The Problem of Other AI Minds

Boris Babic* and Jessica Wilson†

August 24, 2026‡

Abstract. Are we in position to warrantedly establish whether a given artificial intelligence (AI) system is conscious? In short, can we warrantedly establish whether there is AI Consciousness (AIC)? Here we aim to consider this question from the perspective of the traditional problem of other minds—the philosophical problematic which has been specifically directed at the question of whether and how we might come to be warranted in taking some other entity to have conscious mentality. More specifically: for each of the traditional strategies of response to this problem, we argue that even granting that the strategy works to establish that other humans and certain non-human animals (and perhaps even certain aliens) are conscious, the prospect of the strategy's working to establish that there is AIC is unlikely. Moreover, we argue, the failure of the traditional strategies to apply to the case of AI systems does not in itself support thinking that there is not AIC, in part for reasons which are central to generating the problem of other minds. The upshot is that there a distinctively difficult problem of other AI minds, whose lack of principled resolution offers a new line of support for what McClelland (2025) calls 'agnosticism about artificial consciousness'.

Keywords: Consciousness, Cognition, Artificial Intelligence, Identity, Mind, Problem of Other Minds, Hard Problem of Consciousness

Introduction

The question of whether sufficiently complex artificial intelligence (AI) systems are conscious, in the broad sense of being experientially aware, is of clear normative and political importance. For given that consciousness is commonly seen as prerequisite for (and, according to some, sufficient for) the possession of moral status, whether such attributions (or their denials) can be supported bears on whether AI systems should be seen as moral agents, patients, or persons. Relatedly, the status of AI systems as conscious (or not) bears on the rights, duties, privileges, and liabilities of AI systems, their manufacturers, and their users.

But are we in position to warrantedly establish that a given AI system is conscious (or not)? In short, can we warrantedly establish whether there is AI Consciousness (AIC)?

Here we aim to consider this question from the perspective of the traditional problem of other minds—the philosophical problematic which has been specifically directed at the

*Department of Philosophy and Institute of Data Science, University of Hong Kong; babic@hku.hk
†Department of Philosophy, University of Toronto; jessica.m.wilson@utoronto.ca
‡Substantially revised as of August 24, 2026.

question of whether and how we might be warranted in taking some other entity or entities to have conscious mentality.

We proceed as follows. We first lay out the version of the problem of other minds, and the conception of AI systems, operative in our discussion (§1). We then consider each of the main strategies of response to the problem of other minds, including argument by analogy (§2), variations on the Turing test (§3), appeal to a theory or theories of consciousness (§4), and inference to the best explanation (§5). For each strategy, we argue that it fails to provide a basis for establishing that there is AIC; throughout we aim to connect considerations made salient by attention to the problem of other minds to certain recent lines of thought pertaining to whether there is AIC (as in, e.g., Butlin et al. 2023, Bayne et al. 2024, McClelland 2025, Udell and Schwitzgebel 2021a). We then argue that the failure(s) of these strategies to establish that there is AIC does not suffice to establish that there is *not* AIC, in part for reasons which are central to generating the problem of other minds (§6). We conclude that—at least going by the usual responses to the problem of other minds—we are not presently in position to warrantedly establish whether there is AIC, and modulo some unforeseen new strategy, are unlikely to ever be able to do so. The upshot is that there a distinctively difficult problem of other AI minds, whose lack of principled resolution offers a new line of support for what McClelland (2025) calls 'agnosticism about artificial consciousness'.[1] (§7).

1 Preliminaries

1.1 The operative problem of other minds

Schematically, the problem of other minds may be expressed as follows:

> Given that each of us has introspective access only to our own consciousness, how can any of us warrantedly establish that another entity is conscious?[2]

As Avramides (2020) notes, the problem admits different readings. We will distinguish these as 'skeptical', 'justificatory', and 'categorical'. The skeptical reading focuses on the threat of solipsism in light of certain broadly Cartesian scenarios:

> **Skeptical.** Given that I have introspective access only to my own consciousness, can I warrantedly establish that another entity is conscious in the face of certain skeptical scenarios in which everything appears to me as it presently does, but no other conscious entities exist?

[1] McClelland's route to agnosticism abut AIC highlights the hard problem of consciousness, whereas ours highlights the problem of other minds; his discussion is consonant with ours, however, and we will draw upon it at various points in what follows. As McClelland observes, existing literature is mainly divided into camps of AIC-advocates and AIC-deniers; he moreover observes that one can distinguish 'soft' and 'hard' versions of agnosticism, reflecting one's stance about whether we might someday be in position to warrantedly infer to AIC (or its absence). We, like McClelland, are 'hard-ish' agnostics, who aim to render salient the existence of serious obstacles to knowledge about AIC, where it is unclear how these obstacles might be overcome.

[2] The problem of other minds is typically specifically directed at whether we are in position to ascertain whether some other entity (or category of entities) has conscious or experientially present mental states. Of course, there are a wide variety of such states; for present purposes it suffices to present the problem and its strategic responses in terms of consciousness, or conscious awareness.

The justificatory reading sets aside skeptical considerations, rather assuming that individual humans are warranted in taking certain other entities to be conscious, and focuses instead on the question of *which* justificatory procedures are involved in such warrant:

> **Justificatory.** Given that each of us has introspective access only to our own consciousness, and is also warranted in taking certain other entities to be conscious, by which justificatory procedure(s) do we come to have such warrant?

The categorical reading, like the justificatory reading, puts aside skeptical considerations, but removes the assumption that we are antecedently warranted in taking the target entities to be conscious:

> **Categorical.** Given that each of us has introspective access only to our own consciousness, are there justificatory procedures by which we can warrantedly establish that certain other entities are conscious, and if so, what are these?

Our focus here is the categorical version of the problem of other minds, and henceforth by 'problem of other minds' we intend to refer to this version.

While the relevance of the problem of other minds to the prospects for determining that, or whether, there is AIC might seem obvious, that connection remains surprisingly underexplored.[3] But as we hope to show, exploring that connection can be very fruitful, for raising to salience the full range of historically important strategies of response and for appreciating the distinctive challenges faced by attempts to determine whether there is AIC, despite the clear urgency of doing so.

1.2 The operative conception of AI systems

We next sketch the core elements of AI systems as we are understanding such systems here. Very generally, an AI system is a statistical learning model which uses a (typically convex) optimization procedure in order to estimate a (typically high dimensional) function on the basis of the data it is trained on (Hastie et al. 2009). Such a model is ordinarily updated (or fine-tuned) on new data over time. And such a model can be used to classify new samples, make predictions, or generate new content through sampling from the estimated function (Rubinstein and Hastie 1997, Ng and Jordan 2001). A case-in-point would be a generative supervised learning model such as a feed-forward neural network which estimates the joint probability of the label and the data. Indeed, under certain weak assumptions, such a model can be seen as a universal approximator of any function (Hornik et al. 1989).

[3]There are certain exceptions. Harnad (1991) states: "Any attempt to explain the mind by building machines with minds must confront the other-minds problem" (44); however, he focuses only on whether any variation on the Turing test can serve as a response to the problem. Schneider (2019) mentions the problem of other minds, but only by way of suggesting that 'the problem of AI Consciousness' is not best seen as a variation on this problem, since "the popular [analogical] solution to the problem of other minds is ineffective in the context of the Problem of AI Consciousness". Of course, many of those who are concerned with the question of whether and how we might test for AIC, including Butlin et al. (2023) and Bayne et al. (2024), while not specifically adverting to the problem of other minds, can be seen as indirectly concerned with this problem. We will engage with various aspects of the aforementioned literature below.

This is a broad definition. One could restrict the operative understanding of AI systems to those models which are trained on sufficiently large data, or to those which generate new content; however, these conditions are fairly arbitrary. For example, it is not always the case that a model's capability increases linearly with increases in the size of either the dimensionality or the sample: that claim requires certain further assumptions. And the distinction between classifying samples and generating new content is ultimately a distinction without a difference; for once an estimate of the joint feature-label probability is obtained, one can use it either to bin samples into groups (classification) or to sample from the estimated distribution (generation) (Ng and Jordan 2001). It is also worth observing that while the notions of 'pre-training' and 'fine-tuning' are often assumed to correspond to a technical property of AI models, in fact this terminology more accurately reflects that as computer scientists started training increasingly larger and larger models, using energy intensive GPUs, it became computationally too intensive to train a model before each application. Further limited refinements were thus described as 'fine-tuning'. Hence, we stick with (though occasionally expand on) the broad definitional sketch offered here.

We provide this sketch so as to have some concrete framing of the kinds of systems we have in mind, but nothing in our argument hangs on a particular definition of AI systems. One caveat: our discussion excludes radically different 'intelligent' systems whose performance is completely unlike the systems we have today. As above, our focus is restricted to the kinds of statistical learning models that constitute today's AI systems, and their anticipated near-to-medium-term progenitors. Today's models (for example, state of the art large language models built on deep neural networks, trained via backpropagation and gradient descent, and implemented using the transformer architecture) have evolved from the single layer perceptron of the 1960s. Such models have grown increasingly large, but there are resource constraints to such scaling—transistors, for example, cannot keep shrinking forever. What future AI systems might transpire as a result of such limitations is anyone's guess. Correspondingly, we do not aim to argue for agnosticism about the consciousness of AI systems radically different from those existing today or in the foreseeable future. Even so, our results are significant: at least so far as present-day AI systems and their anticipated descendants are concerned, and by lights of the traditional philosophical problematic which has been specifically directed at the question of whether some other creature has conscious mentality, we are not in position to warrantedly establish that, or whether, any such systems are conscious. Or so we will now argue.

2 Argument by Analogy

2.1 The general strategy

One historically common response to the problem of other minds involves an argument by analogy, whereby certain similarities between an entity known to be conscious (such as one's self, via introspection) and the target (such as another human) are taken to support the target entity's also being conscious. For example, Mill argues that certain bodily and behavioural similarities between him and other humans provide a justificatory basis for his taking them to be conscious:

> **Mill's analogical argument.** First, they have bodies like me, which I know in my own case, to be the antecedent condition of feelings; and because, secondly, they exhibit the acts, and outward signs, which in my own case I know by experience to be caused by feelings. (Mill 1865, 243)

And as Russell (1948) put it, in addressing the problem of other minds, "it is clear that we must appeal to something that may be vaguely called 'analogy". The behaviour of other people is in many ways analogous to our own, and we suppose that it must have analogous causes. What people say is what we should say if we had certain thoughts, and so we infer that they probably have these thoughts. [...] As it is clear to me that the causal laws governing my behaviour have to do with "thoughts", it is natural to infer that the same is true of the analogous behaviour of my friends" (501-2). See also Hampshire 1952.

Contemporary advances in neuroscience and evolutionary biology have identified further biological (including evolutionary), physical, and functional similarities associated with human consciousness. For example, humans are physically similar in displaying characteristic patterns in electroencephalograms when awake (not present when asleep or under anesthesia); humans are biologically and evolutionarily similar in that "the basic neurophysiologic mechanisms supporting consciousness in humans are found at the earliest points of vertebrate brain evolution" (Mashour and Alkire 2014); and humans share similar internal functional architectures (e.g., the association of cortical activity with recurrent local feedback). Such empirically established similarities provide additional support to the general analogical strategy.

To be sure, there remains the further question of how, exactly, the features at issue, internal and external, are (constitutively, causally, correlationally?) related to consciousness. Luckily, applying the argument by analogy to the target case of humans doesn't require commitment on these further scores. Rather, it suffices to show that the features are clearly relevant, somehow or other, to human consciousness, and that they are in place in cases of other (normally functioning) humans. So far, so good, then—at least some have thought—for purposes of an adequate response to the problem of other human minds.[4]

Such analogical considerations plausibly extend to at least some non-human animals. Many non-human animals behave in ways that "I know by experience" to be caused by conscious states, and they frequently have bodies which, though not exactly "like me", are close variations on the theme. Moreover, they have brains or associated neurobiological substrates which are similarly internally functionally organized. As Mashour and Alkire (2014) observe:

> [F]unctionally similar cognitive systems may arise from neurobiologically distinct structures. For example, the mammalian cortex and avian pallium are histologically distinct ... but may subserve similar network functions that can be quantitatively assessed and compared with human findings. (10361)

[4]The analogical strategy, like the other strategies we will consider, is the target of various objections and associated responses (see Avramides 2020 for discussion). Here and elsewhere considerations of space prevent us from entering into these further details; in any case our aim is to show that even granting the viability and successful application of the strategies to certain cases, they cannot be applied to the case of AIC.

Related considerations entered into the creation and signing, by a prominent group of cognitive neuroscientists and related scholars, of the 'Cambridge Declaration on Consciousness':

> We declare the following: "The absence of a neocortex does not appear to preclude an organism from experiencing affective states. Convergent evidence indicates that non-human animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviors. Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Non-human animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates." (Low et al. 2012)

The argument in the Declaration is a clear variation on an analogical response to the problem of other minds—one on which certain physical, behavioural, biological, and functional similarities are taken to warrant the positive attribution of consciousness, not just to other humans, but also to certain non-human animals. Plausibly, then, it is commonly granted that the analogical strategy can work for other humans and some non-human animals. But might this strategy work for AI systems?

We next argue that it will not. To prefigure: our argumentative strategy differently targets different of the similarities at issue. For the case of physical and biological/evolutionary similarities, we argue that such similarities are not in place. For the case of behavioural similarities, we argue that granting that certain such similarities are in place, they do not support the analogical inference. And for the case of functional similarities, we argue, first, that (since neuroscientists suppose that the functions relevant to consciousness are instantiated in neurobiological structures), the relevant similarities are not in in place; second, that (putting aside the requirement of neurobiological substrate) the opacity of AI systems currently and (as attention to the usual strategies for 'explaining' or 'interpreting' AI processing indicates) for the foreseeable future renders it unclear how we might establish that AI system processing is functionally similar to that relevant to consciousness in humans and non-human animals; and third, even if these barriers can somehow be overcome, similarity in just this respect is too thin a reed upon which to rest an argument by analogy.[5]

2.2 Physical similarity

We start by observing that one would be hard-pressed to find a compelling basis for taking AI systems to have 'bodies like us' or like any non-human animals. The bodies of conscious human and non-human animals are concrete, spatiotemporally localized and bounded particulars, which admit of comparatively clear individuation and persistence conditions, and which have parts that are empirically associated with consciousness. But it is unclear whether AI systems have bodies at all. Perhaps most plausibly, AI systems are ontologically on a par with computer programs, musical scores, or other comparatively abstract processes,

[5]Down the line we will consider whether some or other version of computational functionalism, coupled with some future identification of similar functional architecture in AI systems, might support a non-analogical inference to AIC.

which in admitting of multiple realizations at a time, are more like repeatable types or universals than token concrete particulars. Hence it is, for example, that AI systems are taken to be capable of being synchronously run on multiple distinct hardware systems.

Nor does it help to focus on a specific instantiation of an AI system. If there is anything answering to 'body' in the vicinity of such an instantiation, it lies in the physical hardware needed for the system to run. Unlike the organic, spatiotemporally localized and bounded bodies of human and non-human animals, however, this hardware is widely distributed and massively parallelized, it fails to admit of clear individuation and persistence conditions, and we have no antecedent reason to associate consciousness with any of its parts. The latter dissimilarity is of particular importance, since it is because (per Mill) parts of my body, and (per the Cambridge Declaration) more specifically neurobiological parts, are empirically associated with consciousness, that similarity in bodily/neurobiological respects does real work in the associated argument by analogy.[6]

Of course, an AI system might be embedded in a robotic body, and correspondingly be said to 'have' a body that is concrete and spatiotemporally bounded. But this isn't enough to support there being a physical analogy, given that no parts of an inorganic robot body are empirically associated with consciousness. An analogical argument based on physical similarity might be resurrected in the case of an organic 'robot' which physically duplicates a conscious human or non-human body (including relevant neurobiological parts thereof); but in that case the analogical support for consciousness would attach to an organic duplicate of a conscious animal, not to any AI system that might somehow be associated with that duplicate.

2.3 Behavioural similarity

What about behaviour? AI systems do exhibit what might be described as linguistic behaviour: they can answer questions and more generally engage in something resembling human conversation. However, as we'll now argue, since the source of this behaviour is (a) clearly very different from the source of human linguistic behaviour, and (b) such that we have no antecedent reason to think involves consciousness, the similarity at issue provides no analogical support for the posit of AIC.

To appreciate this concern, it is worth expanding on the role that behavioural similarity plays in analogical arguments for the consciousness of other humans. Crucially, Mill's argument rests not just on the behaviour's being similar, but on this similar behaviour's being potentially generated in a way "which in my own case I know by experience to be caused by feelings". Similarly, Russell is explicit that it is not just the similar behaviour, but that (we suppose that) the behaviour "must have analogous causes", which is driving the argumentative strategy. If Mill or Russell justifiably believed that the similar behaviour of the target did not "have analogous causes", but was rather generated by a hidden puppeteer or some other source very different from the source in their own cases, not requiring or implicating

[6] At this point it is worth noting that talk of AI systems as 'neural' nets or networks is misleading, for there is no physical similarity, in either physical constitution or physical processing, between human/non-human animal brains and so-called 'neural' networks. We return to the question of AI processing and associated functional architecture below.

the consciousness of the target, then the behavioural similarity would be irrelevant to the analogical argument.

More generally, any version of an analogical argument appealing to behavioral similarities requires not just that the target have features similar to those had by conscious entities, but that it be plausible and/or probable that the similar behavioural features have the same cause or source—namely, the target's consciousness. The moral is that at least some first-order similarities of the sort entering into an analogical solution to the problem of other minds require, in addition, that it be plausible or probable that the source of the first-order similarity in the target is the same as the source of the first-order similarity in the base case—that is, that it be plausible, or at least a live possibility, *that the first-order similarities are generated by relevant second-order similarities*.

For AI systems, however, the second-order similarities is not in place. For the first-order linguistic behaviours of AI systems are a consequence of these probabilistic prediction models' being trained on massive amounts of human-generated text, in a way very different from the ways in which humans engage or learn to engage in linguistic behaviour, and—most importantly—which is a matter of computational processes which we have no antecedent reason to think involve consciousness. So it is neither plausible nor probable that the linguistic behaviours manifested by AI systems are generated in the same ways as these behaviours in humans. As Russell would put it: it is neither plausible nor probable that the cause of similar linguistic (or other) behaviour in AI systems involves the internal mental states that we know in our own case to be the causes of such behaviours. Hence it is that, notwithstanding the impressive linguistic similarities associated with AI systems, the absence of second-order similarities for these first-order similarities blocks taking the latter to be properly input into an associated argument by analogy for AIC.[7]

The concern highlighted here—call it the 'second-order similarity concern'—is related to, but different from, concerns that the linguistic and other behaviours of AI systems cannot be taken as indicators of consciousness, since these behaviours are designed or 'built in' to AI systems, or (relatedly) that these behaviours 'mimic' certain human behaviours. Variations on these other concerns, for linguistic or other forms of behaviour associated with AI systems, are common. For example, Schneider (2019) maintains that 'if [AI systems] do behave like humans, it may be because they are programmed to behave as if they feel, so we can't infer from their behavior that they are conscious" (129), and Butlin et al. (2023) maintain (following Andrews and Birch 2023; see also Schwitzgebel and Pober 2026) that "AI systems may be trained to mimic human behaviour while working in very different ways, thus "gaming" behavioural tests" (18).

To start, the second-order similarity concern is not (just) that AI systems have been designed or trained to engage in certain linguistic and other behaviours; for if the imple-

[7]For a concrete example, consider AI 'Chain of Thought' models. Early such models tended to be too verbose, such that people would distrust them for, so to speak, yapping too much. Correcting for this tendency required explicit tweaking; hence during pre-release testing of OpenAI's o3 model it was found that the system uses a "yap score" to prevent it from being overly verbose, so as to convey the impression of trustworthiness (Chowdhury et al. 2025). Granting that OpenAI's model engages in linguistic behaviour of the sort that in the case of humans inspires the attribution of trustworthiness, awareness of how the appearance of linguistic trustworthiness was achieved undercuts such behaviour's providing analogical support for the model's being conscious.

mentation of the design (or training) proceeded via effective duplication of the physical and biological basis of these behaviours in humans (or other entities justifiably believed to be conscious), then the similar behaviour would be relevantly input into an analogical argument for the entity's consciousness. Nor is the second-order similarity concern (just) that AI systems are in some sense 'mimicking' human linguistic behaviours; for as Schwitzgebel and Pober (2026) point out, current (e.g., biological) accounts of mimicry allow that a given mimic (say, a butterfly mimicking the wing pattern and colouring of a poisonous Monarch) might in fact be poisonous. Rather, the concern is that we have reason to believe, for the case of AI systems (and also, according to Schwitzegebel and Pober, the case of associated inorganic robots), that the cause of the first-order behavioural similarities is different from any which would serve to support the relevance, for purposes of a positive attribution of consciousness, of the first-order similarities.

2.4 Biological (including evolutionary) similarity

Conceptions of biological similarity typically fall into three categories: first, homologous similarity, whereby entities possess similar (e.g., physical or molecular) traits inherited from a common evolutionary ancestor; second, analogous similarity, whereby entities possess similar traits which, while not inherited from a common ancestor, naturally evolved as a result of independent common responses to evolutionary pressures; third, genetic similarity, measuring the extent to which the genomes (DNA sequences) of entities match or overlap. Each of these respects of biological similarity presupposes that the entities at issue evolved via natural selection; since AI systems did not so evolve, none of these respects of similarity are in place.

These failures of biological similarity also undercut the import of certain other first-order similarities for analogical purposes. Compare: both humans and non-human animals naturally evolved in circumstances where consciousness of internal states and external environments was plausibly selected for (Mashour and Alkire 2014); in turn, this evolutionary commonality serves as another consciousness-relevant second-order similarity—a 'meta-cause', if you will—which supports taking various first-order (e.g., behavioural) similarities to be relevant to an argument by analogy to some other entity's being conscious. Conversely, that AI systems did not naturally evolve further supports the results of the previous subsection, according to which the similar linguistic and other behaviours of AI systems, in being generated by processes very different from those generating the behaviours in humans or other entities warrantedly taken to be conscious, are irrelevant to an analogical argument to AIC.

Note that these results—that AI systems are biologically (to wit: evolutionarily) very different from organisms (humans, certain animals) warrantedly believed to be conscious, and that this biological dissimilarity undercuts the analogical import of certain first-order (e.g., linguistic behavioural) similarities which might be thought to be in place—are not impacted by results suggesting that "in computer simulations of evolution, AI evolves in a manner comparable to that of organisms" (McClelland 2025, 2). The import of natural evolution is that the consciousness of complex middle-sized organisms is plausibly implicated in adaptive behaviours resulting in survival and fitness, since such adaptive behaviours clearly aren't explained by anything answering to in-built corrective programming. Since the 'evolution' of AI systems does proceed as a matter of in-built corrective programming, that AI systems

correspondingly iteratively or otherwise 'evolve' to produce more 'successful' behaviour or outcomes is no more analogically useful than the fact that AI systems engage in linguistic behaviour similar to that of humans.

2.5 Functional similarity

So far we have found no substantive or relevant similarities between AI systems and creatures warrantedly taken to be conscious which might serve as the basis for an argument by analogy. One potential similarity remains—namely, similarity in internal functional architecture. Might similarity in this respect provide the basis for an analogical argument for AIC?

Here we offer three arguments for a negative conclusion.

The first is that while (as above) neuroscientists allow that "functionally similar cognitive systems may arise from neurobiologically distinct structures", they nonetheless suppose that the diverse structures implementing the functions are neurobiological. Relatedly, the functional architecture at issue in the Cambridge Declaration is specifically supposed to involve "neuroanatomical, neurochemical, and neurophysiological" substrates. AI systems do not have any such substrates. Hence, on the usual neuroscientific understandings of the internal functional architecture(s) associated with consciousness, AI systems are not similar to conscious humans or certain non-human animals in this respect.

The second is that, even putting aside the neuroscientific supposition that the functions at issue are neurobiologically implemented, there are severe, commonly acknowledged difficulties in identifying in any detail how AI systems 'work', hence in identifying what abstract functional architecture is implemented by a given AI system. (This problem will also be relevant in §4.)

An initial indication of the difficulty here stems from the complexity and opacity of AI systems. Even for a simple feed-forward neural network with only a few hidden layers, it is difficult to make credible and reliable statements about how each associated feature contributes to the prediction. This is because each input variable is passed through multiple combinations of weights, so that the final assessment about what—causally, within the model—produces the output is difficult to evaluate (see Babic et al. 2021, Babic and Cohen 2023). Moreover, any such assessment is non-robust, as the weights change continuously through backpropagation. The opacity here attaches to even a tiny model—say, one with a few dozen parameters, where the estimated function may even be approximately linear. For modern large models involving billions (Llama 2 and Llama 3) or even trillions (GPT 4) of parameters, for which the estimated functions are massively non-linear, highly non-compressible, and largely recursive (Rudin 2019, 207), there is little hope of meaningfully understanding how they reach their predictions. In that case, however, there is no clear path to establishing that AI systems have an abstract functional architecture similar to those of creatures warrantedly believed to be conscious.

One might wonder whether this pessimistic line of thought fails to take into account progress made on explainable AI ('XAI', for short) (see, e.g., Ribeiro et al. 2016, Lundberg and Lee 2017, Lipton 2018, Adadi and Berrada 2018, Alain and Bengio 2016, Arrieta et al. 2020). Why, one might wonder, couldn't XAI provide the basis for establishing that AI system processing involves a certain internal functional architecture—perhaps even one abstractly similar to that in conscious humans and non-human animals (see, e.g., Goldstein

and Levinstein 2024)? While we cannot aim to comprehensively anticipate every direction or result that XAI investigations into XAI might produce, in what follows we will argue that two of the main families of approaches advanced by proponents of XAI—namely, the counterfactual and surrogate approaches—fail to provide a basis for identifying the internal functional architectures of AI systems.[8] We will also argue that the most recent progenitors of surrogate approaches, along lines of what is called 'mechanistic interpretability', face similar obstacles.

The first, broadly counterfactual approach to XAI involves an assessment of an input perturbation's effect on the output (Wachter et al., 2018). Wachter et. al attempt to explain an AI system's prediction by considering the smallest change in input that would have led to a change in the output. For example, suppose that a given AI system uses a student's SAT score and GPA to determine whether they are admitted to a certain college. Suppose further that a student, Henry, is rejected. Henry now wants an explanation of why he was rejected. Henry's GPA was 3.7 and his SAT score was 1200. A counterfactual explanation looks for the smallest change in inputs that would have led to Henry's being admitted. For example, it may be that had his SAT score been 1400, he would have been admitted.

Such a counterfactual strategy does not, however, help us understand the functional internal architecture of the system. It simply shows us one, non-unique way in which the result could have been different. The route is non-unique because there are many equidistant perturbations given a measure of distance, and moreover there are many possible measures of distance. In our example, there are two input variables and associated parameters (SAT and GPA). It may be that Henry would have been admitted with an SAT score of 1400, leaving GPA unchanged, or a GPA of 3.9, leaving SAT unchanged, or a mixture of increases in both (for example, an SAT score of 1300 and a GPA of 3.8 , or an SAT score of 1310 and a GPA of 3.79, etc.).

Correspondingly, the counterfactual strategy simply provides a post hoc rationalization of how Henry could have been on the other side of the classification boundary, without helping us to understand the internal functional architecture that led to the classification boundary in the first place. In effect, these strategies provide a non-unique and non-stable approximation of feature importance. And we can easily generate infinitely many such approximations: for example, in Henry's case, we would choose any $\phi \in [0, 1]$ such that $\phi\text{SAT} + (1 - \phi)\text{GPA}$ is above the admission threshold. Further, as new data comes in, parameter values change, and counterfactual strategies must be re-applied to get new estimates of feature importance. An additional problem concerns the assumption that parameters are independent of each other, required for counterfactual strategies to make sense; for it is known that parameters in large models are massively colinear. Other 'family resemblance' variations on the theme of a counterfactual approach share the same limitations.

That counterfactual correlations underdetermine actual causal or functional architecture is unsurprising. It's not just that these approaches aren't explicit about the underlying architecture; it's that, as is familiar in discussions of the underdetermination of theory by correlational evidence in philosophy of science,[9], any pattern of input-output correlations

[8] We highlight difficulties with these central approaches to XAI here, since in our view, these difficulties are underappreciated. As noted above, however, the argument of this section does not hinge on the supposition that these difficulties can never be overcome.

[9] See Stanford 2023.

is going to be compatible with indefinitely many mutually incompatible functional architectures.[10] Some working on AI 'explainability' acknowledge the underdetermination problem, but incorrectly suggest that it applies only to actual input-output correlations. Hence Grosse et al. (2023) observe that "[taking] a top-down approach [starts] with the model's input-output relationships and zooming in. [...] Unfortunately, it is difficult to draw firm conclusions simply from looking at model samples and probabilities because any particular output is consistent with many different pathways" (4). Gross *et al.* go on to suggest that this problem can be largely resolved by working with 'influence functions', where one considers counterfactuals asking "how would the model's behaviors change if a given sequence were added to the training set?" (4)—effectively, as per the counterfactual approach above; but even if attention to counterfactual input-output correlations operates to rule out some hypotheses as regards what internal processes generated the original correlations, an infinite number of other hypotheses will remain.

The second, 'surrogate' approach to XAI involves taking the original prediction from a black box AI model and fitting a simple and intuitive model to the original inputs and that prediction.[11] Consider again our previous example, whereby an AI system uses a student's SAT score and GPA to determine whether they are admitted to college. What explains why Henry was rejected? On the surrogate approach, one asks: had a simple model been used to produce that rejection on the basis of his SAT and GPA, which parameter values associated with SAT and GPA could have produced Henry's rejection in the simple model?

Such 'surrogate' approaches aim to replicate the results of a given AI model. A prominent case-in-point of a surrogate approach involves 'probing' (see Goldstein and Levinstein 2024), where the idea is to fit a (typically linear) model between the output and the individual layers of the neural network. LIME and SHAP are also instances of surrogate approaches (Alain and Bengio 2016, Ribeiro et al. 2016, Lundberg and Lee 2017). One can distinguish between local and global surrogate models, where local models replicate the result for individual samples (e.g., Henry), and global models replicate the result for all samples.

What makes a model simple or intuitive enough to allow interpretability is a matter of dispute, but it is generally agreed that linear regressions, decision trees and other additive models may be considered white boxes.[12] In any case, the main point for our purpose here is that post-hoc strategies are not useful for purposes of identifying the internal functional architecture of a given AI system in a way that would enable us to make analogical headway on the question of AIC. First consider global surrogates. If we think about a learning task as a function estimation problem, where the underlying function is non-linear in a high dimensional space, then global surrogate models are clearly either inaccurate representations or else are themselves opaque. There is, necessarily, no middle ground.[13] Next consider local surrogates. Local surrogate models need not be opaque, but they offer "explanations that

[10] Another relevant literature pertains to attempts to extract causal relationships from correlations in interventionist causal models; as Cartwright (1979) convincingly argues, "no causation in, no causation out".

[11] See Rudin (2019) for discussion.

[12] See Babic and Cohen 2023. See also Martín et al. 2023, 'Interpretability Techniques: State of the Art', which offers linear regression and decision trees as examples of white box models.

[13] Note that no global model can be a perfectly accurate representation of the underlying blackbox model, for if it were, it would just be the same model.

are not faithful to what the original model computes" (Rudin 2019, 207). In other words: local surrogate models are just different models altogether.

Consider an everyday example. Alice, a student, is late to class. We want to know why Alice is late to class. We ask you: Why is Alice late? And you reply: it is consistent with the data (the data being that Alice is not in class) that Alice slept in. You may also reply: it is consistent with the data that Alice is stuck in traffic. By providing a post-hoc rationalization, you are substituting your own world model of how things could be (even though you have no affirmative reason to believe that is how things are) for the actual world of how things are (i.e., the true cause of Alice's lateness). Thus, when we are given any of these infinitely many possible replies—these post-hoc explanations which are consistent with the data—we are not any closer to understanding the actual causal or functional mechanisms of Alice's lateness. The analogous problem plagues the explanatory power of local surrogate models.

A related criticism is advanced in philosophy of science against the hypothetico-deductive model of refuting falsifiable hypotheses—namely, that by refuting one of infinitely many hypotheses, I am not thereby any closer to knowing the truth (Hempel 1945a, Hempel 1945b). From this perspective, post-hoc explainability involves committing the same fallacy as one who takes observations of non-black non-raven things to confirm (the thesis consistent with these observations) that all ravens are black.

Another 'white box' (so-called 'inherent interpretability') strategy is discussed in contexts where AI interpretability/explainability is at issue: namely, where one takes the original input data and creates an interpretable model from the outset (see Rudin 2019). While this is a strategy for reducing opacity in algorithmic decision making, by hypothesis it does nothing for understanding the internal functional architecture of complex models, since the explicit goal of that strategy is to simply stop using such models altogether.

One might wonder whether more recent techniques—the most influential of which involve what is known as 'mechanistic interpretability'—can be more useful for understanding consciousness. Unfortunately, they cannot be. In short, contemporary mechanistic interpretability is an extension to large language models of what we earlier described as local surrogacy approaches. One of the foundational projects on mechanistic interpretability is advanced in Olah et al. 2020. The authors, who were then all working at OpenAI, suggest that linear combinations of neurons can correspond to human-interpretable concepts, and that neurons within a network wire together into sparse, localized subgraphs, or 'circuits', as they dub them. But the same post-hoc limitation remains: In order to study such circuits, they rely on feature visualization (synthesizing idealized inputs that maximally activate a neuron), which creates an artificial proxy image *after* the model is already trained, rather than demonstrating how the model in fact processes noisy real world data.

Similarly for mechanistic interpretability techniques which use sparse autoencoders (SAEs) for feature reconstruction, such as that advanced in Cunningham et al. 2023, which explicitly construct an after-the-fact approximation of the original black-box model by training a second, external model to approximate the internal states of the primary model. The role of the SAE is to reconstruct the internal activations of a language model post-training. Earlier XAI—including leading algorithms like SHAP (Ribeiro et al. 2016) or LIME (Lundberg and Lee 2017)—constructed local linear surrogates around input-output pairs. Mechanistic interpretability techniques built around SAEs purport to open the black box by looking at weights and activations, but they still fundamentally rely on linear surrogates as a proxy for

understanding internal activations. Hence the core logic of using post-hoc secondary models as an approximation of the actual black box model remains ever present—we substitute a human legible proxy for the actual underlying high-dimensional computation. This has all the same limitations as our toy example, described earlier. If we ask, "why is Alice late to class?" we are looking for something causal—e.g., the motivating reason behind Alice's lateness or the actual cause of her lateness. But when we substitute this with a proxy—e.g., "it is consistent with Alice being late that she slept in"—we have not improved our understanding of what drove Alice to be late. Correspondingly, taking the output of the SAE to explain the primary network introduces the same post-hoc approximation problem initially described in Babic and Cohen 2023—to wit, applying an interpretive frame post-training (as mechanistic interpretability strategies do) in order to force clean, human legible concepts out of messy high-dimensional computations is a case-in-point of an explainability "bait-and-switch."

At present, then, there is no clear route to the identification of what functional architecture is operative in AI systems, much less to identifying such architecture as relevantly similar to those operative in humans or other entities warrantedly taken to be conscious.

A third difficulty for a functional-architecture-based analogical argument remains even if new and different approaches to XAI somehow succeed in identifying AI system architecture as appropriately functionally similar to ours (and again, putting aside the usual neuroscientific supposition of a neurobiological substrate)—namely, that analogical arguments offered in response to the problem of other minds highlight and rely on the presence of multiple, mutually supporting similarities, including the sort of meta-level similarities which ensure the relevance for analogical purposes of the first-order similarities. Hence even if (we eventually find out that) AI systems instantiate abstract functional architectures similar to that instantiated in humans or other creatures known to be conscious, such a 'one-off' similarity is insufficient for purposes of implementing an argument by analogy to AIC. To be sure, various specific theories of consciousness take the implementation of some or other functional architecture to suffice for or be constitutive of consciousness, and we will address the import of such theories in §4. Our point here, however, is just that the presence of a functional similarity in the absence of any other (first-order or meta-level) similarities is too thin a reed to itself support an argument by analogy to AIC.

We conclude, more generally, that the analogical response to the problem of other minds cannot be implemented for AI systems: there is no analogical argument to there being AIC.

3 The Turing Test and its Variations

We next turn to the Turing test and certain related strategies of response to the problem of other minds. Turing's original suggestion is that if (as in the 'imitation game') a human (the 'interrogator') cannot distinguish between the linguistic responses of another human and a 'machine', then this establishes the machine as 'thinking':

> I believe that in about fifty years' time it will be possible to programme computers [...] to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning. (Turing 1950).

If 'thinking' in the relevant sense requires consciousness, then the test might be seen as a restricted argument by analogy—one according to which linguistic behavioral similarity alone suffices for a positive attribution of consciousness. Even putting aside the question of sufficiency, however, the concerns which undercut the import of linguistic similarity for analogical purposes also undercut the usefulness of the Turing Test for purposes of establishing AIC. Here again, a common complaint is that the fact that AI systems (or associated robots) are designed or programmed, and moreover designed or programmed to mimic human linguistic behaviour. As Schneider (2019) puts it, focusing on seeming utterances about consciousness:

> Even today's robots can be programmed to make convincing utterances about consciousness, and a highly intelligent machine could perhaps even use information about neurophysiology to infer the presence of consciousness in biological creatures. [...] If sophisticated nonconscious AIs aim to mislead us into believing that they are conscious, their knowledge of human consciousness and neurophysiology could help them do so. (46)

As we see it, scenarios such as Searle's 'Chinese room' (Searle 1980), which suggest that "a mindless symbol-manipulator could pass the [Turing Test] undetected" (Harnad 1991), can be seen as identifying considerations that undercut the supposition that the similar linguistic behaviour has a similar consciousness-involving source or cause (per the 'second-order similarity concern', above). Hence, contrary to Turing's original application, the test is not useful for purposes of establishing the consciousness (or experientially informed 'thinking') of any system or machine which is such that the source or cause of its linguistic behaviour is sufficiently different from that which generates linguistic behaviour in humans or others known to be conscious.

Though, more generally, it is commonly supposed that the original Turing test is too weak to do the job of establishing that some other entity is conscious (see Oppy and Dowe 2021), it is worth considering whether two variations on the theme might do better by way of determining that AI systems are conscious.

The first variation is the 'total Turing Test' proposed by Harnad (1989, 1991), which involves the inability to distinguish not just linguistic but any relevant inputs:

> The Total Turing Test (TTT) calls instead for all of our linguistic and robotic capacities; immune to Searle's argument, it suggests how to ground a symbol manipulating system in the capacity to pick out the objects its symbols refer to. (Harnad 1991, 43)

But granting that the TTT triangulates on action in such a way as to sidestep Searle's concern, the problem remains that such further behavioural similarities will be generated in ways that on the face of it do not, and need not, involve consciousness. Indeed, Harnad agrees that "No Turing Test [...] can guarantee that a body has a mind. Worse, nothing in the explanation of its successful performance requires a model to have a mind at all" (1991, 43).

A second variation, directed specifically at AI systems and at addressing the concern that a mindless system might engage in seeming utterances about consciousness, is Schneider and

Turner's (2017) AI Consciousness Test (ACT), which requires that the AI system being tested be trained in such a way that it cannot pass the test simply by mimicking human responses. Specifically, the AI's developers must "deny the AI access to the Internet and prohibit it from gaining too much knowledge of the world, especially information about consciousness and neuroscience" (Schneider 2019, 53–54). The AI system is then asked a series of questions about consciousness-involving concepts such as body-swapping or life after death. If the AI system responds in a fluid and intelligible fashion, then this suffices, according to the ACT, for the system to be conscious. The strategy here might be seen as a variation on the IBE strategy (to be considered in §5), since the idea is that "[w]e can best explain the AI's fluency with such concepts by assuming that it is drawing, as Schneider presumes humans do, upon an introspective familiarity with consciousness" (Udell and Schwitzgebel 2021b, 8).

This is a creative suggestion, but it is subject to two objections, either of which undercuts the ACT's usefulness for determining that there is AIC.

The first, due to Udell and Schwitzgebel (2021b), allows that an AI system might be developed as per the ACT requirement, and that such a system might be fluent with consciousness-implicating questions, but denies that such fluency would be best explained by the system's being conscious, on grounds that "there might be a competing explanation that is equally good [...] solely in terms of lower-level or design-level physical or functional features disconnected from consciousness". A non-consciousness-involving explanation of an AI system's fluency with consciousness-implicating questions may remain more ontologically parsimonious, more theoretically consilient, more systematic, and so on, than a consciousness-involving explanation of such fluency.

The second objection is that there is no hope of developing an AI system in accord with the ACT requirement, since direct and indirect references to consciousness are too pervasively and deeply rooted in our language. In order to fit a model in a way that satisfies the limitations imposed by the ACT requirement, we would be left with extremely scarce training data. Indeed, the strength of today's best models is highly dependent on their ability to fit the model on indiscriminately large bodies of information. But this is impossible to do in a way that satisfies the ACT requirement.

The concern that references to consciousness are too deeply rooted in our language to allow the ACT requirement to be satisfied is not new. Schneider herself considers the concern, and in response offers certain sample questions that do not introduce "expressions like 'consciousness,' 'soul,' and 'mind" (Schneider 2019, 54); and Udell and Schwitzgebel (2021b) press the concern:

> But even employing more neutral language like 'Could you survive the permanent deletion of your program? What if you learned this would occur?' (ACT Sample Question 1, p. 55) requires that the AI understand what 'survival' is, and presumably its opposite. The AI would also need a repertoire of terms through which it could appropriately express the reactions we might expect from it if it were conscious (fear or sadness, perhaps?). The AI can't be entirely prevented from learning about consciousness-adjacent aspects of the world if it's going to be a conversational partner in a test of this sort. But it's unclear how much world knowledge would be too much, in the hands of a sophisticated learning AI. Implementation of the ACT will therefore require careful aim at an epistemic

sweet spot.

Udell and Schwitzgebel go on to assume that such an epistemic sweet spot can be found, focusing rather on the first objection, above. Here we aim to revisit and hammer home the concern that the ACT cannot be implemented, by attention to the 'elementary' question that Schneider and Turner (2017) offer as kicking off the ACT:

> At the most elementary level we might simply ask the machine if it conceives of itself as anything other than its physical self. (2)

Here the supposition is that no terms in this question are 'consciousness-implicating'. But as we'll now argue, at least three terms in this question are consciousness-implicating: 'conceives', 'physical', and 'self'.

First, consider relevant definitions of 'conceiving'. Merriam-Webster, for example, defines 'conceiving' as 'to take into one's mind,' 'to form a conception' or 'to apprehend by reason or imagination'. When one takes an idea into one's mind, one does so by consciously contemplating the idea. Similarly, to apprehend by imagination is to engage in a conscious mental act. Additionally, various terms in these definitions are consciousness-implicating. For example, the Oxford English Dictionary defines 'mind' as 'the element of a person that enables them to be aware of the world and their experiences'.

Next, consider 'physical'. The question of what it is for some goings-on to be physical, especially as input into physicalism, according to which all broadly scientific goings-on are nothing over and above physical goings-on, is a topic of considerable philosophical dispute. This reflects that attempts to define the physical by reference to physics face notorious difficulties, associated most prominently with Hempel's Dilemma (Hempel 1979), according to which the physical cannot be characterized just in terms of physics—not current physics, since this is incomplete and in respects inaccurate, and not future physics, since we don't know what future physics will posit—which opens the door to physics's positing conscious fundamental particles or other goings-on associated with views (panpsychism, strong emergentism) which are supposed to contrast with physicalism. The near-uniform response to these difficulties is to require that, whether or not the characterization of the physical is linked to physics, the physical goings-on cannot be (in particular: fundamentally) mental (see Montero 1999, Wilson 2006, Alter 2024, and many others). Any such characterization of the physical thus implicates consciousness.

Finally, consider 'self'. The Cambridge Dictionary defines 'self' as 'the set of someone's characteristics, such as personality and ability, that are not physical and make that person different from other people.' In this definition, the components of the 'self' are 'not physical'; which as above connotes consciousness. Related definitions make reference to emotions and sensations, which are likewise plausibly consciousness-implicating.

So far we have just considered a single sentence—one supposed by proponents of the ACT to be clearly consciousness-neutral—and found that three of its core terms are, in fact, deeply consciousness-implicating. It is also worth registering just how consciousness-implicating our discourse is. Satisfaction of the ACT constraint would require omission of references to anything involving qualitative experience, including perceived colors and their relations, art, music, food, fashion, architecture—i.e., most human culture. It would require

omission of references to any emotions, including love, hate, joy, sadness, boredom, ecstasy, guilt, gratitude—i.e., most literature, history, sociology, and other expressive and theoretical modes of human expression and exploration for which the emotions are central topics and primary drivers to action. It would require omission of references to anesthesiology, ophthalmology, psychiatry, psychology, child development, and so on. It would require omission of references to any sciences taking the manifest image as a starting point or a constraint on theorizing, and which rely on ultimately perceptual experimental procedures and non-deductive inferential procedures—i.e., all of them. It would require omission of all philosophy of mind, most epistemology, ethics, aesthetics, and any philosophical view taking aspects of human experience or intuition into its purview as part, at least, of what the view aims to explain—i.e., most philosophical views. We conclude that there are no realistic prospects that a model could be fit to a sufficiently large body of data while simultaneously satisfying the ACT requirement, and hence no realistic prospects of hitting the 'epistemic sweet spot' enabling the ACT's implementation.

Summing up: strategies of response to the problem of other AI minds encoded in the Turing Test or its variations are either subject to concerns similar to those undercutting the analogical strategy, or else face insuperable difficulties of implementation. Hence, these strategies do not provide a route to warrantedly establishing that there is AIC.

4 Appeal to Specific Theories of Consciousness

The previous unsuccessful responses to the problem of other AI minds have been broadly 'theory neutral', in not presupposing the truth of any specific theory of consciousness. One might wonder whether progress could be made if, as Udell and Schwitzgebel (2021b) put it, one could "develop the correct theory of consciousness (or a theory close enough to it) and see if the machines fit the bill."

Such a 'theory-heavy' strategy faces the difficulty that there is no consensus about which (if any) such theory is correct, which controversy in turn undercuts attempts to extend any particular theory to the case of AIC. To sidestep this controversy, appeals to theories in contexts where AIC is at issue commonly implement a 'theory-light' strategy,[14] which considers, for some range of neuroscientific candidate theories of consciousness, whether AI systems satisfy certain necessary and/or sufficient conditions specified by the theory. Butlin et al. (2023), for instance, derive "a list of indicator properties" each of which is "said to be necessary for consciousness by one or more theories, and some subsets are said to be jointly sufficient." On this theory-light approach, one might warrantedly determine that a given AI system is (likely to be) conscious (or not) by assessing the extent to which the system has the indicator properties.[15]

Might a theory-light strategy of the sort advanced by Butlin et al. 2023 work as a means of warrantedly establishing that there is AIC? One reason to think not is that the properties which (unlike, e.g., neurobiological conditions) might in principle be had by AI systems are

[14]This strategy extends Birch's (2020) application of a theory-light approach as directed at consciousness in insects and other "biological taxa"; much of what he says about the empirical evidence for neuroscientific theories is consonant with what we say below.

[15]See also Chalmers (2023).

typically broadly functional or computational; but as previously argued, there is no clear route to determining whether AI systems implement the functions at issue.

Here we want to develop another problem for the strategy, aimed at undercutting its operative supposition that neuroscientific theories of human consciousness—the primary source of the indicator properties at issue—are properly seen as bearing on whether there is AIC. More specifically: we will argue that the empirical support for neuroscientific theories (and associated indicator properties) is restricted to their application to humans or other entities with relevantly similar neurobiological substrates. Since AI systems are not neurobiologically similar to humans, none of these theories can be used to show that there is AIC. Hence a 'theory light' approach cannot be used to warrantedly establish that there is (and nor, as we'll discuss in §6, that there is not) AIC. As we'll discuss shortly, McClelland (2025) raises a similar concern, as motivated by attention to the hard problem of consciousness; our aim is to provide further support for this concern, as motivated by attention to the problem of other minds.

To focus ideas, consider Global Workspace Theory (GWT), according to which a state is conscious when it is made globally available to a range of brain processes (see, e.g., Baars 1988, Dehaene et al. 2001, Changeux and Dehaene 2005). Now, as per a recent review of GWT (Mashour et al. 2020), the theory "attempts to account for the main scientific observations regarding the elementary mechanisms of conscious processing in the human brain" in a way presupposing a "neuroanatomical basis".[16] The same is true for all of the neuroscientific theories that Butlin et al. (2023) discuss. In that case, however, Butlin et al.'s claim that these neuroscientific theories "aim to describe correlations between computational processes and consciousness" is incorrect. For as just observed, these theories aim to describe correlations between neurobiological processes and consciousness; and a neurobiological process is not itself a computational process, notwithstanding that the latter might in some sense be instantiated in or realized by the former.

But reflecting Butlin et al.'s claim that "the majority of leading scientific theories of consciousness can be interpreted computationally: that is, as making claims about computational features which are necessary or sufficient for consciousness in humans", could it be that even if claims in and evidence for GWT and other neuroscientific theories of consciousness are not explicitly cashed in substrate-independent computational terms, one can interpret the theories and evidence in such terms, to arrive at (at least some) substrate-independent indicator properties (see, e.g., Dehaene et al. 2017)?

We agree with McClelland (2025) that such extensions of neuroscientific theories go beyond what the evidence supports:

> Even if there is good evidence that the mental states *of organisms* are conscious when broadcast in a global workspace, that is not enough to tell us that a compa-

[16] Hence in Mashour et al. 2020 the "main tenets" of GWT are expressed in human neurobiological terms: "local, specialized cortical processors are linked, at a central level, by a core set of highly interconnected areas (A) containing a high density of large pyramidal neurons with long-distance axons (B). At any given moment, this architecture can select a piece of information within one or several processors, amplify it, and broadcast it to all other processors, thus rendering it consciously accessible and available for verbal report. Recent tracer studies of global feedforward and feedback cortical connectivity confirm a bow-tie architecture with a central core set mostly comprised of parietal and prefrontal areas and forming a structural bottleneck capable of routing information between other cortical processors (C)."

> rable broadcasting of information in an AI would also be conscious. The evidence we have says nothing about whether *non-biological* global workspaces also give rise to phenomenal consciousness. When we take evidence from organic human consciousness and try to apply it to AI, our evidence hits an *epistemic wall*. The conditions of consciousness proposed by GWT are inferred, but the inferences are only warranted if kept within an appropriate scope. AI is beyond that scope. (7)

A similar concern will attach to any broadly computational theory of consciousness.[17]

As McClelland sees it, the problem here reflects the persisting 'hard problem of consciousness', whereby the identification of certain physical and functional correlates of consciousness provides at best shallow explanations of consciousness, incapable of supporting attributions (or denials) of consciousness in cases beyond those (of humans, certain non-human animals) for which the correlations with consciousness have been empirically established. Our variation on or development of the theme of the concern adverts not so much to the explanatory gap associated with the hard problem of consciousness as to the baseline fact driving the problem of other minds, according to which the only direct and decisive evidence of consciousness is first-personal and introspective.

To start: why think that neuroscientific theories only target consciousness in humans or in entities with neurobiological substrates and other features relevantly similar to those of humans? We previously observed that neuroscientists typically presuppose this. But attention to the baseline fact driving the problem of other minds allows us to say more: namely, that this presupposition is non-negotiable. For the empirical evidence for the theories at issue crucially relies on the fact that certain humans—notably: the neuroscientists testing a given theory!—have direct introspective access to their own consciousness—which by analogical inference is warrantedly extended to human test subjects making various reports of or behaving in ways which are known to be correlated with consciousness. It is the introspectively grounded access to consciousness on the part of researchers, warrantedly extended to test subjects which are neurobiologically and otherwise relevantly similar to the researchers for whom consciousness is accessible in first-personal fashion, which provides us with the crucial evidence that this or that functional architecture is associated with consciousness. As such, the support that our 'best supported' neuroscientific theories receive is crucially and ultimately support for this or that functional architecture *as implemented in substrates introspectively known to be associated with consciousness, or substrates sufficiently similar to these*.[18]

The role that direct introspective access to consciousness plays in investigations into and confirmation of 'indicator properties' relevant to consciousness supports McClelland's (2025) observation: even if the evidence supports taking a given function to be necessary and/or sufficient for consciousness when implemented in creatures like us, it does not follow that a similar function, when implemented in entities sufficiently unlike us, is necessary and/or sufficient for consciousness in such entities. As we would put it: from the fact that A and B

[17] Bayne et al. (2024) register, and provide a strategy for responding to, a similar concern under the rubric of the 'generalization problem' for GWT and other neuroscientific theories of consciousness; we address their strategy of response in §5.

[18] See Birch (2020) and Carruthers (2019) for similar observations.

jointly entail C, it doesn't follow that B alone, or B conjoined with something other than A, entails C; and from the fact that function B when implemented in some (relevantly similar) neurobiological substrate A is necessary and/or sufficient for consciousness, it doesn't follow that B alone, or B when implemented in some non-neurological substrate, is necessary and/or sufficient for consciousness.

As a result, granting that "[d]rawing on theories of consciousness is necessary for our investigation because they are the best available guide to the features we should look for", it remains that these theories, and their empirical support, pertain only to creatures introspectively known to be conscious, or to relevantly similar neurobiologically-situated creatures. The further details and controversies about which functions are or are not associated with consciousness don't matter: whatever these may be, in any case the content of and support for these theories is such that there is no warranted basis for their being applied to AI systems. Consequently, a core presupposition of Butlin et al. (2023)'s strategy, according to which neuroscientific theories of human consciousness are properly seen as bearing on whether there is AIC, is reasonably taken to be false. But in that case, there is no theory-based route to warrantedly taking there to be AIC.[19]

5 Inference to the Best Explanation (IBE)

A final response to the problem of other minds appeals to abduction or inference to the best explanation (IBE), as per Harman (1993), Lipton (1991, 2004), Lycan (2002), and Cabrera (2022). In general, an IBE begins by canvassing the relevant facts to be explained; then identifying some appropriately broad range of candidate explanations of these facts; then assessing how well each explanation does by way of satisfying certain theoretical desiderata, including ontological parsimony, general plausibility, theoretical consilience (compatibility with our best sciences), fruitfulness, systematicity, and so on; and finally identifying the explanation that is best in maximizing satisfaction of the desiderata, as that which should be believed.

What facts might an IBE-based strategy of response to the problem of other minds aim to explain? When applied to other humans, the facts to be explained will include those registering the physical, biological (including evolutionary), behavioural (including linguistic behavioural), and functional respects of similarity between humans; in this sense, the IBE strategy enfolds the analogical and Turing test strategies. The facts to be explained will also include facts about empirically confirmed results pertaining to neurological and other theories of human consciousness; in this sense, the strategy enfolds the theory-based strategy. More generally, the IBE strategy as applied to humans can take any relevant facts or data into account, including that the target humans self-report that they are conscious, are treated by others as conscious beings (as when a parent comforts their child after a scraped knee), and manifest various (functional and other) neurological properties which are correlated with consciousness.

The IBE strategy as applied to other humans is arguably successful; as Chalmers (1996) puts it, it seems that an IBE "is as good a solution to the problem of other [human] minds

[19] Modulo the expanded conception of a theory-based approach under the rubric of 'the NK strategy', to be considered in the next section.

as we are going to get" (246). Plausibly, consciousness-involving explanations on which other humans are conscious will fare better than alternative explanations on which other humans are not conscious, notwithstanding that other humans resemble humans known to be conscious in multiple relevant respects.[20] In the case of other humans, non-consciousness-involving explanations of the salient facts might appeal to solipsism, delusion, or skepticism; but such explanations incur substantive theoretical costs not incurred by a consciousness-involving explanation. To consider a few dimensions of theoretical virtue: unlike a consciousness-involving explanation, the alternative explanations are highly revisionary with respect to either the presumed truth-value or presumed content of many claims (e.g., one according to which you and I are each conscious). Unlike a consciousness-involving explanation, the alternative explanations fail to comport with our best scientific theories—both generally, in failing to accommodate there being stable connections between physical bodily processes and states of human organisms, and specifically, in failing to accommodate well-entrenched scientific theories or claims, including that there are specific neural correlates and associated phenomena reliably associated with aspects of conscious experience, and that consciousness has been naturally selected for. Unlike a consciousness-involving explanation, the alternative explanations are unsystematic, offering completely different accounts of attributions of consciousness by one's self and by other humans. It might also be, depending on details, that a consciousness-involving explanation is more parsimonious than the alternative explanations, since the latter have to posit new phenomena in order to explain away the seeming evidence of consciousness in other humans. Here there's no real controversy: IBE clearly supports taking other humans to be conscious.

Might the IBE strategy work for AI systems? We'll now argue that it will not.

To start, much of the data input into the IBE strategy when applied to other humans is simply lacking for AI systems. In particular, as previously noted, relevant physical, biological (including evolutionary), and neurological (substrate-involving) functional similarities are not in place, so the data doesn't include any such facts to be explained. Remaining salient facts would include that AI systems can engage in what appears to be conversation, are treated by some humans as conscious, and that they respond 'yes' when asked whether they are sentient.[21] In addition, and as a sophisticated implementation of the IBE strategy, it has been suggested, per the 'iterative natural kind (NK) strategy' to be discussed shortly, that an IBE-based empirical investigation into consciousness as a natural kind associated with a 'nomological cluster' of indicator properties, which are typically though not always co-instantiated, might warrant taking certain properties that AI systems can or do instantiate to support the posit of AIC.

Focusing first on the remaining salient facts, it seems clear that the best explanation of these facts will not be one according to which there is AIC. Recall that best explanations are ones that are best all things considered, taking various theoretical desiderata into account. To fix ideas, let's focus on ontological parsimony and theoretical consilience. A parsimonious and theoretically consilient explanation of the remaining salient facts plausibly

[20] Certain of the comparative assessments to come are in the same vein as, but perhaps even stronger than, those taken to support an IBE-based response to the problem of external world skepticism (see, e.g., Shepherd (1827) and Russell (1967)).

[21] We put aside that this latter fact isn't especially stable; as Udell and Schwitzgebel (2021b) note, AI systems will also respond 'yes' when asked whether they are not sentient.

runs as follows: that AI systems seem to engage in conversation, and sometimes respond 'yes' when asked if they are conscious, reflects that AI systems are nonlinear probabilistic prediction models, run on non-organic substrates that we have no independent reason to think support consciousness, which are designed to mimic human dialogue, and which are trained and fine-tuned with massive amounts of text; and that AI systems are treated by some humans as conscious reflects our human tendency to anthropomorphize things that mimic us in linguistic or other behavioural respects.[22] This explanation of the remaining salient facts does not involve, require, or support AI systems' being conscious.[23] Alternative explanations of the remaining salient facts which do crucially appeal to AI systems' being conscious will be less ontologically parsimonious, since the posit of AIC involves an ontological posit going beyond those accepted in the previous explanations of the facts at issue.[24] Such AIC-involving explanations will also be less theoretically consilient than explanations not involving AIC, since again, we have no antecedent motivation (intuitive or scientific) for taking any non-organic systems to be conscious.[25]

Some might see this as enough to show that the IBE-based strategy won't work to warrantedly establish that there is AIC. But one final gambit remains. Recall McClelland's concern that the empirical evidence input into our theories of consciousness can't be assumed to apply beyond the humans (or close neurobiological cousins) from whom the evidence was drawn (a concern we endorse, and offered further support for). After registering McClelland's concern under the rubric of the 'the generalization problem', Bayne et al. (2024) suggest that "the iterative natural kind (NK) strategy [...] has the tools to address the generalization problem that troubles the theory-based approach" (461–2). As they describe the strategy:

> [The NK strategy] involves treating consciousness as a natural kind (Block 2007, Shea and Bayne 2010). A natural kind category is a category that cuts the world 'at its joints', grouping items based on commonalities in their underlying nature (as opposed to superficial similarities, arbitrary conjunctions of features, or mere human interests) [69–72]. If consciousness is a natural kind, then conscious systems will share an underlying nature, and in principle that nature should be identifiable via the kinds of iterative procedures that have succeeded in uncovering the underlying nature of other natural kinds, such as heat [...].

As Block (2007) notes, this methodological strategy proceeds via inference to the best explanation:

> The alternative I have in mind is just the familiar default "method" of inference to the best explanation, that is, the approach of looking for the framework that

[22] See, e.g., Epley et al. (2007) and Urquiza-Haas and Kotrschal (2015).

[23] In §6, we'll discuss why this explanation should not be understood as entailing, moreover, that AI systems are not conscious.

[24] What exactly this ontological posit comes to will depend on further details—perhaps it involves a new type of consciousness, perhaps it involves a new type of realizer for consciousness of an already-accepted type.

[25] Supporting these judgments are the sorts of comparative assessments operative in the nearly uniform judgments by those working in the field of AI, to the effect that, e.g., the conversational interactions between Blake Lemoine and LaMDA do not support taking LaMDA to be sentient, given that LaMDA is a fitted statistical model like any other and given that Lemoine is, like humans generally, susceptible to anthropomorphic projection. See Metz (2022) for a relevant survey.

> makes the most sense of all the data, not just reports (Harman 1965; Peirce 1903, Vol. V, p. 171). (486)

As per the reference to 'reports', Block's paper was directed at identifying a test for consciousness that would non-trivially apply to humans who for whatever reason (notably: being in a vegetative state) were unable to engage in self-reports or behaviours standardly taken be evidence of consciousness in normal humans; and similarly for Shea et al.'s paper. Rather than take these indicators to be criterial of consciousness (implausibly, given that infants are similarly situated), the authors suggested that one rather proceed on the supposition that consciousness is a natural kind, which like other natural kinds (e.g., heat, H_2O) is associated with, and explanatory of, 'nomological clusters' of properties which are commonly, but not necessarily, co-instantiated (Shea and Bayne 2010, 471). Such a methodology might make room for warrantedly taking the natural kind to be instantiated even in the absence of the usual indicator properties (as in the case of H_2O in the atmosphere of a distant planet). And the authors identify several broadly neurological properties that 'cluster together' with reports in normal humans, and which at least some persons in vegetative states can instantiate.

So far, so good, for investigations into the consciousness of humans in non-standard circumstances; but how does the NK strategy assist with the generalization problem? Block and Shea et al. suggested, and Bayne et al. (2024) endorsed, the possibility that we can discover new indicator or evidential properties, which could then be iteratively added to the nomological cluster relevant to confirming the presence of the natural kind. Supposing so, the more recent suggestion is that we might ratchet our way to the identification of indicator properties of consciousness capable of being instantiated in AI systems. As Bayne et al. (2024) describe the process:

> If we can identify an underlying feature (or features) that accounts for the pre-theoretical measures of consciousness with which we began, we would have a principled basis for generalizing beyond 'consensus cases' (roughly, human beings who are capable of producing reports) and applying our C-tests to the members of other populations [...]. The idea here is that C-tests must first be validated in 'neighboring' populations before being applied to more 'alien' populations, because the closer a population is to 'us', the more credence we should have in our pre-theoretical measures of consciousness.
>
> In embracing this hierarchical approach to validation, we leave open the important question of which populations are 'closest' to the consensus population, for proximity can be measured in various ways (e.g., behaviorally, functionally, neurophysiologically).

We reply that implementing the iterative NK strategy will not work to overcome the generalization problem to the extent required to warrant a claim of AIC.

To start, the analogy with natural kinds such as heat and H_2O only goes so far. In these other cases, there is a single underlying physical substance/structure/process whose generation of different indicator properties in different circumstances is just a matter of physics. In the case of consciousness, it is very unclear what 'underlying feature (or features)' (beyond

the bare posit of consciousness) might be generative of indicator properties even in normal human subjects, much less in increasingly alien populations. Indeed, as Block and Shea et al. initially implemented the strategy, the idea was not so much to use the NK strategy to identify what underlies and accounts for indicator properties in order to track consciousness in non-standard contexts, but rather to suppose that, since consciousness qua natural kind would be associated with diverse indicator properties, one might reasonably—again, as an IBE—take someone instantiating some but not all of the usual indicator properties to be conscious. Hence Block and Shea et al. maintain that, given that other similarities and certain indicator properties of consciousness are in place as between normal humans and those in vegetative states, consciousness might be warrantedly attributed to the latter. The reference to similarities and other indicator properties is crucial; for as Shea and Bayne (2010) go on to note, it is difficult to apply the NK methodology in cases of entities which are "very different from us":

> Suppose, the critic might say, that consciousness can be realised by different physical structures and in different cognitive architectures. If that is so, then how can we use physical and functional measures with which consciousness happens to be associated in us——measures that might not be essential to consciousness as such——as a guide to the presence of consciousness?
>
> The question is a good one, but it needs to be handled with some care. Worries about multiple realisability have most force when we are dealing with creatures that are very different from us, such as Martians, robots, or even octopuses. In such cases, the natural kind methodology will be extremely difficult to apply (Block [2002]).

So the NK strategy in itself does not overcome the generalization problem. Nor is it clear how an iterative version of this strategy might do so. Initial applications of the NK strategy are, as above, to populations that are similar to normal humans in various salient respects ("e.g., behaviourally, functionally, neurophysiologically") (Bayne et al. (2024). Moreover, these initial applications did not advance any new indicator properties or associated 'tests' of consciousness—rather, it was just deemed explanatorily plausible, given the presence of similarities and some indicator properties, that (some) members of the neighbouring population might be conscious. All this is compatible with the iterative NK strategy requiring that more distant populations instantiate at least some of the usual indicator properties—which AI systems don't.[26] Perhaps the idea is that we can iterate the process to populations which are even less similar—say, to certain non-human animals who both lack the ability to report on their consciousness and are somewhat different neurobiologically. And perhaps in some such cases we might come to identify some new indicator properties to add to the cluster associated with consciousness. In turn, we might hope that, in principle, AI systems could instantiate the new indicator property. But so far, this is just speculation. What would such

[26] AI systems aren't similar in neurophysiological respects, the best explanation of their similar linguistic behaviour does not advert to consciousness, and even if we were in position to determine that AI system processing is functionally similar to that operative in human consciousness (which we presently aren't and may never be) the generalization problem calls into question whether mere functional similarity counts as a salient dimension of similarity.

indicator properties be? If these are functional properties associated with neurobiological or other organic bases, then (and once again granting that it could be determined that AI systems implement the function), the generalization problem will remain in place, and the original verdict that "the natural kind methodology will be extremely difficult to apply" will undercut the prospects of running a natural kind-based IBE to to the warranted posit of AIC.

We conclude that an IBE-based strategy, whether as directed at the limited set of remaining salient facts above, or as implemented via the iterated NK strategy, does not provide a route to warrantedly establishing that there is AIC.

6 The Bearing of the Previous Results on Negative Attributions of AIC

We have argued that none of the strategies for responding to the problem of other minds for other humans or certain non-human animals provide a basis for responding to the problem for AI systems. Does it follow that we can conclude that AI systems are *not* conscious? No, for two reasons illuminated by attention to the problem of other minds.

First, recall the baseline fact driving this problem, according to which each of us has direct epistemic access only to our own consciousness, via introspection. Traditional responses to this problem build on this access, in ways plausibly supporting positive attributions of consciousness to certain other entities. For example, analogical arguments rely on targets' having features similar to those empirically associated with our own introspectively available conscious states. But this line of positive support does not entail that the failure of some other entity to have many, some, or even any of these features suffices to establish that the target entity is (likely to be) *not* conscious. To draw such a conclusion would appear to involve an inappropriately anthropocentric overgeneralization: why suppose that only creatures relevantly similar to ourselves in respects correlated with our consciousness are conscious? Indeed, philosophical discussions of mentality standardly assume that conscious mental states are multiply realizable.

Note that the case being made here is not that from the failure of the traditional strategies to apply to a target, one can never infer to a negative attribution of consciousness in the target. On the contrary, in cases where—as with other humans or certain non-human animals—we have good empirical evidence of the typical connection between satisfaction of the conditions and consciousness, the absence of the conditions in a token member of the target category might well provide a warranted basis for taking the token member to lack consciousness (see McClelland 2025). But in the case of AI systems and any other systems not typically associated with those conditions, and given that the ultimate arbiter for a system's being conscious consists in the system's having direct introspective experience of consciousness, the failure of such systems to manifest the conditions provides no route to their not being conscious.

Second and relatedly, though what it takes for a process to be introspective is controversial (see Schwitzgebel (2024)), for present purposes what is crucial is that a positive introspective finding of consciousness will trump any negative finding output by any indirect

(non-introspective) route.

For example, consider the IBE strategy. For purposes of tracking what is actually the case, a given IBE is only as good as the data (the facts to be explained) upon which it operates. In particular, if a given IBE operates upon incomplete data, then the best explanation of this data might well not be the best, taking all the data into account (see Biggs and Wilson 2017). There's nothing unusual about this. Even the conclusions of modus ponens arguments are only as good as their premises, so far as tracking what is actually the case. But this feature of IBE takes on special importance in the context of the problem of other minds. For this problem gets started precisely because our access to some clearly relevant data—namely, that encoding whether a given entity is directly introspectively aware of itself as conscious—is necessarily restricted to our own case.

Now, responses to the problem of other minds aim to sidestep this asymmetry by identifying alternative modes of justification, besides direct introspection, for taking another entity to be conscious; and in the case of humans and certain non-human animals the IBE strategy plausibly succeeds in doing this. But it remains that the data upon which this strategy operates is necessarily incomplete, since it does not include the evidence of introspection on the part of the target entity; and it moreover remains that if AI systems did turn out to have introspective access to their own consciousness, this fact alone would undercut the negative conclusion of the previous IBE.

Hence it is that the introspective epistemic underpinning of the problem of other minds itself blocks our being able to infer, from the failure of IBE to support the consciousness of AI systems given the relevant data *to which we have access*, to the further conclusion that AI systems are not conscious. And more generally, notwithstanding that considerations of analogy, variations on the Turing test, specific theories, and IBE all fail to provide positive support for AIC, it remains that AI systems might be introspectively conscious, and that such a positive introspective finding would trump these or any other negative findings. Hence from the failure of historically salient responses to the problem of other minds to warrantedly establish that there is AIC, it does not follow that we are warranted in taking AI systems *not* to be conscious. Correspondingly, none of the historically salient means of establishing that, or whether, some target entities are conscious serve to establish that, or whether, there is AIC.

7 Concluding Remarks

We have argued that, for any existing strategy of response to the problem of other minds as directed at AI systems, there are no prospects of the strategy's establishing that there is AIC. Even so, we have also argued, from the failure of any or all of these strategies, it does not follow that we can conclude that AI systems are not conscious. Though these strategies fail, one and all, AI systems might still be conscious. We just can never know; and correspondingly, we should be agnostics about AIC. The upshot is that there a distinctively difficult problem of other AI minds, whose lack of principled resolution offers a new line of support for what McClelland (2025) calls 'agnosticism about artificial consciousness', and which is likely, in turn, to stymie attempts to offer any justified answers to questions about the personal, moral, and agential status of AI systems.

References

Adadi, A. and M. Berrada (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access 6*, 52138–52160.

Alain, G. and Y. Bengio (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Alter, T. (2024). Physicalism and fundamental mentality. *Synthese 204*(2), 1–18.

Andrews, K. and J. Birch (2023). To understand AI sentience, first understand it in animals. *Aeon Essay*.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion 58*, 82–115.

Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

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

Babic, B. and I. G. Cohen (2023). The algorithmic explainability 'bait and switch'. *Minnesota Law Review 108*, 857–909.

Babic, B., S. Gerke, T. Evgeniou, and I. G. Cohen (2021). Beware explanations from AI in health care. *Science 373*(6552), 284–286.

Bayne, T., A. K. Seth, M. Massimini, J. Shepherd, A. Cleeremans, S. M. Fleming, R. Malach, J. Mattingley, D. K. Menon, A. M. Owen, M. A. K. Peters, A. Razi, and L. Mudrik (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences 29*.

Biggs, S. and J. M. Wilson (2017). The a priority of abduction. *Philosophical Studies 174*(3), 735–758.

Birch, J. (2020). The search for invertebrate consciousness. *Noûs 56*(1), 133–153.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences 30*(5), 481–548.

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon, and R. VanRullen (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Cabrera, F. (2022). Inference to the best explanation - an overview. In L. Magnani (Ed.), *Handbook of Abductive Cognition*, pp. 1–34. Springer.

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

Cartwright, N. (1979). Causal laws and effective strategies. *NOÛS 13*(4), 419–438.

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

Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review 1*.

Changeux, J. and S. Dehaene (2005). Ongoing spontaneous activity controls access to consciousness: A neuronal model for inattentional blindness. *PLoS Biology 3*(5), e141.

Chowdhury, N., D. Johnson, V. Huang, J. Steinhardt, and S. Schwettmann (2025, April). Investigating truthfulness in a pre-release o3 model.

Cunningham, H., A. Ewart, L. Riggs, R. Huben, and L. Sharkey (2023). Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Dehaene, S., M. Kerszberg, and J. Changeux (2001). A neuronal model of a global workspace in effortful cognitive tasks. *Pnas 95*(24), 14529–14534.

Dehaene, S., H. Lau, and S. Kouider (2017). What is consciousness, and could machines have it? *Science 358*(6362), 486–492.

Epley, N., A. Waytz, and J. Cacioppo (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review 114*(4), 864–86.

Goldstein, S. and B. A. Levinstein (2024). Does ChatGPT have a mind? *arXiv preprint arXiv:2407.11015*.

Grosse, R., J. Bae, C. Anil, N. Elhage, A. Tamkin, A. Tajdini, B. Steiner, D. Li, E. Durmus, E. Perez, E. Hubinger, K. Lukošiūtė, K. Nguyen, N. Joseph, S. McCandlish, J. Kaplan, and S. R. Bowman (2023). Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Hampshire, S. N. (1952). The analogy of feeling. *Mind 61*(January), 1–12.

Harman, G. (1993). Can science understand the mind? In G. Harman (Ed.), *Conceptions of the Human Mind: Essays in Honor of George A. Miller*, pp. 111–121. Hillsdale, NJ: Lawrence Erlbaum.

Harnad, S. (1989). Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence 1*(4), 5–25.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines 1*(1), 43–54.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Volume 2. New York: Springer.

Hempel, C. (1979). Comment at a symposium on Nelson Goodman's *Ways of Worldmaking*. In *Presented at the 76th Annual Meeting of the American Philosophical Association*.

Hempel, C. G. (1945a). Studies in the logic of confirmation (i.). *Mind 54*(213), 1–26.

Hempel, C. G. (1945b). Studies in the logic of confirmation (ii.). *Mind 54*(214), 97–121.

Hornik, K., M. Stinchcombe, and H. White (1989). Multilayer feedforward networks are universal approximators. *Neural networks 2*(5), 359–366.

Lipton, P. (1991). The best explanation. *Cogito 5*(1), 9–14.

Lipton, P. (2004). *Inference to the Best Explanation*. Routledge/Taylor and Francis Group.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.

Low, P., J. Panksepp, D. Reiss, D. Edelman, B. Van Swinderen, and C. Koch (2012). The Cambridge declaration on consciousness. In *Proceedings of the Francis Crick Memorial Conference*, Volume 7, pp. 1–2. England: Cambridge.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, Volume 30, pp. 4768.

Lycan, W. G. (2002). Explanation and epistemology. In P. K. Moser (Ed.), *The Oxford Handbook of Epistemology*, pp. 413. Oxford University Press.

Martín, J. C., M. Ángel Guzmán Caba, S. J. Láinez, and L. F. Peña (2023). Explainable artificial intelligence (XAI) challenges of model interpretability. *Management Solutions*.

Mashour, G. and M. Alkire (2014). Evolution of consciousness: Phylogeny, ontogeny, and emergence from general anesthesia. In J. C. A. C. J. Cela-Conde, R. G.L̃ombardo (Ed.), *In the Light of Evolution: Volume VII: The Human Mental Machinery*. Washington, D. C.: National Academies Press.

Mashour, G., P. Roelfsema, J. Changeux, and S. Dehaene (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron 105*(5), 776–798.

McClelland, T. (2025). Agnosticism about artificial consciousness. *Mind & Language n/a*(n/a).

Metz, R. (June 14, 2022). No, Google's AI is not sentient. *CNN*.

Mill, J. S. (1865). *An Examination of Sir William Hamilton's Philosophy*. London: Longman, Green, Reader and Dyer.

Montero, B. (1999). The body problem. *NOÛS 33*(2), 183–200.

Ng, A. and M. Jordan (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in neural information processing systems (NeurIPS)*, Volume 14, pp. 841–848.

Olah, C., N. Cammarata, L. Schubert, G. Goh, M. Petrov, and S. Carter (2020). Zoom in: An introduction to circuits. *Distill*. https://distill.pub/2020/circuits/zoom-in.

Oppy, G. and D. Dowe (2021). The Turing Test. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 ed.). Metaphysics Research Lab, Stanford University.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Volume 16, pp. 1135–1144. Association for Computing Machinery.

Rubinstein, Y. D. and T. Hastie (1997). Discriminative vs informative learning. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD)*, Volume 5, pp. 49–53. Stanford University.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence 1*(5), 206–215.

Russell, B. (1912/1967). *The Problems of Philosophy*. Oxford: Oxford University Press.

Russell, B. (1948). *Human Knowledge: Its Scope and Limits*. London and New York: Routledge.

Schneider, S. (2019). *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press.

Schneider, S. and E. Turner (2017). Is anyone home? A way to find out if AI has become self-aware. *Scientific American Blog Network*.

Schwitzgebel, E. (2024). Introspection. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024 ed.). Metaphysics Research Lab, Stanford University.

Schwitzgebel, E. and J. Pober (2026). The mimicry argument against AI consciousness. *Philosophers' Imprint 26*(18).

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences 3*(3), 417–57.

Shea, N. and T. Bayne (2010). The vegetative state and the science of consciousness. *British Journal for the Philosophy of Science 61*(3), 459–484.

Shepherd, M. (1827). *Mary Shepherd's Essays on the Perception of an External Universe*. New York, NY: Oxford University Press.

Stanford, K. (2023). Underdetermination of scientific theory. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 ed.). Metaphysics Research Lab, Stanford University.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind 59*(236), 433–460.

Udell, D. B. and E. Schwitzgebel (2021a). Susan schneider's proposed tests for AI consciousness: Promising but flawed. *Journal of Consciousness Studies 28*(5-6), 121–144.

Udell, D. B. and E. Schwitzgebel (2021b). Susan Schneider's proposed tests for AI consciousness: Promising but flawed. *Journal of Consciousness Studies 28*(5-6), 121–144.

Urquiza-Haas, E. G. and K. Kotrschal (2015). The mind behind anthropomorphic thinking: attribution of mental states to other species. *Animal Behaviour 109*, 167–176.

Wachter, S., B. Mittelstadt, and C. Russell (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology 31*(2), 841–887.

Wilson, J. M. (2006). On characterizing the physical. *Philosophical Studies 131*(1), 61–99.