

# The Problem of Other AI Minds

Boris Babic\* and Jessica Wilson†

April 11, 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)? We argue for a provisionally pessimistic answer—probably not—by attending to the traditional problem of other minds. For each of the main response strategies to this problem, we argue that even if the strategy works to establish that other humans and some non-human animals are conscious, the prospect of the strategy’s working to establish that there is AIC—and more generally, to establish whether there is AIC—is unlikely. The upshot is that there a distinctively difficult problem of other AI minds, whose lack of principled resolution is likely to undermine attempts to answer questions about the personal, moral, and agential status of AI systems.

**Keywords:** Consciousness, Cognition, Artificial Intelligence, Identity, Mind

## Introduction

The question of whether sufficiently complex artificial intelligence (AI) systems are conscious, in the broad sense of being experientially aware, is methodologically interesting, normatively important, and politically salient. Methodologically, the advent of AI systems for which the attribution of consciousness is not completely absurd constitutes a new source of data as input into our theorizing about consciousness. Whether such attributions (or their denials) can be supported bears on questions of normative ethics (concerning, e.g., moral agency, moral personhood, and moral patienthood) and philosophy of action (concerning, e.g., intentionality and goal directedness). Meanwhile, from a legal, ethical, and regulatory perspective, the status of AI systems as conscious (or not) bears on the rights, duties, privileges, and liabilities of AI systems and their manufacturers. But can we warrantedly establish whether a complex AI system is conscious—i.e., whether there is AI consciousness (henceforth, ‘AIC’)?

---

\*Department of Philosophy and Institute of Data Science, University of Hong Kong; [babic@hku.hk](mailto:babic@hku.hk)

†Department of Philosophy, University of Toronto; [jessica.m.wilson@utoronto.ca](mailto:jessica.m.wilson@utoronto.ca)

In what follows, we argue for a provisionally pessimistic answer: we are not now, and we are unlikely to be in the foreseeable future, in position to establish whether there is AIC.

The scope of our argument is broad, but not so broad as to imply a true negative universal; for 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 models might transpire as a result of such limitations is anyone’s guess. Correspondingly, we do not aim to argue that we will not be in position to warrantably establish that AI models radically different from those which we have today are (are not) conscious: any such argument would be unduly speculative. Even so, our results are significant: at least so far as present-day AI systems and their anticipated descendants are concerned, we are not in position to warrantably establish that, or whether, any such systems are conscious. Or so we will argue.

Our discussion proceeds by attending to the traditional problem of other minds, and to a range of strategies that have been or might be offered in response to this problem. Schematically, the problem of other minds may be framed as follows:

Given that each of us has introspective access only to our own consciousness, how can any of us warrantably establish that another entity is conscious?

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 warrantably 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?

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 warrantably 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. Note that traditional versions of the problem of other minds in the first instance focus on what it would take to establish of some other entity *that* it is conscious (i.e., to establish an affirmative result), as opposed to what it would take to establish of some other entity *whether* it is conscious (i.e., to establish either an affirmative or a negative result). As we will argue, however, the two questions are related, in that for any given response strategy to the problem of other minds, failure of some target entity to be deemed conscious by lights of the strategy does not suffice to establish that the target entity is *not* conscious, for reasons stemming from the baseline fact that our only direct access to consciousness is via introspection. Accordingly, we can consider our epistemic prospects for establishing *whether* there is AIC by considering our epistemic prospects for establishing *that* there is AIC, taking available responses to the problem of other minds as a guide.

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.<sup>1</sup> But exploring that connection can be very fruitful. For example, as we discuss below, doing so is relevant to assessing the bearing of the emerging literature on AI explainability and interpretability<sup>2</sup> on the prospects for there being AIC (especially in light of some recent positive proposals).<sup>3</sup>

---

<sup>1</sup>Harnad (1991) is an exception, stating: “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”. We will revisit certain aspects of Harnad’s and Schneider’s discussions below.

<sup>2</sup>See, e.g., Rudin (2019), Babic et al. (2021), and Babic and Cohen (2023).

<sup>3</sup>See, e.g., Schneider (2019), Butlin et al. (2023), and Chalmers (2023).

We proceed by considering the main strategies of response to the problem of other minds. For each strategy, we argue that even granting that the strategy works to establish that other humans and some non-human animals are conscious, the prospects are dim that the strategy will work to establish that there is AIC. These strategies include appeals to an argument by analogy (§1), appeals to inference to the best explanation (§2), appeals to variations on the Turing test (§3), and appeals to one or more substantive theories of consciousness (§4). In each case, we argue, the failure of the strategy to establish that there is AIC does not suffice to establish that there is *not* AIC, and so the strategy fails also to establish *whether* there is AIC. The upshot is that we are not presently in position to warrantably establish whether there is AIC, and modulo some unforeseen strategy, are unlikely to ever be able to do so.

We conclude (§5) that there is a distinctively difficult ‘problem of other AI minds’, whose lack of principled resolution is likely to stymie attempts to offer justified answers to questions about the personal, moral, and agential status of AI systems. We also draw an underappreciated moral of the problem of other minds—namely, that it is unlikely that we will ever be in position to identify any necessary non-introspective conditions for consciousness or its attribution.

## 1 Argument by Analogy

### 1.1 The general strategy

The most common response to the traditional 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.<sup>4</sup> As applied to other humans, bodily and behavioural similarities are salient, as in Mill’s analogical argument:

**Millian 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)

Contemporary advances in neuroscience and evolutionary biology have provided a further rich source of relevant physical, biological, and functional respects of similarity in humans.

---

<sup>4</sup>See, e.g., Blackburn (1994), Avramides (2020), and many others.

Many scholars point to empirical evidence pertaining to, e.g., identification of distinct characteristic patterns in electroencephalograms in humans when awake or when asleep or under anesthesia, differential physical impairments of consciousness, and evidence suggesting that cortical activity is associated with recurrent local feedback ([Mashour and Alkire, 2014](#)). While these advances add more texture to Mill’s argument, the basic analogical strategy remains the same.

To be sure, there remains the further question of how, exactly, these features are related to consciousness; and similarly for the behaviours that are typical concomitants of specific conscious states or experiences. Luckily, applying the argument by analogy 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, for purposes of an adequate response to the problem of other (human) minds.

Such analogous 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 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 subservise similar network functions that can be quantitatively assessed and compared with human findings. (pg. 10361)

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 the Millian analogical argument—one on which certain similarities are taken to warrant the positive attribution of consciousness, not just to other humans, but also to certain non-human animals. Further, judgment calls as to which similarities are relevant and which are not are closely intertwined to our introspective judgments about which features are closely tied to consciousness. We can thus plausibly grant 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 almost certainly will not.

## 1.2 The core elements of AI systems

We start by providing a sketch of the core elements of AI systems. 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, this kind of model can be seen as a universal approximator of any function (Hornik et al., 1989).

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 simply reflects that as computer scientists started training increasingly larger and larger models, using energy intensive GPUs, it became computation-

ally 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.<sup>5</sup>

### 1.3 Physical constitution

We now argue that there are in-principle difficulties with extending the Millian analogical argument to the case of AIC.

First, and perhaps most obviously, 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, which in admitting of multiple realizations at a time, are more like repeatable types or universals than token concrete particulars. Hence, for example, it is natural to speak of Open AI’s GPT 4 as 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; but unlike the organic, spatiotemporally localized and bounded bodies of human and non-human animals, this hardware is widely distributed and massively parallelized; it does not 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

---

<sup>5</sup>Modulo, that is, our aforementioned exclusion of inconceivable or radically different ‘intelligent’ systems whose performance is completely unlike the systems we have today—e.g., omniscient computing systems capable of deterministically predicting the future.

associated argument by analogy.

It is moreover 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 ‘neural’ networks. Take the simplest such network: the single layer perceptron ([Rosenblatt, 1957](#)). This is an approximately linear/additive model with a weight and bias term informing a prediction. For example, suppose that one wants to predict a person’s weight in pounds on the basis of their height in centimeters. After making some observations, one might estimate that a good prediction multiplies the person’s height by 1.1 and subtracts 10 pounds, such that, e.g., a person who is 180cm tall would be predicted to weigh 188 pounds. A single layer perceptron stops with these two parameters. When more layers are added to a model, as in the case of modern deep learning neural networks, the model takes many weighted combinations of the initial parameter values in order to capture additional latent (not directly observed) structure of the data. When new data is processed, every weight is updated, using backpropagation and gradient descent ([Hastie et al., 2009](#)). Two points of difference are salient. First, neural networks are fit using digital processing systems; but neural processing in human/non-human animals is analog; hence there’s no constitutional physical similarity underpinning the use of ‘neural’ in ‘neural network’.<sup>6</sup> Second and perhaps relatedly, neural processing in conscious humans and non-human animals proceeds very differently from backpropagation and updating via gradient descent. Regarding backpropagation: When a human sees an image, for example, the input does not and indeed cannot hit every neuron in the brain (corresponding to the weights in a ‘neural’ network)—that would be paralyzing. There is clearly nothing in the brain resembling such a process. Regarding gradient descent: it would be at best metaphorical to argue that we as humans implicitly use an optimization procedure to locally minimize an implicit cost function. Hence, talk of AI systems as ‘neural’ networks fails to serve as a basis for a physical analogy between such systems and creatures warrantedly taken to be conscious.

Of course, an AI system can be embedded in a robotic body, and to this extent might 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 the dissimilarities between how AI systems

---

<sup>6</sup>[Arvan and Maley \(2022\)](#) moreover argue that analog processing is likely critical for macro-conscious experience.

(whether or not robotically embedded) and conscious animals process information, and—most importantly—given that no parts of an inorganic robot body are empirically associated with consciousness.<sup>7</sup>

## 1.4 Outward behaviour

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. But since AI systems have been specifically designed to mimic human linguistic behaviour, in a way bearing no resemblance to the way that humans engage or learn to engage in linguistic behaviour, and which we have no antecedent reason to think involves consciousness, the similarity at issue provides little if any independent support for the posit of consciousness.<sup>8</sup>

To see why this is so, it is worth expanding on the role that behavioural similarity plays in the Millian analogical argument. Crucially, the behavioural analogy 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”. If Mill justifiably believed that the similar behaviour was generated by a hidden puppeteer, then the behavioural similarity wouldn't support the analogy. Hence, there is an underlying logic in the Millian analogical argument as to which similarities matter and why. More generally, the (potential) success of an argument by analogy to some target entity's being conscious requires not just that the target have features similar to those had by conscious entities, but that it be plausible that the similar features have the same cause or source—namely, consciousness.<sup>9</sup>

For 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

---

<sup>7</sup>The analogy might be resurrected in the case of an organic ‘robotic’ body which physically duplicates a conscious human or non-human body (or relevant neurobiological parts thereof)—but in that case the support for consciousness would attach to the robotic creature, not the AI system somehow associated with that creature.

<sup>8</sup>This line of thought also undercuts the viability of the Turing test for purposes of warrantably establishing that there is AIC, to be further discussed in §3.

<sup>9</sup>This observation serves also to deflect concerns that there are always indefinitely many similarities between any two entities.

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.

To be sure, that a given behavioural similarity has been ‘built in’ might not completely undercut its import, if the design’s implementation proceeded via effective duplication of the physical and biological basis of some entity justifiably believed to be conscious. But for AI systems, linguistic mimicking is accomplished via a probabilistic prediction model trained on massive amounts of human-generated text, in a way bearing no resemblance to the processes via which humans engage or learn to engage in linguistic behaviour, and which is a matter of computational processes which we have no antecedent reason to think involve consciousness.<sup>10</sup>

The underlying dissimilarity here is related to another. Both humans and non-human animals naturally evolved in circumstances where consciousness of internal states and external environments was plausibly selected for; and this similarity serves as a ‘meta-level’ similarity which supports taking the first-order similarities to be relevant to an argument by analogy. Conversely, that AI systems are the product of design aimed at generating behaviours and features similar to those of conscious humans undercuts taking the first-order similarities to be relevant to the argument by analogy.

## 1.5 Functional architecture and AI explainability

It remains to consider whether a final potential respect of similarity—namely, one according to which AI systems have an internal functional architecture similar to that operative in conscious humans and some non-human animals—might support an argument by analogy for AIC.<sup>11</sup> Here again, the prospects are dim.

---

<sup>10</sup>This is the linguistic application of Schneider’s (2019) observation 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.’s (2023) observation (following Andrews and Birch (2023)) that “AI systems may be trained to mimic human behaviour while working in very different ways, thus “gaming” behavioural tests” (18).

<sup>11</sup>In §4, we will consider a strategy in the vicinity of this one, which more specifically presupposes the truth of computational functionalism, which we define later.

To start, 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.

Moreover, even if one relaxes the usual neuroscientific assumption and focuses on some comparatively abstract computational correlate of neuroscientific functional architecture, the prospects are equally poor for an argument by analogy for AIC.

An initial problem stems from the opacity of AI systems. Even for a simple feed forward neural network with not too many hidden layers, it is difficult to make credible and reliable statements about how each of the features 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 make (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 4o) of parameters, for which the estimated functions are massively non-linear, highly non-compressible, and largely recursive (Rudin, 2019, 2027), there is very little hope of meaningfully understanding how they reach their predictions. Hence there is no clear path to establishing that AI systems have an abstract functional architecture similar to those of creatures warrantably believed to be conscious.

One might object as follows. There is now a large literature 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). So perhaps XAI will provide the basis for establishing that AI system processing involves a certain internal functional architecture—perhaps even one computationally similar to that in conscious humans and non-human animals. More generally: perhaps the analogical argument for AIC will fare better with the benefit of XAI. Indeed, Goldstein and Levinstein (2024) argue that LLMs

satisfy certain important desiderata requisite for having a mind by drawing on XAI (and on ‘probing’, in particular). Unfortunately, the XAI strategy will not work to establish AIC, as we next argue by paying attention to the two main approaches (counterfactual and surrogate) to XAI.

### 1.5.1 Counterfactual approaches to XAI

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.

But such a counterfactual strategy does not 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 causal/functional architecture is unsurprising. As [Cartwright \(1979\)](#) convincingly argues as regards attempts to extract causal relationships from correlations in interventionist causal models, “no causation in, no causation out”. More generally, 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,<sup>12</sup> any pattern of input-output correlations is going to be compatible with indefinitely many mutually incompatible functional architectures.

### 1.5.2 Surrogate approaches to XAI

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.<sup>13</sup> Again take 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.<sup>14</sup> One can distinguish between local and global surrogate models, where local models replicate the

---

<sup>12</sup>See [Stanford \(2023\)](#). 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.

<sup>13</sup>See [Rudin \(2019\)](#) for discussion.

<sup>14</sup>Probing, as discussed in ([Goldstein and Levinstein, 2024](#)), is an instance of a surrogate approach, as the idea in probing 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](#)).

result for Henry only, and global models replicate the result for all samples. 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.

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.<sup>15</sup> In any case, the main point for our purpose here is that post-hoc strategies are useless for purposes of identifying the internal functional architecture of a given AI system. 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 it is clear that global surrogate models are either inaccurate representations, or they are themselves opaque (there is, necessarily, no middle ground). Meanwhile, local surrogate models need not be opaque, but they offer “explanations that are not faithful to what the original model computes” (Rudin, 2019, 207). In simple terms: local surrogate models are just different models altogether.

Consider an everyday example. Alice, a student, is late to class. I want to know why Alice is late to class. I 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 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,b). 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.<sup>16</sup>

---

<sup>15</sup>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.

<sup>16</sup>One other ‘white box’ (so-called ‘inherent interpretability’) strategy has been discussed in contexts where

To sum up, neuroscientific claims about the functional internal architecture of conscious humans and non-human animals presuppose that this architecture is implemented in a neurobiological substrate, in which case no argument by analogy based on similarity in this respect will establish that there is AIC. Even bracketing the commitment to neurobiological substrate, no such argument by analogy seems available. In particular: efforts to explain or interpret AI system processing are either (a) limited to tweaking inputs and outputs in a way that sheds no understanding on their internal architecture, or (b) involve the post-hoc identification of functional architectures which are compatible with, but drastically underdetermined by, the data. We conclude that there is no realistic prospect of establishing that there is AIC via analogical considerations based on internal functional architecture.

## 1.6 Absence of Consciousness' Evidence is Not Evidence of its Absence

We have considered the spectrum of respects of similarity operative in arguments by analogy offered in response to the problem of other minds, and argued that none of these respects of similarity are in place 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. The analogical response to this problem builds on this access, in ways plausibly supporting positive attributions of consciousness to certain other entities, on grounds that the targets have features similar to those empirically associated with our own conscious states. But this line of positive support does not entail that the failure of some other entity to have some or even any of these features suffices to establish that the target entity is *not* conscious. To draw such a conclusion would appear to involve an inappropriately anthropocentric overgeneralization: why suppose that only creatures relevantly similar to ourselves might be conscious? Indeed, philosophical discussions of mentality standardly assume that conscious mental states are

---

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.

multiply realizable.

Second, 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. Consequently, notwithstanding that analogical considerations 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 the negative analogical finding. Hence from the failure of analogical considerations to warrantably establish that AI systems are conscious it does not follow that we are warranted in taking AI systems *not* to be conscious. Consequently, the analogical strategy does not provide a route to warrantably establishing either that, or whether, there is AIC.

## 2 Inference to the Best Explanation (IBE)

A second 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 \(2022\)](#). Such an IBE would start 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.

The IBE strategy is related to the analogical strategy, since some facts to be explained might advert to respects of similarity between the target entity and entities justifiably believed to be conscious. But the IBE strategy doesn't *require* that respects of similarity be in place; other facts or data can be taken into account. When other humans are the target, these facts may include their self-reports of consciousness, and their treatment of other humans as conscious beings.

The IBE strategy is promising as regards providing a warranted basis for the existence of other conscious humans and non-human animals. Plausibly, 'consciousness-involving explanations' on which other humans and certain non-human animals are conscious will fare

better than alternative explanations.<sup>17</sup> In the case of other humans, alternative explanations might appeal to solipsism, delusion, or skepticism; but these each invoke substantive theoretical costs not invoked 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 an organism, and specifically, in failing to accommodate well-entrenched scientific theories or claims, including that there are specific neural correlates 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 others. It might also be that a consciousness-involving explanation is more parsimonious than the alternative explanations.

Might the IBE strategy work for AI systems? Since relevant internal and external similarities between humans and AI systems aren’t in place, the facts to be explained will presumably be along the lines, e.g., that AI systems can simulate human conversations, that they are treated by some humans as conscious, and that they respond ‘yes’ when asked whether they are sentient.<sup>18</sup> That’s not much to go on, but in any case 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 previous facts plausibly runs as follows: that AI systems can mimic human dialogue, including sometimes responding ‘yes’ when asked if they are conscious, reflects that AI systems are nonlinear probabilistic prediction models 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 instinct to

---

<sup>17</sup>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 \(1967\)](#)).

<sup>18</sup>We put aside that this fact isn’t especially stable; as [Udell and Schwitzgebel \(2021\)](#) note, AI systems will also respond ‘yes’ when asked whether they are not sentient.

anthropomorphize things that mimic us in linguistic respects.<sup>19</sup> Alternative explanations of these facts which 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.<sup>20</sup> Such alternative explanations will also be less theoretically consilient, since we have no antecedent or independent scientific motivation for taking non-organic computer systems (or associated software programs) to be conscious, again by way of contrast with a non-consciousness-involving explanation of the facts at issue. We could hammer more nails into the coffin, but these considerations already make clear that there's no realistic hope of an IBE-based argument for AIC.

Supporting this judgment 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.<sup>21</sup>

Can the previous considerations show that AI systems are *not* conscious, and so provide a route to establishing whether, if not that, there is AIC? The answer is again no, for reasons having to do both with the nature of IBE and the introspective starting point of the problem of other minds.

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.<sup>22</sup> 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 identi-

---

<sup>19</sup>See, e.g., [Epley et al. \(2007\)](#) and [Urquiza-Haas and Kotrschal \(2015\)](#).

<sup>20</sup>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.

<sup>21</sup>See [Metz \(2022\)](#) for a relevant survey.

<sup>22</sup>See [Biggs and Wilson \(2017\)](#) for discussion.

fying 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. The upshot is that the IBE strategy, like the analogical strategy, does not provide a route to warrantably establishing either that, or whether, there is 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. Relatedly, 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 consciousness-involving source or cause. And as Schneider (2019, 43) 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 (Schneider, 2019, 46).

Hence, contrary to Turing’s original application, the test is not useful for any machine which has been purposely designed to linguistically interact with humans. Though 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, or whether, 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 argument, the problem remains that such further behavioural similarities will be designed and 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” (Harnad, 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 closer to the IBE than the analogical strategy, 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, 2021, 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 (2021), 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 consistent, 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.

To give an example of just how deeply rooted references to consciousness are in our language, consider the ‘elementary’ question that Turner and Schneider 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.

Here the supposition is that no terms in this question are ‘consciousness-implicating’. But at least three terms in this question appear to be 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. Any such characterization of the physical thus implicates consciousness.<sup>23</sup>

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 and found that three of its core terms are consciousness-implicating. As such, it’s hard to see how any model could be fit to a large body of data while simultaneously satisfying the ACT requirement. Out the window must go references to anything involving qualitative experience, including perceived colors and their relations, art, music, food, fashion, architecture—i.e., most human culture. Out the window must go 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

---

<sup>23</sup>See [Montero \(1999\)](#), [Wilson \(2006\)](#), [Alter \(2024\)](#).

topics and primary drivers to action. Out the window must go references to anesthesiology, ophthalmology, psychiatry, psychology, child development, etc. Out the window must go all the sciences which take 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. Out the window goes all of 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.

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 analogy- or IBE-based strategies, or else face insuperable difficulties of implementation. Hence, these strategies do not provide a route to warrantably establishing either that, or whether 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 \(2021\)](#) 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 such theory is correct, and theory-heavy strategies have repeatedly faced this roadblock in their attempts to extend any particular theory (e.g., global workspace theory) to the study of AI consciousness in particular. This difficulty inspires a ‘theory-light’ strategy,<sup>24</sup> which considers, for some range of candidate theories of consciousness, whether AI systems meet the proposed necessary and/or sufficient (typically functional) conditions specified by the theory. [Butlin et al. \(2023\)](#), for instance, derives “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 approach, to judge whether an existing or proposed

---

<sup>24</sup>Here we extend Birch’s (2020) terminology. Birch’s application of a theory-light approach is directed at insects and other “biological taxa”; much of what he says about the empirical evidence for neuroscientific theories is consonant with what we say below.

AI system is a serious candidate for consciousness, one should assess whether it has or would have these properties.<sup>25</sup>

Might this strategy work as a means of warrantedly establishing that, or whether, there is AIC? One reason to think not is that, as we previously argued, the prospects are dim for determining whether AI systems implement the functions at issue. But 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.

We will argue that the empirical support for neuroscientific theories is restricted to their application to humans or other entities with neurobiological substrates. Since AI systems are not similar to humans in these respects, none of these theories can be used to show that there is AIC. Conversely, since these neuroscientific theories are not properly seen as offering general necessary conditions on consciousness, the failure of any or all such theories to apply to AI systems does not entail that AI systems are not conscious. Hence a ‘theory light’ strategy fails to warrantedly establish that, or whether, there is AIC.

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 adverts to direct introspective access in individual humans (notably: the neuroscientists testing a given theory!), which by analogical or IBE-based means is extended to test subjects making various reports of (or justifiably taken to have) consciousness. It is introspectively grounded reports which provide us with the crucial evidence that this or that functional architecture is associated with consciousness, which evidence can then, given that appropriate neurobiological substrates and other similarities are in place, provide a basis for warrantedly taking other humans and certain non-human animals to be conscious. 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

---

<sup>25</sup>See also [Chalmers \(2023\)](#).

to these.<sup>26</sup>

Relatedly, the evidence for the theories that [Butlin et al. \(2023\)](#) discuss is uniformly drawn from experiments on humans or non-human animals. In these and other neuroscientific theories that [Butlin et al. \(2023\)](#) discuss, claims that a given functional architecture is necessary and/or sufficient for consciousness, and empirical results offered as supporting these claims, are restricted to humans whose consciousness is introspectively and analogically established (perhaps extended to relevantly similar non-human animals).

Nonetheless, [Butlin et al. \(2023\)](#) claim that these neuroscientific theories “aim to describe correlations between computational processes and consciousness”. That’s incorrect, however; as just observed, these theories aim to describe correlations between brain processes and consciousness; and a brain process is not itself a computational process, notwithstanding that the latter might in some sense be instantiated in the former. Now, [Butlin et al. \(2023\)](#) also say 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”; so perhaps the suggestion is that even if claims in and evidence for neuroscientific theories are not cashed in computational terms, one can interpret the theories and evidence in such terms.

There are two problems with this suggestion, however. First, as previously the empirical evidence for these theories crucially relies on our direct introspective access to our own consciousness, since such access provides a basis (via analogical or IBE-based strategies) for taking ‘immediate consciousness reports’ at face value. Hence even if this 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 not like us, is necessary and/or sufficient for consciousness in such entities. From the fact that  $A$  and  $B$  jointly entail  $C$ , it doesn’t follow that  $B$  alone, or  $B$  conjoined with something other than  $A$ , entails  $C$ ; from the fact that  $A$  and  $B$  jointly cause  $C$ , it doesn’t follow that  $B$  alone, or  $B$  coupled with something other than  $A$ , causes  $C$ ; and from the fact that function  $B$  when implemented in some neurobiological process  $A$  is necessary and/or sufficient for consciousness, it doesn’t follow that  $B$  alone, or  $B$  when implemented in some non-neurological process, is necessary and/or sufficient for consciousness. And here again the baseline fact is salient: for even if  $B$  when implemented in some non-neurobiological

---

<sup>26</sup>See [Birch \(2020\)](#) and [Carruthers \(2019\)](#) for similar observations.

process *is* associated with consciousness, we are not now, and plausibly will never be, in position to confirm this fact, for we are not now, and plausibly will never be, in position either to engage in the requisite introspection, or to infer by means of analogy or IBE to the consciousness of the non-neurobiological entity at issue.

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 are irrelevant: 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 false.

This result, by the way, does not presuppose that computational functionalism is false. For all we have said, computational functionalism might be true. Our point, instead, is the weaker one that the only empirical support we have for this or that function’s being relevant to consciousness is one which is inextricably mixed up with the messy, carbon-based, naturally-evolved neurobiological substrate which serves as the basis for our own, introspectively available, conscious mentality. That’s enough to show that it is false that certain neuroscientific theories of human consciousness are properly seen as bearing on whether there is AIC.

## 5 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, 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, AI systems might still be conscious. We just can never know. There is therefore a distinctively difficult ‘problem of other AI minds’, whose lack of principled resolution is likely to stymie attempts to offer any justified answers to questions about the personal, moral, and agential status of AI systems.

Beyond the specific case of AIC, attention to the problem of other AI minds has revealed an underappreciated moral of the problem of other minds: namely, that it is unlikely that, by attention to our own case and any of the associated features thereof, we will ever be in position to arrive at any *necessary* non-introspective conditions for, much less a general theory of, consciousness or its attribution. Indeed, proper appreciation of the baseline fact driving the problem of other minds suggests that ultimately there is, at best, only one necessary (and indirectly circular) criterion of consciousness: that one introspectively has it.

## 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.
- Arvan, M. and C. J. Maley (2022). Panpsychism and AI consciousness. *Synthese* 200(3), 1–22.
- Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.
- 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.

- 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.
- Blackburn, S. (1994). *The Oxford Dictionary of Philosophy*. Oxford: Oxford University Press.
- 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. (2023). Could a large language model be conscious? *Boston Review* 1.
- Chowdhury, N., D. Johnson, V. Huang, J. Steinhardt, and S. Schwettmann (2025, April). Investigating truthfulness in a pre-release o3 model.
- 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*.

- 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. Lombardo (Ed.), *In the Light of Evolution: Volume VII: The Human Mental Machinery*. Washington, D. C.: National Academies Press.
- 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.
- 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.
- Rosenblatt, F. (1957). *The Perceptron, a Perceiving and Recognizing Automaton:(Project Para)*. Cornell Aeronautical Laboratory.
- 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.

- 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.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3), 417–57.
- 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 (2021). 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.