitive consuming systems. This is easy enough to apply to cases of conscious mental states in humans—if a participant is able to use information about a stimulus in thought, planning, memory, and so on, in other words, if it was globally available for cognition, then it was consciously perceived. But as a number of authors have pointed out, in its current form, the global workspace theory doesn't tell us much about what qualifies as a global workspace. As a result, it is not straightforward to apply the global workspace theory to answer questions about the distribution of consciousness throughout the animal kingdom (Carruthers, 2019; Birch, 2022; Schwitzgebel, 2020; Shevlin, 2021). Schwitzgebel's example of the garden snail makes the point nicely. In snails, information does travel broadly through the central nervous system, enabling coordinated action. Is that global workspace enough? Or is something more sophisticated required? As Schwitzgebel points out, "without the help of snail introspections or verbal reports, it is unclear how we should then generalize such findings to the case of the garden snail" (Schwitzgebel, 2020). Similar issues arise for other theories that focus on what makes a mental state conscious rather than not (Mudrik et al., 2023).

By contrast, for those theories that aim to specify what distinguishes conscious from non-conscious creatures, the converse is true. For example, Merker's mid-brain theory tells us that a creature is conscious if it possesses a self-model that integrates information about the environment together with the allostatic needs of the organism in order to guide self-preserving behaviour (2007). If that is right, then, it tells us that insects are probably conscious (Klein & Barron, 2016). But it is not clear what implications Merker's theory has for debates about the boundary between conscious and unconscious perception. Does it imply that conscious perception overflows attention and cognitive access? Or might attention still constrain when and how information is integrated into the organisms self-model, and thereby, whether or not it is conscious?

Moreover, when it comes to the question of consciousness in artificial intelligence, existing theories of consciousness provide only limited guidance—a point Butlin and colleagues are careful to point out in their recent report on consciousness in artificial intelligence (Butlin et al., 2023). While existing theories of consciousness are largely compatible with computational functionalism about consciousness—the view that consciousness depends on the functional (computational) organization of a system and not the mechanisms implementing those computations—this is typically because they are largely silent on the extent to which mechanistic details matter. However, the truth of computational functionalism remains, hotly contested (Godfrey-Smith, 2020; Seth, 2024; Mckilliam, forthcoming a).

These complications suggest that the theory-driven approach to the epistemological dimension of consciousness research may be less help we might have hoped. Even if theory testing is able to converge on one theory as clearly superior to all others, there are open questions about the extent to which we will be able to leverage that theory to answer questions about the distribution of consciousness.

A second concern is that it is not clear that theory-testing in consciousness science will converge on a single theory. Theory-testing works best when competing theories are unambiguously theories of the same phenomenon. This is not obviously the case in consciousness science—existing theories of consciousness are homing in on quite different features of our neurocognitive mechanisms. Global workspace theories associate consciousness with information being made globally available for cognition (Mashour et al., 2020). Higher order theories associate consciousness with certain metacognitive abilities (Brown et al., 2019). Lamme's recurrent processing theory and IIT associate consciousness with the integration of information (Lamme, 2006; Tononi et al., 2016). Merker's mid-brain theory associates consciousness with a minimal form of self-modelling (Merker, 2007); and so on. These are all real phenomena. And, as a result, we should expect progressive research programs to emerge around each of them (Lakatos, 1978). The question will remain, which theory better tracks the mechanisms responsible for subjective experience?

This wouldn't be such a problem if we had a consciousness meter. If we had a consciousness meter—if we already knew how to detect consciousness in the cases where introspective reports are either suspect or unavailable—then we could simply test which of these theories does a better job of tracking consciousness. But we do not have a consciousness meter. The worry then, is that any empirical test capable of arbitrating between competing theories will also be one in which a degree of uncertainty is warranted as to whether or not consciousness is present. And in the face of uncertainty, we should not be surprised if researchers turn to their preferred theory for guidance—in fact it is arguably rational for them to do so. But if they do, then we should expect theory testing in consciousness science to lead to divergence rather than convergence, with advocates of competing theories becoming ever more deeply entrenched in their preferred theory (Irvine, 2013; Mckilliam, 2024).

Admittedly, this may not be an insurmountable problem. Some authors are optimistic that theory-testing may allow for convergence if we adopt more rigorous testing methods and emphasize the collaborative rather than the adversarial aspect of adversarial collaborations (Melloni, 2022; Corcoran et al., 2023; Negro, 2024). But it is worth keeping in mind that adversarial collaborations do not have a strong track record of delivering convergence elsewhere in science (Kahneman & Klein, 2009; Latham, et al., 1988; Mellers, et al., 2001). And, when we take a look at the state of theory testing in consciousness science today, divergence, rather than convergence, appears to be currently well under way (Yaron et al., 2023; Cleeremans 2023; Cogitate, 2023).

Another optimistic possibility here is that the kind of disagreement we see today may not persist in the next generation. In other words, we might converge on a single theory of consciousness via the Kuhnian route (Kuhn, 1962). Even if the current generation of researchers cling to their preferred theory, the next generation may see one theory as clearly superior to the others. But as an anonymous reviewer has pointed out, here we face a slightly different worry. When scientific consensus is honestly won it is diagnostic of credibility—that's why we care about consensus. But one might worry that in consciousness studies, consensus is just as likely (perhaps even more likely) to indicate that people go to similar conferences, publish in the same journals, have training that makes them prone to the same failures of imagination, or a host of other non-epistemic sources of consensus. In other words, consensus via the Kuhnian route may turn out not to be a reliable indicator that we've actually arrived at the correct theory.

So where does that leave us? Arguably, with a considerable degree of uncertainty about the prospects of a theory-driven approach to the epistemological challenges in consciousness science. In their current guise, most existing theories of consciousness only provide a partial guide to questions about the distribution of consciousness. And, at this stage at least, it is not clear that theory testing in consciousness science will reveal one theory to be clearly superior to all others.

3.2. Epistemic Iteration and Natural Kind Reasoning.

An alternative approach is to try to bootstrap our way to better measurement procedures via epistemic iteration and natural kind reasoning (Michel, 2022; Birch, 2022; Bayne et al., 2024; Mckilliam, 2024). An analogy with the scientific study of temperature can help illustrate the idea.

We didn't always have thermometers. Initially, our only procedures for measuring temperature were sensations—perceptual judgments of hot and cold. Experimentalists were able to bootstrap their way to thermoscopes and thermometers—instruments capable of correcting sensations—via a process Hasok Chang has called "epistemic iteration", in which "successive stages of knowledge, each building on the preceding one" are created "in the absence of assured foundations" (2004, p. 45; see also 2017, p. 231).

Simplifying matters somewhat, experimentalists interested in the measurement of temperature noticed that how hot or cold something felt to the touch systematically correlated with how much it caused a vial of fluid to expand and contract. This systematic correlation was best explained by the hypothesis that temperature causes both i) sensations of hot and cold, and ii) fluids to expand and contract. With this hypothesis in place, experimentalists could then leverage the expansion and contraction of fluids to build thermoscopes and themometers that are capable of both *extending* and *correcting* the initial standards of measurement. Pots of boiling water may have been too hot to measure reliably via sensations, but not via thermometers. And in cases where your perceptual judgement diverges from the thermometer, unless you can find some explanation for why the usually reliable thermometer is malfunctioning in this particular case, explanatory considerations suggest that it is probably your perceptual judgement that is in error.

In recent years there has been an increased interest in leveraging epistemic iteration and natural kind reasoning to make progress in consciousness science too (Michel, 2022; Bayne et al., 2024; Mckilliam, 2024). The general idea is that consciousness science can begin with clear-cut cases where its presence is undisputed and identify a cluster of cognitive abilities that appear to be facilitated by conscious processing. We can then appeal to this cluster of cognitive abilities to both extend and correct the initial criteria for detecting consciousness that we relied on in the clear-cut cases.

For example, the ability to over-ride primed responses (Debner & Jacoby, 1994), certain cross-modal interaction effects such as the McGurk Effect (Palmer & Ramsey, 2012), and the ability to quickly realize that a learned patten no longer holds (Travers, et al., 2018), all correlate with the presence of consciousness as measured by subjects' introspective judgements. A tidy explanation for this correlation is that our brains engage in a distinct mode of information processing that i) feels like something from a subjective perspective and also facilitates, among other things, ii) the over-riding of primed responses, iii) certain forms of cross-modal integration, iv) rapid reversal learning.

With this in place, we can leverage this cluster of abilities to detect consciousness even in creatures that lack the sophisticated kinds of metacognition necessary for introspection and report. Suppose we find that fish can override primed responses and exhibit rapid reversal learning, and that these capacities can be switched off through masking in much the same way that the cluster of consciousness related cognitive abilities is switched off by masking in humans. If that turns out to be the case, it would provide compelling evidence that fish engage in the same kind of information processing that, in us, gives rise to subjective experience. Considerations of plausibility, then, would suggest that it also feels like something to be a fish (see Shea & Bayne, 2010; Bayne et al., 2024; Birch, 2022; Mckilliam, 2024).

We can also leverage this methodology to detect and correct for introspective errors (Mckilliam, forthcoming b). Consider the case of aphantasia. In people who report conscious experiences of mental imagery, imagery appears to i) prime which stimulus will resolve first in binocular rivalry (Keogh & Pearson 2018; 2024) ii) influence pupil size (Kay et al., 2022), and iii) amplify the emotional salience of passages of text—those who report no conscious imagery have a significantly dampened fear response to text-based scary stories even though their response to scary images is no different (Wicken et al., 2021). If we now encounter someone who reports conscious imagery but displays none of these behavioral markers, the best explanation for all the data may be that their introspective report is unreliable.

In this way, iterative natural kind reasoning can, in principle, allow us to move beyond a naïve reliance on introspection.

However, there are substantial open questions about the prospects of this strategy succeeding in consciousness science.[2] Inverstigating consciousness in this way is undoubtedly going to be considerably more complicated than the case of temperature. In the case of temperature, there was a single phenomenon—a single property cluster—to home in on. The clustering in the vicinity of consciousness is unlikely to be so well-behaved. As mentioned above, there are multiple real phenomena in the vicinity of consciousness—global availability, self-monitoring, perceptual processing, and so on—and these are likely to be causally related in complex and overlapping ways. As a result, we should not be surprised to find multiple overlapping property clusters in the vicinity of consciousness, each underpinned by distinct aspects of neural processing.

[2] There may even be open questions about whether this approach is appropriate for investigating consciousness. Some researchers might object for reasons similar to those raised against the Cornell Realists' proposal to apply scientific methods to morality. I am grateful to Tim Bayne for bringing this to my attention.

By itself this isn't a deep problem—progress in science often requires untangling complex causal networks. However, it may become a problem when we take into account the fact that researchers appear to have subtly different (and sometimes not so subtly different) intuitions about how much evidential weight to ascribe the various pretheoretical marks of consciousness.

To get a sense of the problem, consider Matthias Michel's recent discussion of calibration in consciousness science. Michel argues that research on conscious perception relies on two basic principles. The first: "if you are able to meaningfully respond to stimuli, you are more likely to be conscious of them than if you are unable to do so" (2021, p. 831). The second: "people can usually tell whether they are conscious of something or not" (2021, p. 838). He points out that so long as we treat both of these basic principles as initial guesses capable of error, then there is no in principle difficulty with using each to calibrate the other (2021, p. 839).

Unfortunately, things are unlikely to be quite as straightforward as Michel's discussion implies. The difficulty is that there are live disagreements about how much evidential weight to ascribe each of these basic principles. Consider the recent controversy over blindsight for example. While Michel suggests that the fact that blindsight subjects report no visual experiences in their blind hemifield is "good reason" to distrust signal detection theoretic measures in this case (Michel 2021, p. 835), Ian Phillips disagrees. As he sees things, the data are better explained by the hypothesis that blindsight actually involves conscious, though degraded, vision, and that the negative reports are due to a conservative response bias (Phillips, 2021a; 2021b). Michel and Lau disagree (2021).

The details of this debate a subtle, but they are not essential for grasping the deeper methodological issue here. Epistemic iteration is a proceedure for homing in on phenomena in the absence of sure foundations. It works well when there is a single phenomenon to home in on. When that is the case, so long as our initial guesses are at least in the right ballpark, disagreement about where to start will wash out in the end. But when there are multiple phenomena in the vicinity, subtle disagreements about where to start can be compounded, rather than resolved by epistemic iteration.

This issue is not obvious, but an analogy with Newton's method for approximating the root of a function can help to bring it out.[3] In Newton's Method, we begin with an initial guess, x_n, for where the root might be. We then calculate the slope of the function at that point and extend a tangent line down to the x-axis. If our initial guess is at least in the right ball park, the x-

[3] The analogy was suggested by Hasok Chang (2017).

coordinate where this tangent intersects the x-axis, x_{n+1}, will be a better approximation of the true root (see fig. 1).

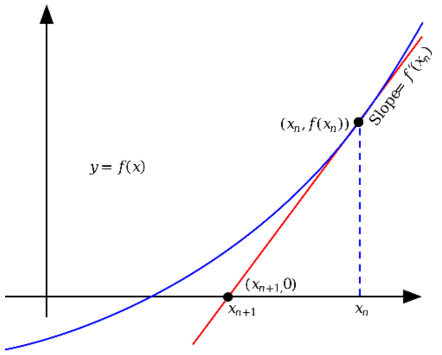

Figure 1: An illustration of the first step in Newton's method for approximating the root.

By iterating this process—using x_{n+1} as our next guess—we can refine our estimate, converging on the true root to whatever degree of accuracy we like (Chang, 2017).

Newton's method works well in cases where there is only a single solution—where the function only crosses the x-axis once. However, for more complex functions with multiple crossings we can run into a problem. If, for example, the function crosses the x-axis in two places then our initial guess constrains where we will end up.

The worry is that we may be in an analogous situation in consciousness science. If we start with a higher credence in subjective reports, epistemic iteration might lead us to home in on a cluster of properties associated with global availability and/or self-monitoring. Conversely, if we start with a higher credence in above chance performance on forced choice tasks, epistemic iteration might lead us to associate consciousness with a cluster of properties associated with perceptual processing and sensory integration.

This problem about starting points was initialy pointed out by Ian Phillips (2018). Elsewhere, I have suggested that there migth be more aggreement on starting with introspective judgments than first meets the eye, but this is, admittedly, a speculative point (Mckilliam, 2024). It is currently unclear whether there is sufficient agreement on how much evidential weight to ascribe the various pre-theoretical principles grounding subjective and objective measures of consciousness for epistemic iteration to deliver consensus.

Conclusion

In this paper, I have suggested that our ability to achieve a scientific understanding of consciousness may be profoundly limited—not just for the familiar reason that facts about subjective experience do not appear to be logically entailed by facts about cognition or neurobiology, but also because methodological challenges may limit our ability to construct models that would enable us to reliably predict and control consciousness. While there is little reason to doubt clear-cut cases where human subjects provide unambiguous introspective reports, many cases warrant a significant degree of uncertainty. At present, it remains unclear whether the methods used in consciousness science can resolve this uncertainty—both theory-driven and measurement-focused approaches face serious challenges. If these challenges cannot be overcome, our ability to achieve a scientific understanding of consciousness may be profoundly limited.

My aim in this paper has not been to sew pessimism. On the contrary, I am actually quite optimistic about the prospects of making progress on these issues. But they will not be resolved if they continue to be swept under the rug. My hope is that by bringing these difficulties out into the daylight, we might begin to make progress on them. Much work remains to be done.

References

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., ... & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in cognitive sciences.* https://doi.org/10.1016/j.tics.2024.01.010

Bayne, T., Frohlich, J., Cusack, R., Moser, J., & Naci, L. (2023). Consciousness in the cradle: on the emergence of infant experience. *Trends in cognitive sciences.* https://doi.org/10.1016/j.tics.2023.08.018

Bayne, T., & Williams, I. (2023). The Turing test is not a good benchmark for thought in LLMs. *Nature Human Behaviour*, *7*(11), 1806–1807. https://doi.org/10.1038/s41562-023-01710-w

Birch, J. (2022). The search for invertebrate consciousness. *Noûs, 56*(1), 133–153. https://doi.org/10.1111/nous.12351

Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI.* Oxford University Press. Oxford. https://doi.org/10.1093/9780191966729.001.0001

Block, N. (1997). Begging the question against phenomenal consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The nature of consciousness: philosophical debates* (pp. 176–179). MIT Press.

Block, N (2002). The harder problem of consciousness. *Journal of Philosophy 99*, 391–425. https://doi:10.2307/3655621

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30*(5-6), 481–499. https://doi.org/10.1017/s0140525x07002786

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in cognitive sciences.* https://doi.org/10.1016/j.tics.2019.06.009

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. https://doi.org/10.48550/arXiv.2308.08708

Carruthers, P. (2019). *Human and animal minds: the consciousness questions laid to rest.* Oxford University Press. https://doi.org/10.1093/oso/9780198843702.001.0001

Carruthers, P. (2020). Stop caring about consciousness. *Philosophical Topics*, *48*(1), 1–20. https://www.jstor.org/stable/48628583

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.

Chalmers, D. (1996). *The conscious mind: in search of a fundamental theory*. Oxford University Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*: Oxford University Press. https://doi.org/10.1093/0195171276.001.0001

Chang, H. (2017). Epistemic iteration and natural kinds: Realism and pluralism in taxonomy. In K. Kendler, and J. Parnas (Eds.) *Philosophical issues in psychiatry IV: Psychiatric nosology* (pp. 229–245). Oxford University Press. https://doi.org/10.1093/med/9780198796022.003.0029

Cleeremans, A. (2022). Theory as adversarial collaboration. *Nature Human Behaviour, 6*(4), 485–486. https://doi.org/10.1038/s41562-021-01285-4

Coffa, J. A. (1974). Hempel's ambiguity. *Synthese*, 141–163. https://doi.org/10.1007/BF00485232

Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., ... & Melloni, L. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *bioRxiv*, 2023–06. https://doi.org/10.1101/2023.06.23.546249

Corcoran, A. W., Hohwy, J., & Friston, K. J. (2023). Accelerating scientific progress through Bayesian adversarial collaboration. *Neuron, 111*(22), 3505–3516. https://doi.org/10.1016/j.neuron.2023.08.027

Craver, C. F. (2014). The ontic account of scientific explanation. In M. I. Kaiser, O. R. Scholz, D. Plenge, & A. Hüttemann (Eds.), *Explanation in the special sciences: The case of biology and history* (pp. 27–52): Springer Netherlands. https://doi.org/10.1007/978-94-007-7563-3_2

Craver, C. F. (2019). Idealization and the ontic conception: A reply to Bokulich. *The Monist, 102*(4), 525–530. https://doi.org/10.1093/monist/onz023

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press. https://doi.org/10.1093/oso/9780190652913.001.0001

de Regt, H. W., Leonelli, S., & Eigner, K. (Eds.). (2009). *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press. https://doi.org/10.2307/j.ctt9qh59s

Debner, J. A., & Jacoby, L. L. (1994). Unconscious perception: attention, awareness, and control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **20**(2), 304. https://doi.org/10.1037/0278-7393.20.2.304

Del Pin, S., Skóra, Z., Sandberg, K., Overgaard, M., & Wierzchoń, M. (2021). Comparing theories of consciousness: why it matters and how to do it. *Neuroscience of Consciousness, 2021*(2). https://doi.org/10.1093/nc/niab019

Dung, L. (2023). Tests of Animal Consciousness are Tests of Machine Consciousness. *Erkenntnis*, 1–20. https://doi.org/10.1007/s10670-023-00753-9

Flanagan, O. J. (1992). *Consciousness reconsidered*. MIT press. https://doi.org/10.7551/mitpress/2112.001.0001

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*(1), 5–19. https://doi.org/10.2307/2024924

Ginsburg, S., & Jablonka, E. (2019). *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*. MIT Press. https://doi.org/10.7551/mitpress/11006.001.0001

Godfrey-Smith, P. (2020). *Metazoa: Animal life and the birth of the mind*. Farrar, Straus and Giroux.

Goff, P. (2017). *Consciousness and fundamental reality*. Oxford University Press. https://doi.org/10.1093/oso/9780190677015.001.0001

Goldman, A. (2004). Epistemology and the Evidential Status of Introspective Reports. *Journal of Consciousness Studies, 11*(7), 1–16.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*, 101–118. https://doi.org/10.1023/a:1008290415597

Grimm, S. (2009). Reliability and the Sense of Understanding. In H. de Regt; S. Leonelli, & K. Eigner (Eds.) *Scientific understanding: Philosophical perspectives*, (pp. 83–99). University of Pittsburgh Press. https://doi.org/10.2307/j.ctt9qh59s.8

Hempel, C. G. (1965). *Aspects of scientific explanation*. The Free Press.

Hempel, C. G. (2001). Valuation and objectivity in science. In J. H. Fetzer (Ed.), *The philosophy of Carl G. Hempel: Studies in science, explanation, and rationality* (pp. 372–396): Oxford University Press.

Irvine, E. (2013). *Consciousness as a scientific concept: A philosophy of science perspective*. Springer. https://doi.org/10.1007/978-94-007-5173-6

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American Psychologist, 64*(6), 515. https://doi.org/10.1037/a0016755

Kay, L, Keogh, R., Andrillon, T., & Pearson, J. (2022). The pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength. *Elife, 11,* e72484. https://doi.org/10.7554/eLife.72484

Keogh, R. & Pearson, J. (2018). The blind mind: No sensory visual imagery in aphantasia. *Cortex, 105*, 53–60. https://doi.org/10.1016/j.cortex.2017.10.012

Keogh, R. & Pearson, J. (2024). Revisiting the blind mind: Still no evidence for sensory visual imagery in individuals with aphantasia. *Neuroscience Research 201*, 27–30. https://doi.org/10.1016/j.neures.2024.01.008

Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press. https://doi.org/10.1017/9781108164276

Kim, J. (1994). Explanatory knowledge and metaphysical dependence. *Philosophical Issues, 5*, 51–69. https://doi.org/10.2307/1522873

Klein, C., & Barron, A. B. (2016). Insects have the capacity for subjective experience. *Animal Sentience*, *1*(9), 1. https://doi.org/10.51291/2377-7478.1113

Klein, C., & Barron, A. B. (2020). How experimental neuroscientists can fix the hard problem of consciousness. *Neuroscience of Consciousness, 2020*(1), niaa009. https://doi.org/10.1093/nc/niaa009

Kouider, S., De Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences, 14*(7), 301–307. https://doi.org/10.1016/j.tics.2010.04.006

Kuhn, T. S. (1962/2012). The Structure of Scientific Revolutions. University of Chicago press. https://doi.org/10.7208/chicago/9780226458144.001.0001

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes: Philosophical papers volume 1*. (J. Worrall & G. Currie, Eds.). Cambridge University Press.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Latham, G. P., Erez, M., & Locke, E. A. (1988). Resolving scientific disputes by the joint design of crucial experiments by the antagonists: Application to the Erez–Latham dispute regarding participation in goal setting. *Journal of Applied Psychology, 73*(4), 753. https://doi.org/10.1037/0021-9010.73.4.753

Lawler, I., Khalifa, K., & Shech, E. (Eds.). (2022). *Scientific understanding and representation: Modeling in the physical sciences*. Routledge. https://doi.org/10.4324/9781003202905

Levine, J. (2001). *Purple haze: The puzzle of consciousness*. Oxford University Press. https://doi.org/10.1093/0195132351.001.0001

Lipton, P. (2009). Understanding without explanation. In H. de Regt; S. Leonelli, & K. Eigner (Eds.) *Scientific understanding: Philosophical perspectives*, (pp. 43–63). University of Pittsburgh Press. https://doi.org/10.2307/j.ctt9qh59s.8

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron, 105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

McGinn, C. (1989). Can we solve the mind—body problem? *Mind, 98*(391), 349–366. https://doi.org/10.1093/mind/xcviii.391.349

Mckilliam, A. (2024). Natural kind reasoning in consciousness science: An alternative to theory testing. *Noûs*. https://doi.org/10.1111/nous.12526

Mckilliam, A. (forthcoming a). Do mechanisms matter for inferences about consciousness? *Australasian Journal of Philosophy*.

Mckilliam, A. (forthcoming b). Detecting Introspective Errors in Consciousness Science. *Ergo*.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science, 12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

Melloni, L. (2022). On keeping our adversaries close, preventing collateral damage, and changing our minds. Comment on Clark et al. Journal of Applied Research in Memory and Cognition, 11(1), 45–49. https://doi.org/10.1037/mac0000009

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioral and Brain Sciences, 30*(1), 63–81. https://doi.org/10.1017/s0140525x07000891

Michel, M. (2021). Calibration in consciousness science. *Erkenntnis*, 1–22. https://doi.org/10.1007/s10670-021-00383-z

Michel, M., & Lau, H. (2021). Is blindsight possible under signal detection theory? Comment on Phillips (2021). https://doi.org/10.1037/rev0000266

Miracchi, L. (2017). Generative explanation in cognitive science and the hard problem of consciousness. *Philosophical Perspectives*, *31*, 267–291. https://doi.org/10.1111/phpe.12095

Mudrik, L., Mylopoulos, M., Negro, N., & Schurger, A. (2023). Theories of consciousness and a life worth living. Current Opinion in Behavioral Sciences, 53, 101299. https://doi.org/10.1016/j.cobeha.2023.101299

Nagel, T. (1986). *View from nowhere*: Oxford University Press.

Negro, N. (2024). (Dis)confirming theories of consciousness and their predictions: towards a Lakatosian consciousness science. *Neuroscience of Consciousness, 2024*(1), niae012. https://doi.org/10.1093/nc/niae012

Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting awareness in the vegetative state. *Science*, *313*(5792), 1402–1402. https://doi.org/10.1126/science.1130197

Palmer, T. D., & Ramsey, A. K. (2012). The function of consciousness in multisensory integration. *Cognition*, **125**(3), 353–364. https://doi.org/10.1016/j.cognition.2012.08.003

Papineau, D. (2002). *Thinking about consciousness*. Clarendon Press. https://doi.org/10.1093/0199243824.001.0001

Phillips, I. (2018). The methodological puzzle of phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1755), 20170347. https://doi.org/10.1098/rstb.2017.0347

Phillips, I. (2021a). Blindsight is qualitatively degraded conscious vision. *Psychological Review, 128*(3), 558. https://doi.org/10.31234/osf.io/gdk6m

Phillips, I. (2021b). Bias and blindsight: A reply to Michel and Lau (2021). *Psychological Review, 128*(3), 592–595. https://doi.org/10.1037/rev0000277

Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226507194.001.0001

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press. https://doi.org/10.1515/9780691221489

Salmon, W. C. (2006). *Four decades of scientific explanation* (1st ed.). University of Pittsburgh Press.

Schwitzgebel, E. (2015). If materialism is true, the United States is probably conscious. *Philosophical Studies*, *172*, 1697–1721. https://doi.org/10.1007/s11098-014-0387-8

Schwitzgebel, E. (2020). Is there something it's like to be a garden snail? *Philosophical Topics, 48*(1), 39–64. https://doi.org/10.5840/philtopics20204813

Seth, A. (2021). *Being you: A new science of consciousness*. Penguin.

Seth, A. (2024). Conscious artificial intelligence and biological naturalism. *PsyArXiv preprint.* https://doi.org/10.31234/osf.io/tz6an

Shea, N. (2012). Methodological encounters with the phenomenal kind. *Philosophy and Phenomenological Research, 84*(2), 307–344. https://doi.org/10.1111/j.1933-1592.2010.00483.x

Shea, N., & Bayne, T. (2010). The vegetative state and the science of consciousness. *British Journal for the Philosophy of Science, 61*(3), 459–484. https://doi.org/10.1093/bjps/axp046

Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, *36*(2), 297–314. https://doi.org/10.1111/mila.12338

Smith, R. (2014). Explanation, understanding, and control. *Synthese*, *191*, 4169-4200. https://doi.org/10.1007/s11229-014-0521-3

Stoljar, D. (2006). *Ignorance and imagination: The epistemic origin of the problem of consciousness*. Oxford University Press. https://doi.org/10.1093/0195306589.001.0001

Taylor, E. (2016). Explanation and the explanatory gap. *Acta Analytica*, *31*(1), 77–88. https://doi.org/10.1007/s12136-015-0260-1

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016) Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience 17*, 450–461 (2016). https://doi.org/10.1038/nrn.2016.44

Travers, E., Frith, C. D., & Shea, N. (2018). Learning rapidly about the relevance of visual cues requires conscious awareness. *Quarterly Journal of Experimental Psychology*, **71**(8), 1698–1713. https://doi.org/10.31234/osf.io/7becr

Trout, J. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69*(2), 212–233. https://doi.org/10.1086/341050

Trout, J. D. (2005). Paying the price for a theory of explanation: De Regt's discussion of Trout. *Philosophy of Science, 72*(1), 198–208. https://doi:10.1086/426849

Trout, J. (2007). The psychology of scientific explanation. *Philosophy Compass, 2*(3), 564–591. https://doi.org/10.1111/j.1747-9991.2007.00081.x

Wicken, M., Keogh, R., & Pearson, J. (2021). The critical role of mental imagery in human emotion: Insights from fear-based imagery and aphantasia. *Proceedings of the royal society B*, *288*(1946), 20210267. https://doi.org/10.1098/rspb.2021.0267

Wilkenfeld, D. A. (2013). Understanding as representation manipulability. *Synthese*, *190*, 997–1016. https://doi.org/10.1007/s11229-011-0055-x

Wright, W. (2007). Explanation and the hard problem. *Philosophical Studies*, *132*, 301–330. https://doi.org/10.1007/s11098-005-2220-x

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press. https://doi.org/10.1093/0195155270.001.0001

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour, 6*(4), 593–604. https://doi.org/10.1101/2021.06.10.447863