# Do Mechanisms Matter for Inferences About Consciousness?

Andy Mckilliam
National Taiwan University

*[This paper is forthcoming in AJP. Please cite the published version.]*

What should we make of systems that behave just like conscious creatures but operate via mechanisms that are profoundly different from our own? How should we even think about mechanistic similarity and difference in this context? To answer these questions, I take a closer look at the inferential machinery that allows us to justifiably draw conclusions about consciousness in others. I argue that inferences about consciousness in others are best viewed as involving analogical inferences grounded in explanatory considerations. I conclude that profound mechanistic difference severs the evidential link between behavioural evidence and subjective experience. I then draw on recent developments within the mechanistic philosophy of science to motivate a conception of mechanistic similarity that is not wedded to similarity of physical substrate and outline a strategy for resolving disputes about whether it is cognitive similarity or neural similarity that matters for inferences about consciousness.

**Keywords:** animal consciousness; artificial consciousness; natural kind reasoning; argument from analogy; inference to the best explanation; mechanisms

## Introduction

To gather evidence about consciousness in ourselves we can rely on introspection—reflection on our own case. Naturally, we cannot gather evidence about consciousness in others in this way. Instead, we draw inferences about the experiences of others based on two sources of objectively observable data—behaviour and brain processes. What should we do when these two lines of evidence diverge? In particular, what should we make of systems that behave just like conscious creatures but operate via mechanisms that are profoundly different from our own? How should we even think about mechanistic similarity and difference in this context?

One approach to these questions is theory heavy (Birch 2022). The theory heavy approach is guided by the thought that in order to answer questions about the distribution of consciousness throughout the world we first need a theory of consciousness, one that tells us what consciousness is and what kinds of systems can have it (Tononi and Koch 2015; Carruthers 2019). This is not the approach I adopt here. There are currently many theories claiming to be theories of consciousness. Many of these theories are targeting different, but equally real features of our neurocognitive mechanisms and it is unclear how to decide between them in the absence of consensus concerning how to measure consciousness (Irvine 2017; Seth and Bayne 2020; Mckilliam 2024). Moreover, many of these theories are silent on the question of whether and to what extent mechanisms matter for consciousness (Butlin, Long, *et al.* 2023).[[1]](#footnote-1) Global Workspace Theory, for example, can be interpreted as either binding consciousness to a certain computational architecture or to a particular neural implementation of that architecture. Why should we prefer the former to the latter?

Answering this question has been dubbed the “harder problem of consciousness” (Block 2002) and some authors worry that it may be empirically unresolvable (Michel 2019).[[2]](#footnote-2) That may be so if we insist on adopting a theory heavy methodology. But there is an alternative.

Rather than starting with a theory of consciousness, we can start with epistemology and see if we can make progress that way. What exactly do I mean by starting with epistemology? Well, we can think of the question of consciousness in others—be they human or otherwise—as just a special case of a broader set of questions in the epistemology of science concerning how to draw conclusions about unobserved states of affairs. The epistemological approach to the question of other minds, which subsumes the “theory-light” approach (Birch 2022) and “natural kind method” (Shea and Bayne 2010), asks: What are the conditions that allow us to draw conclusions about unobserved states of affairs in general? And how can they help us decide whether mechanisms matter for inferences about consciousness?

One advantage of this approach is that it is relatively neutral about contested issues concerning the nature of consciousness—reductionists, functionalists, and dualists can all get on board. But it is not entirely neutral. There are two assumptions that we need. First, we need to assume that introspective judgements about consciousness are not wildly unreliable. We do not need them to be infallible. Or even close to it. But if introspection is no more reliable as a guide to consciousness than, say, Tarot is as a guide to the future, then we are in trouble.

The second assumption we need is that consciousness is a natural phenomenon. What do I mean by that? Well, at a first pass, I mean that the category of conscious entities is more like the category of ‘things that are alive’ than the category of ‘things that are cool, hip, or trendy’. There is a real partitioning in the world that separates living things from non-living things. That partitioning may be fuzzy about the edges—viruses appear to be edge cases—but it is no less real because of it. Moreover, that partitioning exists irrespective of how we choose to deploy the term ‘life’. By contrast, what is and is not ‘cool’ is, in a sense, up to us—or, rather, it is up to the cool kids who live in the trendy part of town. To say that consciousness is a natural phenomenon is to say that there is a real partitioning in the world that separates those entities that are conscious from those that are not, it is not something that we simply get to decide. That partitioning may be fuzzy about the edges, but it is no less real because of it.

There is more to be said about each of these assumptions. But both are widely, though not universally, accepted within the literature. For example, Schwitzgebel has at times questioned whether introspection is sufficiently reliable for the purposes of consciousness science (2011), though see Mckilliam (forthcoming). And Carruthers has questioned whether consciousness really is a natural phenomenon (2019). Rather than attempt to defend either of these assumptions here, I want to explore the implications of accepting them.

The paper proceeds as follows. The first half of the paper aims to develop a particular view of the nature of inferences about consciousness in others—they are analogical inferences grounded in explanatory considerations. The second half argues that, if this is right, then mechanisms do matter for inferences about consciousness, and that the relevant mechanisms are cognitive mechanisms.

I begin by introducing a distinction between explanatory abduction and analogical abduction—analogical inferences grounded in explanatory considerations. In section three I make the case that inferences about consciousness in others are analogical abductions. In section four I explain how the iterative application of inference to the best explanation to home in on and leverage regularities in nature—sometimes dubbed iterative natural kind reasoning (Shea and Bayne 2010; Bayne, Seth, *et al.* 2024; Mckilliam 2024)—can resolve traditional objections to arguments from analogy. In section five I explain why, if I am right that inferences about consciousness are analogical abductions, then evidence of profound mechanistic difference undercuts behavioural evidence of consciousness. And in section six I suggest that we can make progress on the question of which mechanistic differences undercut behavioural evidence by considering which details best explain why the cluster of consciousness-related capacities hang together.

## Analogical abduction and explanatory abduction

Traditionally, analogical reasoning and inference to the best explanation have been taken to provide alternative solutions to the problem of other minds (Pargetter 1984; Hyslop 1995; Avramides 2020). This strikes me as unfortunate. To my mind, the most attractive account requires combining the two (Melnyk 1994). Inferences about consciousness in others are analogical inferences grounded in explanatory considerations. To make sense of this idea, it will be useful to distinguish two modes of inference involving appeals to explanatory considerations that I call analogical abductions and explanatory abductions. Both appeal to explanatory considerations for justification, but they do so in different ways. Let me explain.

Sometimes we infer the existence of a phenomenon that we have not observed on the grounds that its existence would neatly explain the data we have observed. Suppose I wake to find that the street is wet and infer that it rained last night. I did not directly observe it raining, and the fact that the street is wet does not entail that it rained last night—perhaps the neighbours decided to have a 3am water fight. Nonetheless, the hypothesis that it rained last night explains why the street is wet and does so better than the alternatives. As a result, I am justified in concluding that it probably rained last night. In this case, the phenomenon I’m drawing an inference about is doing explanatory work. The rain explains why the streets are wet.

Explanatory inference is a common strategy in science. When 19th-century mathematicians Urbane Le Verrier and John Couch Adams inferred the existence of Neptune, they did so because the existence of an eighth planet provided the best explanation of the peculiarities of Uranus’ orbit (Baum and Sheehan 2013). When Joseph Goldberger inferred that pellagra—a disease characterized by diarrhea, dermatitis, dementia and ultimately death—was caused by nutritional inadequacies and not by the transmission of germs, he did so on the grounds that that hypothesis best explained why his attempts to infect well-fed individuals failed (Bollet 1992). And when neurologists infer the location of a lesion based on specific behavioural deficits, they do so on the grounds that, together with background theory, a lesion in that location would explain the deficits. In each case, the phenomenon that one is drawing an inference about is doing explanatory work.

By contrast, when we make analogical abductions, the property, event, or thing we are drawing an inference about is not doing explanatory work. Examples help illustrate this point. Suppose you are in the market for a new book. You walk into a bookstore and pick up a copy of the intriguingly titled “50 Shades of Grey” by E.L. James. You read the blurb and flick through some of the pages, but ultimately decide that this is not the book for you; the characters, the plot, and the writing style, just don’t appeal. You then turn to the next book on the shelf, also titled “50 Shades of Grey” and written by the same author. This time, you do not pick it up and flick through the pages. You infer that it too will not be to your liking and instead search for a book with a different title.

This seems to be a perfectly reasonable thing to do. But notice that in this case the properties you are projecting—characters, plot structure, writing style, and so on—are not doing explanatory work—at least not if we accept the orthodox view that explanations cite causes (Salmon 1984; Woodward 2003; Lipton 2004). The fact that the book has a plot and writing style that you will not enjoy does not explain why you did not pick it up. Rather, it is your belief that the book has a plot and writing style that you will not enjoy that explains why you did not pick it up. What then, justifies your belief?

Again, explanatory considerations do, but in this case, rather than doing explanatory work, the properties being projected come along for the explanatory ride. In the case of books, we know there to be a similarity-generating mechanism explaining why some of the properties of books—title, author, plot structure, characters, and writing style—reliably cluster together. Books, or rather, the stories they contain, are an example of what Ruth Millikan has called historical kinds—they descend via a process of copying from a single point of creation (Millikan 1999; 2000; see also Sober 2000). This copying process explains why books that have the same title and author tend also to have the same characters, and plot structure. So, if you observe that one book titled “50 Shades of Grey” has a writing style and plot structure that is not to your liking, you can justifiably infer that other books titled “50 Shades of Grey” will also have a writing style and plot structure that is not too your liking.

This inference is defeasible of course. If you were to learn that E.L. James authored two very different books both titled “50 Shades of Grey”, this would undercut your inference. But unless you have a reason to think that “50 Shades of Grey” is unusual in this regard, it is reasonable to prefer the simpler hypothesis that the two books will have the same plotline, characters, and writing style. In this case, even though the properties being projected is not doing explanatory work, explanatory considerations—simplicity, for example—are still at play in justifying your inference.

Consider another example. Suppose you stumble across a chilli bush with a bounty of bright red chillies ready for harvest. You pick one, take a bite, and note that it is quite spicy. You infer that the rest of the chillies on the bush will be similarly spicy. This seems like a perfectly reasonable inference to make, but again, the property you are projecting is not doing explanatory work—the chillies being spicy does not explain why they are all growing on the same bush. Nonetheless, your inference can still be justified by appealing to explanatory considerations. You know that spiciness is primarily determined by the presence and quantity of the chemical, capsaicin. And you know there to be a similarity-generating mechanism that would neatly explain why all the chillies on the bush have a similar amount of capsaicin—namely, they have all grown on the same plant under the same growing conditions. So, you infer that each of the chillies will contain a similar amount of capsaicin and will therefore be similarly spicy. Again, this inference is defeasible. If you were to learn that someone had grafted a different strain of chillies onto one side of this plant this may defeat your inference. But in the absence of a reason to think that this chilli bush is unusual in this regard, you are justified in accepting the simpler hypothesis.

## Inferences about consciousness in others are analogical abductions

Turning now to consciousness, are inferences about subjective experience explanatory abductions or analogical abductions? Does subjective experience do explanatory work, or does it come along for the explanatory ride? Those who appeal to inference to the best explanation when drawing inferences about consciousness are often unclear on this. Indeed, sometimes they appear to jump between the two. Consider Michael Tye’s discussion.

I see you lying on the road with blood trickling down your arm and your bicycle with a bent front wheel. You are groaning and you have a tense look on your face. You ask me if I have any aspirin… The pain hypothesis [the hypothesis that you are in pain] offers a straightforward explanation of your behavior, an explanation better than any other available. So I am justified in embracing it. (Tye 2017: 58)

In this passage, Tye is clearly endorsing the claim that ‘pain’ is doing explanatory work. The hypothesis that you are in pain provides the best explanation of your behaviour. However, as Tye goes on to explain, the concept ‘pain’ has two aspects to it.

On the one hand, there is the unpleasant, subjective felt quality: pains hurt. On the other hand, there is the causal role pain plays: pain is a state that is normally caused by bodily damage and normally causes a variety of psychological and physical reactions. (Tye 2017: 58)

Which aspect of pain does Tye take to be doing explanatory work? At times he seems to suggest that the felt quality—the subjective ouchiness—does explanatory work. He writes: “It is a truism that we want to get rid of pain and that we behave in ways that reduce or eliminate it *because of how it feels*” (2017: 75 my italics). And elsewhere: “our experiences have distinctive functional roles as well as subjective aspects. But they function in the ways they do precisely *because* of how they feel” (2017: 93). But in other places, he seems to suggest that subjective experience merely comes along for the explanatory ride. This is clearest in his response to the solipsistic hypothesis that only he has subjective experience. He notes that “the hypothesis that you undergo ersatz pain [a functional analogue of pain without the pain experience] is not as good an explanation of the facts as the hypothesis that you experience pain just as I do” (Tye 2017: 60–61). In this passage, Tye appeals to explanatory considerations to justify belief in the hypothesis that other humans are like him in the sense that they have subjective experiences. But in this case, subjective experience isn’t recruited to do work in explaining behaviour. Rather, it comes along for the explanatory ride.

So, which is it? Is subjective experience a theoretical construct postulated to do explanatory work? Or does it come along for the explanatory ride? I am inclined to say that subjective experience comes along for the explanatory ride. To see why you might want to say that, ask yourself the following question: Would it be rational for an intelligent but non-conscious being to postulate the existence of subjective experience to explain the behavioural capacities of conscious creatures? It is not clear that it would. It might postulate all manner of psychological constructs—beliefs, desires, predictions, action plans, self-models, and so on—but given the widely held assumption that the behaviour of conscious beings can be exhaustively explained in psychological, neurocognitive terms that are purely third-personal, from the perspective of a non-conscious agent, postulating that some of these psychological states also feel like something from a subjective perspective would be superfluous. It would be assuming the existence of an additional entity that does no explanatory work. When deciding between a purely third-person account of human psychology and one that also postulates that some psychological states also feel like something from a first-person perspective, a non-conscious agent should presumably prefer the former.

The same is not true for us of course. We are conscious. We know from our own introspective reflections that some psychological states feel like something from a subjective perspective. And so, we will have reason to prefer the explanations that take some of these psychological states to be accompanied by subjective experiences.

To be clear, in saying that subjective experience comes along for the explanatory ride I do not mean to imply that we cannot appeal to subjective experience in causal explanations. Nor do I mean to suggest that subjective experience is epiphenomenal. If it turns out that subjective experiences *just are* brain states, then they are not epiphenomenal. All that is being claimed here is that subjective experience is not something we postulate the existence of in order to do explanatory work.

One might object at this point, that some subjective experience can do explanatory work in some cases, in particular in those cases involving reports of what Chalmers’ calls “problem intuitions” (Chalmers 2018; see also Schneider 2019). Perhaps a non-conscious agent would postulate the existence of subjective experience to explain why we humans sometimes report being puzzled about the possibility of spectrum inversions for example. But even if subjective experience can do explanatory work in the limited range of cases involving reports of “problem intuitions”, in most cases, particularly those relevant to consciousness in non-human animals, subjective experience is best treated as coming along for the explanatory ride.

If that’s right, then at for the least the majority of cases, inferences about consciousness in others are analogical abductions. We reflect on our own case, recognize that in us, some psychological states feel like something from a subjective perspective, and then we project this feeling like something from a subjective perspective onto other systems that we judge to be in the same psychological state on the grounds that their being in that psychological state best explains their behaviour.[[3]](#footnote-3)

## How explanatory iteration can help resolve the traditional objections to arguments from analogy

While researchers are generally happy to allow for analogical inferences about books and chillies, many doubt whether the same strategy can be extended to the case of consciousness (Hyslop 1995; Tye 2017; Bartha, 2019). In the everyday cases, like those involving books and chillies, we already know which similarities are relevant and which are not. We know this in part because we have illuminating explanations of the properties in question, and in part because we can appeal to previous successful inferences involving chillies and books. In the case of consciousness, not only do we lack an illuminating explanation of the phenomenon, we also only have access to a single case—our own. How then, are we supposed to figure out which similarities are relevant, and which are not?

I take it that this is what Tye has in mind when he writes:

“The trouble with the analogical inference to other minds is that the evidence pertains only to me, whereas the conclusion pertains to others, to people who are not me. How can I be confident that the similarities between myself and other people are relevant, given that all the evidence pertains to me alone?” (Tye 2017: 56)

Once we allow explanatory considerations to guide our analogical inference, these concerns begin to lose their bite. I happen to be 183cm tall and physically located in Melbourne. But neither of these facts seem to be relevant to whether or not I am conscious. Although we cannot definitively rule out the possibility that I was not conscious when I was a child and shorter than I am now, or that I might suddenly cease to be conscious when I move to Taiwan later this year, explanatory considerations strongly speak against odd-ball hypotheses like these.

This is a start, but there are a wide range of factors that are not so clear cut. How do we address those? A number of papers have emerged in the last few years exploring how iterative natural kind reasoning can allow us to home in on a cluster of behavioural markers of consciousness (Shea and Bayne 2010; Bayne and Shea 2020; Birch 2022; Bayne, Seth, *et al.* 2024; Brown, Paul, and Birch 2024; Mckilliam 2024). The general idea is that through reflecting on one’s own case we can come to notice a host of correlations between consciousness and various cognitive abilities. States that feel like something to be in are generally also available for thought, planning, verbal report, and so on. What might explain this correlation? One plausible hypothesis is that our brains engage in a distinct kind of information processing—conscious processing—that both feels like something from a subjective perspective and facilitates each of these cognitive abilities.

We can further refine the contours of the cluster of capacities that are facilitated by conscious processing by leveraging masking paradigms to probe the differences between conscious and unconscious perception in humans. For example, in us humans, the ability to i) override priming effects in the performance of goal directed actions (Debner and Jacoby 1994), ii) rapidly learn pattern reversals (Travers, Frith, and Shea 2018), and iii) integrate information across sensory modalities (Palmer and Ramsey 2012) all seem to require, or at least be facilitated by conscious processing—masking a stimulus so that it is registered by the brain but not consciously perceived systematically switches off each of these capacities. We can further fine-tune the list of behavioural markers by investigating whether a subset of this cluster also come online all at once in infant development and across phylogenetic time. In this way we can home in on a cluster of behavioural capacities that are the relevant ones for making analogical abductions about consciousness.[[4]](#footnote-4)

The picture that is emerging here is that there is a distinct kind of information processing that our brains engage in that can serve as a similarity-generating mechanism for drawing analogical abductions about consciousness. If a creature possesses a cluster of behavioural capacities that are systematically switched-off by masking in roughly the same way that they are in us, that is compelling evidence that they too engage in the mode of information processing that in our own case happens to feel like something from a subjective perspective.

There are already numerous papers leveraging this strategy to draw inferences about consciousness in infants (Bayne, Frohlich, *et al.* 2023) and non-human animals (Birch 2022). What remains relatively unclear however, is how mechanistic details fit into this picture. For the most part, those developing this strategy have focussed on behavioural markers.[[5]](#footnote-5) One notable exception is Crump and Birch, who suggest that neuroscientific evidence of profound mechanistic differences may undercut the cluster of behavioural markers as evidence of consciousness (Crump and Birch 2022: 122). In the remainder of the paper, I want to unpack why this might be the case and sketch one way of addressing the question of which mechanistic details might qualify as revealing profound mechanistic differences.

At this point, it will be useful to introduce some terminology from the literature on defeasible reasoning (Pollock 1987). An *undercutting* defeater provides a reason to distrust what is typically a reliable evidential link. Suppose I was to lend you my car for the weekend and upon getting in you noticed that the fuel gauge was reading low. Naturally, you would infer that the car was running low on fuel and that you would need to stop at a petrol station before going too far. But suppose I were to tell you that the fuel gauge is faulty and that it has been stuck on empty for months. This new information would *undercut* the evidential link between the fuel gauge and the fuel tank. It does not actually give you a reason to think that there *is* fuel in the tank. The tank might, in the end, be empty after all. All it tells you is that the fuel gauge is not a reliable guide to how much fuel is in the tank.

In the next section, I explain why, if inferences about consciousness in others are analogical abductions, then profound mechanistic difference is like this. Learning that a system operates via profoundly different mechanisms from our own undercuts the evidential link between behavioural capacities and subjective experience.

## Why profound mechanistic difference undercuts behavioural evidence of consciousness

It is often useful to group things together in a way that explicitly ignores similarity-generating mechanisms and instead focuses on what things can door on the functional roles they play. If I am stranded on a desert island, all I care about is that I have some way off. Whether that is by jet-ski, hot air balloon, or teleportation, I do not really mind. For this reason, we might say that jet-skis, hot-air balloons and teleportation devices all belong in the category ‘things that can get me off the island’. Similarly, if my heart’s natural pacemaker is on the fritz, all I care about is whether or not the artificial one can play the same role and successfully keep my heart beating. In this case, we might think of both my heart’s natural pacemaker and the artificial one as members of the same category ‘things that can keep my heart beating’.

Given that these categories are formed with explicit indifference to similarity-generating mechanisms, it is natural to suppose that they will not be good candidates for analogical inferences. And indeed, they are not. There is very little that one can infer about jet-skis based on observations of hot-air balloons, and there is very little that one can infer about my heart’s natural pacemaker based on observations of artificial ones.

Some authors have taken a hard line on this point and argued that categories that abstract away from mechanistic detail do not licence any analogical inferences whatsoever. For example, Ruth Millikan has suggested that the fact that “a variety of objects all exhibit the same functionalist property for the same reason would not seem, by itself, to imply that they are alike in any other respect” (1999: 59). On reflection though, this is implausible. There is a limited range of analogical inferences that one can justifiably make based on, say, functional similarity, due to the fact that membership in a functionally defined category requires being capable of performing the function in question.

This is true even for Jerry Fodor’s classic example—money (1974). Fodor pointed out that a wide range of things can be used as money in an economic exchange. Cowry shells, slips of paper, pieces of metal, strings of wampum, and bits on a computer have very little in common besides the fact that they have all been used as currency at some point or another. However, even here some analogical inferences are supported. To successfully play the role of money, something needs to be tricky to obtain outside economic interactions. If I could gather fistfuls of cowry shells just by strolling down to the beach, why would I trade my valuable apples for them? And it also needs to be relatively stable over time. Ideal forms of money not only enable people to turn one thing into another through trade, they also enable people to store and accumulate wealth. So, even if you have no idea what strings of wampum are, after learning that they were used as currency by indigenous populations in North America, you can reasonably infer that they are both relatively stable over time and were not easy to come by outside of economic interactions for indigenous populations in North America.

The general idea here is that categories that abstract away from details about similarity generating mechanisms do support some analogical inferences, but only about properties that are constraints on membership into the category in question.

Turning now to inferences about consciousness. We saw above that researchers are homing in on a cluster of behavioural capacities that can serve as markers of consciousness on the grounds that in our own case these capacities are systematically correlated with subjective experience. We might then, take this cluster of behavioural markers to define a category of potentially conscious creatures that explicitly abstracts away from similarity generating mechanisms. Should we project subjective experience to all members in this category? Or only to the subset of the category that possesses the cluster of markers by virtue of mechanisms that are like our own?

To answer this question, we need to ask whether subjective experience is a constraint on possessing this cluster of behavioural markers. It doesn’t seem to be.

There are two avenues one might take here to try to establish that subjective experience isa constraint on possessing this cluster of behavioural markers. In my view, neither is very promising. First one might try to argue that subjective experience is a constraint on facilitating cognition in the way that conscious processing does in us on *a priori* grounds. We can know that being unmarried is a constraint on being a bachelor just by reflecting on the meaning of the terms involved. And we can know that being three-sided is a constraint on being a triangle just by reflecting on the meaning of the terms involved. One might try to argue that we can know that feeling like something from a subjective perspective is a constraint on facilitating cognition just by reflecting on the concepts involved.

There are familiar reasons from the literature for thinking that this strategy is unlikely to succeed. If we can know, just by reflecting on the concepts involved, that feeling like something from a subjective perspective is a constraint on possessing this cluster of behavioural markers of consciousness, then it follows that zombies—beings that are physically and functionally just like us but who lack any subjective experiences—are conceptually incoherent. Zombies might not be metaphysically possible, but arguably they are conceptually coherent (Chalmers 1996; Carruthers 2019; Seth 2021).

Alternatively, one might suggest that we may come to learn via empirical investigation that subjective experience is a constraint on possessing this cluster of behavioural markers. The difficulty with this strategy is that it is not entirely clear how one would go about doing so. Our initial source of evidence about consciousness comes from reflecting on our own case. From there we draw inferences about consciousness in others based on analogical abductions. But analogical abductions about consciousness, so the argument above suggests, would only be warranted in creatures that have mechanisms profoundly different from our own if subjective experience was a constraint on possessing these behavioural capacities. It is not clear how one could break out of this in a non-question-begging way.

I do not think this conclusion is terribly controversial. All it implies is that behavioural evidence of consciousness is defeasible, and this has long been accepted. If a behavioural analogue of us turned out to be a marionette (Tye 2017) or an AI gerrymandered specifically to possess the capacities in question (Shevlin 2021; Dung 2023) that would undercut the behavioural evidence. But what the analysis in this section does is suggest an underlying justification for these intuitions and generalizes beyond the specific cases. It suggests that learning that a system’s behaviour is governed by profoundly different mechanisms than our own severs the evidential link between behavioural evidence and subjective experience.

It is worth unpacking this idea a little more by considering an example. Suppose we find that a particular class of artificial agent or non-human animal possessed a cluster of cognitive abilities associated with conscious processing in us and that these capacities can be collectively switched off through masking. This provides *prima facie* compelling evidence that the agents engage in the kind of processing that, in our own case feels like something from a subjective perspective, and so, provides *prima facie* warrant for inferring that they are conscious. But suppose we were to then look at their ‘brain’ and find only a single mode of processing whereas in our own case conscious and unconscious processing are distinct. This latter fact would undercut the behavioural markers as evidence of consciousness because we would no longer have good reason to suppose that this cluster of capacities is underpinned by the same kind of mechanism—the one that serves as a similarity generating mechanism in this case—that underpins them is us.

In this section I have explained why evidence of profound mechanistic difference undercuts behavioural evidence of cognitive abilities. In the final section I will say a little more about what kinds of mechanistic differences might qualify as ‘profound’ in this context.

## Which mechanistic differences undercut behavioural evidence?

Researchers disagree about which mechanistic differences might undercut behavioural similarity as evidence of consciousness. Some seem to think that none would. Doerig and colleagues (2019) for instance, suggest that a feedforward network that computes the same input-output function as a conscious agent—in other words, one that is behaviourally indistinguishable from a conscious agent—might be conscious (see also Herzog, Schurger, *et al.* 2022). Others suspect that details about how neurons work may be crucial (McLaughlin 2003, 2019; Seth 2021). Seth, for instance, speculates that the fact that our brains are composed of cells that are themselves self-maintaining systems, may be important (Seth 2021: 255). Others express views that sit somewhere in between. Godfrey-Smith, for instance, suggest that information processing alone isn’t enough—we won’t be able to upload our minds into the cloud—but also that self-maintaining neurons may not be essential (2020). Perhaps it is the cognitive mechanisms that matter (Shevlin 2021)?

In this final section I want to do two things. First, I will appeal to some recent developments within the mechanistic philosophy of science to clarify what it might mean for two systems to operate via similar cognitive mechanisms. And second, I will provide some initial motivation for the view that it is cognitive mechanisms that matter for inferences about consciousness.

In discussions within the philosophy of mind, the idea of mechanistic similarity has typically been taken to imply similarity of substrate. On this usage, humans and robots operate via profoundly different mechanisms because the latter is composed of silicone chips while our brains are composed of carbon-based neurons (Schwitzgebel 2015; Tye 2017; Schnider 2019). This made sense in the context of traditional debates between reductionists and anti-reductionists. Reductionists held that psychological concepts would eventually be reduced to (Bickel 2003) or replaced by (Churchland 1981) neuroscientific ones. Anti-reductionists insisted that psychological concepts picked out functional roles that are autonomous from the gory mechanistic details (Fodor, 1968, 1974; Putnam 1975). However, since the cognitive neuroscience revolution and the rise of the new mechanistic philosophy of science, the battle between reduction and autonomy has given way to integration. Rather than trying to reduce or replace the models and constructs of psychology and cognitive science with neuroscientific models and constructs or, alternatively, treating them as entirely autonomous from lower-level mechanisms the objective is to integrate them into multi-level mechanistic explanations (Piccinini and Craver 2011; Boone and Piccinini 2016a; Piccinini 2020). And this calls for a more flexible use of the term ‘mechanism’, one that is not, by definition, wedded to physical substrate.

From the perspective of mechanistic cognitive neuroscience, the computational models and box-and-arrow diagrams of cognitive psychology canqualify as mechanistic models just so long as “the variables in the model correspond [at least roughly] to components, activities, properties, and organizational features of the target mechanism” (Kaplan and Craver 2011: 611). Kaplan and Craver call this the model-to-mechanism-mapping criterion, or 3M for short. On this view, psychological and computational models may be idealizations (Boone and Piccinini 2016b), they may be partial and incomplete mechanism sketches (Piccinini and Craver 2011), but so long as they satisfy 3M, they qualify as mechanistic models (see also Piccinini 2020).

I suggest that 3M provides a useful tool for characterizing a more flexible conception of mechanistic similarity—one that allows for the integration of psychology and neuroscience without falling back into either reductionism or anti-reductionism. In particular, it allows us to tease apart mechanistic from non-mechanistic versions of functionalism by considering the different criteria one might adopt when deciding whether or not to attribute particular cognitive states to a system. As far as non-mechanistic functionalists are concerned, cognitive states need only be visible from the perspective of behavioural psychology. So long as the components in a cognitive model serve as useful control variables for predicting and controlling behaviour, then the system really does possess those cognitive states (Campbell 2010; Weiskopf 2017). Mechanistic functionalists on the other hand, suggest that this isn’t quite enough. In addition, we need to be able to ‘pop the hood’ so to speak and map the variables of these computational/cognitive models, at least loosely, onto the components and activities of the system in question (Godfrey-Smith 2008; see also Piccinini and Craver 2011; Kaplan and Craver 2011; Piccinini 2020).

An example will help illustrate this point. At an abstract level, mouse traps can be described as systems that have a component that functions as a mouse detector suitably connected to another component that functions as a mouse detainer such that when a mouse is detected it is subsequently detained. This abstract and idealized functional model can be used to describe very different kinds of mouse traps in a way that satisfies 3M. For example, in snap traps, the kind familiar from Tom and Jerry cartoons and physical comedy sketches, a delicately weighted trigger plays the role of mouse detector, and a spring-wound bar functions as a mouse detainer. But snap-traps are not the only kind of trap that can be described by this abstract functional model in a way that satisfies 3M.

I once spent 6 weeks living in a tent in the woods during a mouse plague. One of my fellow campers had the idea of spearing an empty beer can with a stick, covering it in peanut butter, and balancing it on top of a bucket. His thought was that as the mice crawled onto the can to get at the peanut butter, its weight would cause the can to spin, tipping the mouse into the bucket. Many of us were dubious as to whether it would work, but by morning there were 18 mice in the bucket. Bucket traps can also be modelled by the abstract functional description above in a way that satisfies 3M. In this case, the skewered beer can functions as the mouse detector while the bucket plays the role of mouse detainer. So, despite their many differences, there is a minimal sense in which snap-traps and bucket-traps are mechanistically similar—they can both be described by the same functional model in a way that satisfies 3M.

But not all mouse traps meet this criterion. Glue traps, for instance, arguably do not. Glue traps are sticky sheets that exude, as one brand proudly states, a “tasty meat scent”. The mice are attracted by the scent and get detained by the sticky glue. Glue traps have a mouse detainer—the glue—and a mouse attractor—the tasty meat scent. But arguably they do not have anything that can usefully be treated as a mouse detector. As a result, they can not be described by the abstract model above in a way that satisfies 3M.[[6]](#footnote-6)

3M can serve as a useful minimal criterion for those engaged in comparative consciousness studies. If the cognitive model of conscious processing in us cannot be mapped onto an animal or artificial agent in a way that satisfies 3M, then that is reason to refrain from attributing consciousness to it even if it displays sophisticated behavioral capacities. Alternatively, if an animal or artificial agent has a brain that is strikingly different from ours in some respects, but is similar enough to satisfy 3M, then that is reason to take behavioral markers of consciousness to be reliable indicators. In fact, 3M is arguably already at play in the literature. For example, although fish lack anything homologous to a neocortex (Key 2016), the global workspace model can be mapped reasonably well onto the fish brain—arguably well enough to satisfy 3M (Zacks and Jablonka 2023).

But is similarity of cognitive mechanisms really similarity enough? Or might lower-level details about the neural implementation of these cognitive mechanisms also be important? Part of what makes answering this question hard, according to Block, is that we cannot straightforwardly appeal to inference to the best explanation in the face of the explanatory gap (2002). If neural models provided illuminating explanations of subjective experience while computational models did not, or vice versa, then we would have good reason to think that consciousness comes and goes with the details that provide the best explanation. But, so the thought goes, neither cognitive models nor neural models provide terribly illuminating reductive explanations of subjective experience.

Philosophers of mind have attempted to make progress on this question by constructing thought experiments involving hypothetical space aliens and group agents to reveal the counter intuitive implications of rival views (Block 1978; Schwitzgebel 2015). But it is not clear that this approach is appropriate. On the one hand, researchers appear to have conflicting intuitions, and so appealing to intuitions is unlikely to lead to consensus. But even if intuitions did all point in the same direction here, it is not clear how much evidential weight intuitions about consciousness in others should really be given. This is not because intuitions are always suspect. Introspective judgements that have been honed through practice or evolution are likely to be pretty reliable. Think, for example, of Magnus Carlson’s intuitions about chess moves or a chimpanzee’s intuition about whether a branch will hold its weight. In each case feedback from the environment all but ensures that the intuitions will be reasonably reliable. Our intuitions about consciousness in hypothetical space aliens, group agents, and artificial intelligence do not seem to be subject to similar environmental constraints. And so, it is not clear that we should really be trusting our intuitions in this context.

The analysis I have provided here suggests another route forward. It suggests that we may be able to leverage inference to the best explanation to make progress on this question even without having to confront the explanatory gap. Rather than considering whether cognitive or neural models better explain subjective experience, we can instead ask which better explains why the cluster of consciousness-related capacities hang together in the way that they do. Is the clustering better explained by a model that includes details at the neural level? Or by a more abstract cognitive model couched in terms of cognitive psychology? If cognitive models better explain the cluster, then we have a principled reason to supposed that cognitive mechanisms are the relevant ones for drawing inferences about consciousness and that neural details may be inessential. Alternatively, if models that include details at the neuronal level turn out to better explain the cluster, then we would have reason to suppose that those details are important for inferences about consciousness.

Exactly how this will play out is an open question. And answering it is unlikely to be straightforward. One potential complication here is that the number of consciousness-related capacities a model explains may covary to some extent with the fineness of grain (Shevlin 2021; Boyle 2023). But with a bit of luck, we may be able to leverage this strategy to converge toward consensus on which mechanisms matter for inferences about consciousness in a principled manner.

## Conclusion

In this paper I have explored an epistemology driven approach to the question of whether mechanisms matter for inferences about consciousness and concluded that they do. I have argued that inferences about subjective experience in others are best understood as analogical abductions and sketched how natural kind reasoning can resolve traditional objections to arguments from analogy. I have also explained why, if inferences about consciousness are analogical abductions, then behavioural similarity is undercut by evidence of profound mechanistic difference. Behavioural similarity is only evidence of consciousness in others in so far as it is evidence of the same mode of information processing—conscious processing—that explains why the cluster of consciousness-related capacities reliably come and go together. And I have sketched a strategy via which we might resolve disagreements about how much mechanistic similarity is sufficient. If cognitive models provide the best explanation for the cluster of consciousness related capacities, then we have reason to suppose that subjective experience comes and goes with cognitive mechanisms. Whether cognitive models do in fact best explain the cluster is an open question, but it is a resolvable one.

## Acknowledgements

I would like to thank Tim Bayne, Eric Schwitzgebel, Jonathon Birch, two anonymous reviewers and members of the Monash Centre for Consciousness and Contemplative Studies for helpful feedback on an earlier version of this paper.

## Funding Information

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

## ORCID iD

https://orcid.org/0000-0001-8223-3817

## References

Avramides, A (2020) ‘Other Minds’, in E N Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2020/entries/other-minds/>.

Bartha, P (2019) ‘Analogy and Analogical Reasoning’, in E N Zalta, ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2022/entries/reasoning-analogy/>.

Baum, R P and Sheehan, W (2013) *In Search of Planet Vulcan: The Ghost in Newton’s Clockwork Universe*. Springer.

Bayne, T and Shea, N (2020) ‘Consciousness, Concepts, and Natural Kinds’. *Philosophical Topics* **48**: 65–84. [doi:10.5840/philtopics20204814](https://doi.org/10.5840/philtopics20204814).

Bayne, T, Frohlich, J, Cusack, R, Moser, J, and Naci, L (2023) ‘Consciousness In the Cradle: On the Emergence of Infant Experience’. *Trends in Cognitive Sciences* **27**: 1135–49. [doi:10.1016/j.tics.2023.08.018](https://doi.org/10.1016/j.tics.2023.08.018).

Bayne, T, Seth, A K, Massimini, M, Shepherd, J, Cleeremans, *et al.* (2024) ‘Tests for Consciousness in Humans and Beyond’.*Trends in Cognitive Sciences* **28**: 454–66.[doi:10.1016/j.tics.2024.01.010](https://doi.org/10.1016/j.tics.2024.01.010).

Bickle, J (2003) *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Springer. [doi:10.1007/978-94-010-0237-0](https://doi.org/10.1007/978-94-010-0237-0).

Birch, J (2022) ‘The Search for Invertebrate Consciousness’. *Noûs* ***56***: 133–53. [doi:10.1111/nous.12351](https://doi.org/10.1111/nous.12351).

Block, N (1978/2007) ‘Troubles with functionalism’. In *Consciousness, Function, and Representation.* The MIT Press. [doi:10.7551/mitpress/2111.003.0006](https://doi.org/10.7551/mitpress/2111.003.0006%C2%A0).

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

Bollet, A J (1992) ‘Politics and Pellagra: The Epidemic of Pellagra in the US in the Early Twentieth Century. *Yale Journal of Biology and Medicine* ***65***: 211–21.

Boone, W and Piccinini, G (2016a) ‘The Cognitive Neuroscience Revolution’. *Synthese* **193:** 1509–34. [doi:10.1007/s11229-015-0783-4](https://doi.org/10.1007/s11229-015-0783-4).

Boone, W and Piccinini, G (2016b) ‘Mechanistic Abstraction’. *Philosophy of Science* **83**: 686–97. [doi:10.1086/687855](https://doi.org/10.1086/687855).

Boyle, A (2023) ‘Disagreement & Classification in Comparative Cognitive Science’. *Noûs* ***58***: 825–847. [doi:10.1111/nous.12480](https://doi.org/10.1111/nous.12480).

Brown, S A B, Paul, E S, and Birch, J (2024) ‘To Test the Boundaries of Consciousness, Study Animals’. *Trends in Cognitive Sciences* **10**: 874–5. [doi:10.1016/j.tics.2024.06.003](https://doi.org/10.1016/j.tics.2024.06.003)

Brown, S A B (2024) ‘How to Get Rich From Inflation’. *Consciousness and Cognition* **117**: 103624. [doi:10.1016/j.concog.2023.103624](https://doi.org/10.1016/j.concog.2023.103624).

Butlin, P, Long, R, Elmoznino, E, Bengio, Y, Birch, J, *et al.* (2023) ‘Consciousness in Artificial Intelligence: Insights From the Science of Consciousness’. [doi:10.48550/arXiv.2308.08708](https://doi.org/10.48550/arXiv.2308.08708).

Campbell, J (2010) ‘Independence of Variables in Mental Causation’. *Philosophical Issues* **20**: 64–79. [doi:10.1111/j.1533-6077.2010.00178.x](https://doi.org/10.1111/j.1533-6077.2010.00178.x).

Carruthers, P (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press.

Chalmers, D (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Chalmers, D (2018) ‘The Meta-problem of Consciousness’. *Journal of Consciousness Studies* **25**: 6–61.

Churchland, P M (1981) ‘Eliminative Materialism and Propositional Attitudes’. *The Journal of Philosophy* **78**: 67–90. doi:[10.2307/2025900.](https://doi.org/10.2307/2025900.)

Crump, A, and Birch, J (2022) ‘Animal Consciousness: The Interplay of Neural and Behavioural Evidence’. *Journal of Consciousness Studies* **29**: 104-128. doi:[10.53765/20512201.29.3.104](https://doi.org/10.53765/20512201.29.3.104).

Debner, J A, and Jacoby, L L (1994) ‘Unconscious Perception: Attention, Awareness, and Control’. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**: 304–317. doi:[10.1037/0278-7393.20.2.304.](https://doi.org/10.1037/0278-7393.20.2.304)

Doerig, A, Schurger, A, Hess, K, and Herzog, M H (2019) ‘The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness’. *Consciousness and Cognition* **72**: 49–59.  [doi:10.1016/j.concog.2019.04.002.](https://doi.org/10.1016/j.concog.2019.04.002)

Dung, L (2023) ‘Tests of Animal Consciousness are Tests of Machine Consciousness’. *Erkenntnis*. [doi:10.1007/s10670-023-00753-9](https://doi.org/10.1007/s10670-023-00753-9).

Fodor, J A (1968) *Psychological Explanation: An Introduction to the Philosophy of Psychology*. Random House.

Fodor, J A (1974) ‘Special Sciences (or: The Disunity of Science as a Working Hypothesis)’. *Synthese* **28**: 97–115.  [doi:10.1007/bf00485230.](https://doi.org/10.1007/bf00485230)

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

Godfrey-Smith, P (2008) ‘Reduction in Real Life’. In J Hohwy and J Kallestrup, eds., *Being Reduced: New Essays on Reduction, Explanation, and Causation:* 52-74. Oxford University Press. [doi:10.1093/acprof:oso/9780199211531.003.0004](https://doi.org/10.1093/acprof%3Aoso/9780199211531.003.0004).

Godfrey-Smith, P (2020) *Metazoa: Animal Life and the Birth of the Mind*. Farrar, Straus and Giroux.

Herzog, M H, Schurger, A, and Doerig, A (2022) ‘First-Person Experience Cannot Rescue Causal Structure Theories From the Unfolding Argument’. *Consciousness and Cognition* **98**: 103261. [doi:10.1016/j.concog.2021.103261](https://doi.org/10.1016/j.concog.2021.103261).

Hohwy, J (2004) ‘Evidence, Explanation, and Experience: On the Harder Problem of Consciousness’. *Journal of Philosophy* **101**: 242–254.  [doi:10.5840/jphil2004101525.](https://doi.org/10.5840/jphil2004101525)

Hyslop, A (1995) *Other Minds*. Springer. [doi:10.1007/978-94-015-8510-1.](https://doi.org/10.1007/978-94-015-8510-1%C2%A0)

Irvine, E (2017) ‘Explaining What?’ *Topoi* **36**: 95–106.  [doi:10.1007/s11245-014-9273-4.](https://doi.org/10.1007/s11245-014-9273-4)

Kaplan, D M, and Craver, C (2011) ‘The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective’. *Philosophy of Science* **78**: 601–627.  [doi:10.1086/661755.](https://doi.org/10.1086/661755)

Key, B (2016) ‘Why Fish Do not Feel Pain’. *Animal Sentience*.  [doi:10.51291/2377-7478.1011.](https://doi.org/10.51291/2377-7478.1011)

Lipton, P (2004) *Inference to the Best Explanation,* 2nd edition. Routledge. doi:[10.4324/9780203470855.](https://doi.org/10.4324/9780203470855)

McLaughlin, B P (2003) ‘A Naturalist-Phenomenal Realist Response to Block's Harder Problem. *Philosophical Issues* **13**: 163–204. doi:[10.1111/1533-6077.00011.](https://doi.org/10.1111/1533-6077.00011)

McLaughlin, B P (2019) ‘Could an Android Be Sentient?’ In A Puatz and D Stoljar, eds., *Blockheads!: Essays on Ned Block's Philosophy of Mind and Consciousness*, 335–374. The MIT Press. doi:[10.7551/mitpress/9196.003.0022.](https://doi.org/10.7551/mitpress/9196.003.0022%C2%A0)

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

Mckilliam, A, (2024) ‘Natural Kind Reasoning in Consciousness Science: An Alternative to Theory Testing’. *Noûs*. doi:[10.1111/nous.12526](https://doi.org/10.1111/nous.12526).

Melnyk, A (1994) ‘Inference to the Best Explanation and Other Minds’. *Australasian Journal of Philosophy* **72**: 482–491. doi:[10.1080/00048409412346281](https://doi.org/10.1080/00048409412346281).

Michel, M (2019) ‘Fish and Microchips: On Fish Pain and Multiple Realization’. *Philosophical Studies* **176**: 2411–28. doi:[10.1007/s11098-018-1133-4.](https://doi.org/10.1007/s11098-018-1133-4)

Millikan, R G (1999) ‘Historical Kinds and the “special sciences”’. *Philosophical Studies* **95**: 45–65.

Millikan, R G (2000) *On Clear and Confused Ideas: An Essay About Substance Concepts*. Cambridge University Press. doi:[10.1017/cbo9780511613296.](https://doi.org/10.1017/cbo9780511613296)

Palmer, T D and Ramsey, A K (2012) ‘The Function of Consciousness in Multisensory Integration’. *Cognition* **125:** 353–364. doi:[10.1016/j.cognition.2012.08.003.](https://doi.org/10.1016/j.cognition.2012.08.003)

Pargetter, R (1984) ‘The Scientific Inference to Other Minds’. *Australasian Journal of Philosophy* **62**: 158–63. doi:[10.1080/00048408412341341.](https://doi.org/10.1080/00048408412341341)

Piccinini, G and Craver, C (2011) ‘Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches’. *Synthese* **183**: 283–311. doi:[10.1007/s11229-011-9898-4.](https://doi.org/10.1007/s11229-011-9898-4)

Piccinini, G (2020) *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press. doi:[10.1093/oso/9780198866282.001.0001](https://doi.org/10.1093/oso/9780198866282.001.0001).

Phillips, I (2018) ‘The Methodological Puzzle of Phenomenal Consciousness’. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**: 20170347. doi:[10.1098/rstb.2017.0347](https://doi.org/10.1098/rstb.2017.0347).

Pollock, J L (1987) Defeasible Reasoning *Cognitive Science* **11**: 481–518. doi:[10.1207/s15516709cog1104\_4.](https://doi.org/10.1207/s15516709cog1104_4)

Putnam, H (1975) *Mind, Language and Reality: Philosophical Papers*. Cambridge University Press. doi:[10.1017/cbo9780511625251.](https://doi.org/10.1017/cbo9780511625251)

Russell, B (1948/2009) *Human Knowledge: Its Scope and Limits*. Routledge. doi:[10.4324/9780203875353.](https://doi.org/10.4324/9780203875353)

Salmon, W C (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton University Press. doi:[10.1515/9780691221489.](https://doi.org/10.1515/9780691221489)

Schneider, S (2019) *Artificial You*. Princeton University Press. doi:[10.1515/9780691197777.](https://doi.org/10.1515/9780691197777)

Schwitzgebel, E (2011) *Perplexities of Consciousness*. The MIT Press. doi:[10.7551/mitpress/8243.001.0001.](https://doi.org/10.7551/mitpress/8243.001.0001)

Schwitzgebel, E (2015) ‘If Materialism is True, the United States is Probably Conscious’. *Philosophical Studies* **172**: 1697–721. doi:[10.1007/s11098-014-0387-8.](https://doi.org/10.1007/s11098-014-0387-8)

Seth, A (2021) *Being You: A New Science of Consciousness*: Penguin.

Seth, A and Bayne, T (2022) ‘Theories of Consciousness’. *Nature Reviews Neuroscience* **23:** 439-52. doi:[10.1038/s41583-022-00587-4](https://doi.org/10.1038/s41583-022-00587-4).

Shea, N (2012) ‘Methodological Encounters with the Phenomenal Kind’. *Philosophy and Phenomenological Research* **84**: 307-344. doi:[10.1111/j.1933-1592.2010.00483.x](https://doi.org/10.1111/j.1933-1592.2010.00483.x).

Shea, N and Bayne, T (2010) ‘The Vegetative State and the Science of Consciousness’. *British Journal for the Philosophy of Science* **61**: 459–84. doi:[10.1093/bjps/axp046](https://doi.org/10.1093/bjps/axp046).

Shevlin, H (2021) ‘Non-Human Consciousness and the Specificity Problem: A Modest Theoretical Proposal’. *Mind & Language,* **36**: 297–314. doi:[10.1111/mila.12338](https://doi.org/10.1111/mila.12338).

Sober, E (2000) ‘Evolution and the Problem of Other Minds’.  *Journal of Philosophy* **97**: 365–86. doi:[10.2307/2678410](https://doi.org/10.2307/2678410).

Thomas, J (2001) ‘Mill’s Argument for Other Minds’. *British Journal for the History of Philosophy,* **9**: 507–23. doi:[10.1080/09608780110072498.](https://doi.org/10.1080/09608780110072498%C2%A0)

Tononi, G and Koch, C (2015) ‘Consciousness: Here, There and Everywhere?’. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**: 20140167. doi: [10.1098/rstb.2014.0167](https://doi.org/10.1098/rstb.2014.0167).

Travers, E, Frith, C D, and Shea, N (2018) ‘Learning Rapidly About the Relevance of Visual Cues Requires Conscious Awareness’. *Quarterly Journal of Experimental Psychology* **71**: 1698–713. doi:[10.31234/osf.io/7becr](https://doi.org/10.31234/osf.io/7becr).

Tye, M (2017) *Tense Bees and Shell-Shocked Crabs: Are Animals Conscious?* Oxford University Press. doi:[10.1093/acprof:oso/9780190278014.001.0001.](https://doi.org/10.1093/acprof%3Aoso/9780190278014.001.0001)

Weiskopf, D A (2017) ‘The Explanatory Autonomy of Cognitive Models’, in D Kaplan ed., *Explanation and Integration in Mind and Brain Science*: 44–69. Oxford University Press. doi:[10.1093/oso/9780199685509.003.0003](https://doi.org/10.1093/oso/9780199685509.003.0003).

Woodward, J (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. doi:[10.1093/0195155270.001.0001.](https://doi.org/10.1093/0195155270.001.0001)

1. In their report on consciousness in artificial intelligence, Butlin, Long, and colleagues address this by considering multiple competing theories at once, and assuming computational functionalism. Here my focus is, in part, to explore how to justify these assumptions. [↑](#footnote-ref-1)
2. I do not discuss Block’s presentation of the issue here, nor the responses to it by McLaughlin (2003) and Hohwy (2004), because his way of framing the issue builds in metaphysical and explanatory gap worries, that i) require a lot of set up—Brian McLaughlin required a full 17 pages just to introduce the problem—and ii) are, in my view, tangential to the epistemological question that is really at issue. [↑](#footnote-ref-2)
3. As an interesting historical aside, there is a case to be made that many of those traditionally associated with the argument from analogy really had something like analogical abduction, rather than analogical induction in mind. For example, Janice Thomas has convincingly demonstrated that a close reading of Mill—often cited as the father of the analogical account—reveals that he clearly took explanatory factors into consideration (2001). And some of Bertrand Russell’s writings—another advocate of the analogical account—look strikingly similar to Michael Tye’s appeal to inference to the best explanation (Tye 2017). For example, in one passage, Russell writes:

From subjective observation I know that A, which is a thought or feeling, causes B, which is a bodily act, e.g. a statement. I know also that, whenever B is an act of my own body, A is its cause. I now observe an act of the kind B in a body not my own, and I am having no thought or feeling of the kind A. But I still believe, on the basis of self-observation, that only A can cause B; I therefore infer that there was an A which caused B, though it was not an A that I could observe. On this ground I infer that other people’s bodies are associated with minds, which resemble mine in proportion as their bodily behaviour resembles my own. (Russell 1948/2009: 428) [↑](#footnote-ref-3)
4. One complication here is that masking may not actually switch consciousness on and off in the way we typically assume. Instead, it might result in degraded, but nonetheless conscious perception that goes unreported due to a conservative response bias (Phillips 2021; Brown 2024). Some worry that this might undermine the iterative natural kind strategy (Phillips 2018), but advocates of the view suggest that this too can be tested via the iterative natural kind method (Mckilliam 2024). [↑](#footnote-ref-4)
5. Shea (2012) suggests including neural markers as well in the cluster but doesn’t comment on how they relate to the behavioral markers. Similarly, Bayne et al (2023) take neural activity into consideration, but do not explore the evidential relationship between these and behavioral markers. [↑](#footnote-ref-5)
6. Thanks to an anonymous reviewer for suggesting this example. [↑](#footnote-ref-6)