Erkenntnis (2025) 90:1323–1342
https://doi.org/10.1007/s10670-023-00753-9

ORIGINAL RESEARCH

Tests of Animal Consciousness are Tests of Machine Consciousness

Leonard Dung[1,2]

Received: 4 May 2023 / Accepted: 18 October 2023 / Published online: 14 November 2023

Abstract
If a machine attains consciousness, how could we find out? In this paper, I make three related claims regarding positive tests of machine consciousness. All three claims center on the idea that an AI can be constructed "ad hoc", that is, with the purpose of satisfying a particular test of consciousness while clearly not being conscious. First, a proposed test of machine consciousness can be legitimate, even if AI can be constructed ad hoc specifically to pass this test. This is underscored by the observation that many, if not all, putative tests of machine consciousness can be passed by non-conscious machines via ad hoc means. Second, we can identify ad hoc AI by taking inspiration from the notion of an ad hoc hypothesis in philosophy of science. Third, given the first and the second claim, the most reliable tests of animal consciousness turn out to be valid and useful positive tests of machine consciousness as well. If a non-ad hoc AI exhibits clusters of cognitive capacities facilitated by consciousness in humans which can be selectively switched off by masking and if it reproduces human behavior in suitably designed double dissociation tasks, we should treat the AI as conscious.

1 Introduction

A being is (phenomenally) conscious iff there is something it is like to be the being in question (Nagel, 1974). That is, it is conscious iff it has subjective experience. Arguably, whether a being has conscious (affective) experience determines whether it has moral status (Dung, 2022b; Jaworska & Tannenbaum, 2021; Shevlin, 2020b).

✉ Leonard Dung
leonard.dung@fau.de

[1] Centre for Philosophy and AI Research, Universität Erlangen-Nürnberg, Werner-von-Siemens-Str. 61, 91052 Erlangen, DE, Germany

[2] Institute of Philosophy II, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, DE, Germany

In addition, as soon as a being is conscious, there arises the risk that it might suffer. The connection between consciousness and suffering sparked a debate on the risks emerging from research which might eventually enable the creation of conscious machines (Metzinger, 2021).[1] Most theories of consciousness are functionalist and thereby entail that a machine with the right causal organization would be conscious (Dehaene et al., 2017; Doerig et al., 2021; Seth & Bayne, 2022). If we build conscious artificial intelligences (AIs) but don't recognize that they are conscious and thus don't treat them appropriately, then we might inadvertently cause – potentially astronomical – suffering (Dung, 2023; Saad & Bradley, 2022; Tomasik, 2014).

Consequently, there arises a demand for tests of machine consciousness. If machines develop consciousness, then we need to be able to detect this. Such tests capitalize on a certain behavior, cognitive capacity or other feature some machines might possess. This behavior, capacity or feature is taken to be an indicator of consciousness (or its absence), at least in machines. Most strongly, the idea might be that the indicator (or its absence) guarantees the presence (or absence) of consciousness. More moderately, the presence (or absence) of the indicator might give us good grounds for believing that the machine in question is conscious (or not conscious). Importantly, the evidential strength of consciousness indicators is frequently asymmetrical. The presence of an indicator may be strong evidence of consciousness while its absence does not speak much against attributions of consciousness and vice versa.

This paper is about *positive* indicators of consciousness, i.e., tests for the presence of consciousness, not its absence. Its aim is to show that some proposed tests of animal consciousness can also serve as tests of AI consciousness. In the next section, I will motivate the view that most tests of animal consciousness which have been proposed in the literature fail as tests of AI consciousness. I also express skepticism regarding other tests of AI consciousness. I suspect that most of them are subject to the problem that AIs can be contrived specifically to exhibit the proposed indicator while lacking the general organization and cognitive complexity which would qualify them as a serious consciousness candidate. In Sect. 3, I propose amending tests of machine consciousness with an additional constraint: To indicate consciousness, those tests must be passed in a way which is not ad hoc. After spelling out this constraint by relating it to the notion of ad hoc hypothesis in science, Sects. 5, 6 and 7 illustrate concrete applications of this methodology. The result will be that, given this constraint, the best tests of animal consciousness serve as useful positive tests of machine consciousness. Section 8 concludes.

2 Animal Consciousness Tests and the Gaming Problem

Before discussing the flaw of currently entertained proposals of machine consciousness, I will first mention two further desiderata for an indicator of artificial consciousness. First, an indicator is better (*ceteris paribus*) if it is more neutral between different theories of the nature, function or physical substrate of consciousness (Udell,

[1] I employ the terms 'machine' and 'AI' in a fairly broad sense which includes robots, merely virtual beings and computational models used, for instance, in consciousness research.

2021).[2] This is desirable because there is widespread disagreement on which theory of human consciousness is true. Due to deep metaphysical (Schwitzgebel, 2020) and methodological (Irvine, 2012) differences, it doesn't seem like a consensus is within reach. Moreover, a theory which correctly describes the mechanisms underlying consciousness in humans and other animals might not apply to machines. Hence, a test of machine consciousness which presupposes that a contentious theoretical view of consciousness is true runs the high risk that this presupposition is false or cannot be extended to non-living beings.

Second, a positive indicator for the presence of consciousness is better (*ceteris paribus*) if it is less demanding and thus applies to more kinds of machines. For this makes the indicator more informative. For instance, one may claim that a machine is probably conscious if it demonstrates behavior mirroring humans in all domains, possesses all cognitive skills humans have to a high degree and has a cognitive architecture and physical implementation sufficient for consciousness according to every influential theory of human consciousness. While this seems plausible, such an indicator is not particularly useful since few machines will satisfy it such that it tells us little and since it does not advance our intuitive practices of ascribing consciousness much.

What could be a good test of machine consciousness? Prima facie, it is not obvious that we need tests specific to machine consciousness. For the science of consciousness has developed many putative measures of consciousness in non-human animals (Birch, 2022; Birch et al., 2020; Dung, 2022a; Ginsburg & Jablonka, 2019; Sneddon et al., 2014; Tye, 2017). Animal and machine consciousness research share similar challenges. In both cases, we have to extend knowledge from human consciousness, which is typically the methodological starting point of consciousness research (Dung & Newen, 2023), to non-human subjects which either are not able to provide verbal reports of their (alleged) experiences, or whose reports we cannot take at face value. Thus, a natural first step when exploring AI consciousness is to think about whether measures of animal consciousness can be applied to machines as well.

However, there are some problems. Applying these tests to AI systems is not straightforward. Obviously, indicators derived from comparative neuroscience cannot be applied to machines because they lack a nervous system. Yet, as noted by Shevlin (2020a), cognitive-behavioral indicators of consciousness rely implicitly on neuro-physiological similarities of humans and animals as well. For the usual route to collecting evidence of consciousness is to find a behavior which is enabled by consciousness in humans and then to search for the same type of behavior in other animals (Tye, 2017). However, that animals display behavior which is caused by consciousness in humans is only credible evidence of animal consciousness if it is plausible that the same processes lead up to the behavior in humans and animals. Due to neuro-physiological similarity and shared evolutionary history, this condition may be fulfilled in respect to many animals but not to machines. For this reason, repurpos-

[2] While a test of machine consciousness should not presuppose a specific view on the function of consciousness, the entire project of finding tests of consciousness does presuppose that consciousness makes some functional difference, i.e., has systematic causal effects.

ing indicators of animal consciousness to machine consciousness is commonly seen to not be fruitful.[3]

While I later argue that some indicators of animal consciousness can serve as reliable indicators of machine consciousness, I will accept the preceding reasoning for now. There is a further argument against using tests of animal consciousness as tests of machine consciousness. Trace conditioning, a form of classical conditioning where the two stimuli are separated by a temporal interval, is sometimes proposed as a potential indicator of animal consciousness since some form of trace conditioning seems to require consciousness in humans (Clark & Squire, 1998, 1999).[4] However, one might plausibly build a simplistic "gerrymandered" (Shevlin, 2020a) AI which is specifically build to be able to do trace conditioning and not much else. The presuppositions for having trace conditioning are the ability to immediately react to certain inputs, to learn simple associative rules connecting them and to briefly store relevant information. Those seem trivial and can be captured in simple If-then rules. Therefore, there is no reason to think that an AI which exhibits trace conditioning needs to be conscious.

In general, when we use a property which is taken to indicate animal consciousness as indicator of machine consciousness "we can game the system by constructing degenerate examples of systems exhibiting that property that we don't intuitively think of as sentient" (Tomasik, 2014). If it is easy to build systems which possess an alleged indicator but are not conscious, then the indicator seems to be flawed. Call this the 'gaming problem' (Birch & Andrews, 2023).

As a first approximation, the gaming problem is that many putative tests of consciousness can be gamed by the designers of AI systems. Designers can build systems with the goal of passing the test, even if the system has none of the capacities the test was thought to track and which may give rise to consciousness. In this case, the AI is 'gerrymandered' with respect to the test. We will later extend this definition of the gaming problem.

I hypothesize that indicators derived from animal consciousness research are generally vulnerable to the gaming problem. Moreover, my suspicion is that most other positive tests of machine consciousness which fare well with respect to our two desiderata succumb to the gaming problem as well.

The general reason is this: If the test of consciousness under consideration is based on a relatively superficial property closely tied to specific behavioral capacities, there are often many ways to enable the same capacity. With sufficient ingenuity, a designer will often be able to find a way to reproduce the capacity without the complex mechanisms which could plausibly be sufficient for conscious experience. However, when taking specific deep structural features as the basis for a consciousness test, a test risks violating the theory-neutrality desideratum, since different theories of consciousness focus on very different features. [5]

[3] Among others, this is evidenced by the fact that the literature on tests for machine consciousness developed in separation from the literature on animal consciousness.

[4] For critical discussion, see Droege et al. (2021) and Mason and Lavery (2022).

[5] For a recent review of proposed tests for machine consciousness, see Elamrani and Yampolskiy (2019). Notice that many tests they review fail the desideratum of theory-neutrality.

Since I believe that theory-neutrality is valuable, I will argue for relatively superficial, behavior-based tests of AI consciousness in this paper. This requires tackling the gaming problem head-on which is the task of the next two sections.

3 The Naturalness Constraint

In this section, I argue that tests of AI consciousness should include rules for which types of machines are allowed to take the test. AI systems which are gerrymandered to pass a particular test do not count as proper subjects (of this test). Thus, tests of consciousness can be valid even if machines *would* pass them due to gaming, as long as the tests indicate consciousness in non-gerrymandered machines. To give this claim content, we need ways to distinguish irrelevant gerrymandered AIs from legitimate counterexamples to the claim that consciousness is necessary to pass a particular consciousness test.

The guiding idea is this: gerrymandered AIs can be understood in analogy to ad-hoc hypotheses in science. Consequently, the criteria for identifying ad hoc hypotheses can inform our understanding of gerrymandered AI. Afterwards, I will examine what good indicators of machine consciousness look like, given that we allow that gerrymandered AIs may possess them. My conclusion will be that the strongest tests of animal consciousness can serve as good tests of machine consciousness as well.

As we have seen, the usual paradigm for positive tests of machine consciousness has the following form: Find a feature which *in general*, if a machine possesses it, provides good grounds for taking this machine to be conscious. Notably, the indicator is supposed to provide good evidence irrespective of the context in which it occurs. Since the machine itself is part of the context, this entails that the indicator should provide good evidence independently of the machine which exhibits it. I reject this paradigm. Instead, I hold that an indicator only provides good grounds for believing in consciousness if the AI was not constructed specifically to exhibit this indicator. I call this an 'ad hoc' construction. In what follows, I will use the term 'ad hoc AI' synonymous with 'gerrymandered AI' and 'AI resulting from gaming'.

Given this view, having an indicator is not sufficient for justifying attributions of machine consciousness. Instead, there are two conditions: We have good grounds for believing that a machine is conscious if (i) it exhibits a valid indicator of machine consciousness and (ii) the indicator was not produced via an ad hoc construction. It follows that a test for machine consciousness can be legitimate, even if there are gerrymandered (ad hoc) AIs which would pass the test and are not conscious.

I will call machines which are not constructed ad hoc (in relation to a certain feature) 'natural'. The preceding constraint which excludes ad hoc constructions from having evidential significance will be called the 'naturalness constraint'.

Naturalness constraint: Passing a test of machine consciousness only provides good grounds for believing that the machine is consciousness, if the machine is natural with respect to that test.

The naturalness constraint is motivated by an analogy to scientific theorizing and the notion of an 'ad hoc hypothesis'. As a first approximation, a hypothesis is ad hoc iff it is constructed solely to accommodate a particular piece of empirical evidence

and has no independent support. According to the dominant view in philosophy of science called 'predictivism', a hypothesis is more strongly supported by evidence if it predicts it rather than accommodating it in an ad hoc fashion (Barnes, 2021). In analogy, I submit that the postulation of machine consciousness is only sufficiently supported by passing a consciousness test if the machine was not constructed ad hoc to pass this test.

In what follows, I will take inspiration from this analogy between ad hoc hypotheses and ad hoc AI to further flesh out the view that indicators only need to provide good grounds for believing in machine consciousness if the AI does not accommodate them ad hoc. This requires an elucidation of when AIs count as producing indicators ad hoc and of why ad hoc accommodation does not provide good grounds for belief in machine consciousness.

Intuitively, an AI is designed ad hoc with respect to a particular consciousness test when it is specifically designed to fit the test. While I will make this characterization more precise in the next section, I will first explain why ad hoc accommodation does not provide good grounds for belief in machine consciousness.

There is a wide range of authors suggesting that predictions of novel evidence have more evidential weight than accommodations of evidence achieved by modifying a scientific theory or introducing a new theory (Barnes, 2021; Hitchcock & Sober, 2004; Howson, 1988; Lipton, 1990; Maher, 1993; White, 2003). This may be the case either because there is something intrinsic to predictions that makes them superior to accommodations (strong predictivism) or because predictions are reliably correlated with other features which increase the probability that the theory making the prediction is true (weak predictivism). In addition, the history of science seems to bear out the suggestion that predictions of novel facts raise the confidence in a scientific theory typically more than accommodation of known observations.[6] One can ask where this epistemic advantage of predictions over accommodations stems from. I wish to remain neutral on this further question.

In any case, that predictions confer an epistemic advantage over accommodations is the received view for science, although the reason for this advantage is contested. I hold that the same is true in the case of AI consciousness. If an AI is constructed with the aim of passing a certain test of consciousness, it is not surprising that it succeeds. We can explain this success by appealing to the fact that it was designed to pass the test. By contrast, it is surprising when an AI passes a test for consciousness even if it was not created with the purpose of excelling at this test. In this case, that the AI passes the test cannot be explained in recourse to the aim of making it able to do so. Comparatively, this makes the competing explanation more likely that the AI passes the test because it is conscious.[7]

In this section, I have argued that it is possible to delineate the class of machines that are constructed ad hoc and that the possession of an indicator which was pro-

[6] See Sober (2015, p. 17) for an account according to which the superiority of Copernican to Ptolemaic cosmology was crucially due to the fact that the former predicted many regularities that the latter could only accommodate. By contrast, Brush (1994) argues that, in many historical episodes, scientists didn't give more weight to novel predictions than to accommodations of known facts, at least if 'novelty' is understood in a purely temporal sense.

[7] For an analogous argument in the context of scientific theories, see White (2003).

duced by ad hoc construction possesses less evidential weight than a natural construction. In the preceding section, I suggested that tests of machine consciousness cannot rule out that machines can be constructed ad hoc to pass the test, despite not being conscious. In conjunction, this suggests that we should allow that valid tests of machine consciousness can be passed by non-conscious AIs via ad hoc means. So, we can be confident that a machine is conscious if it passes an adequate test of machine consciousness *and* satisfies an independent constraint of not being designed specifically to pass this test.

In the next section, I apply these ideas. I clarify what it means for an AI to be ad hoc and explain how we can evaluate whether an AI obeys the naturalness constraint. Subsequently, based on the insight that tests of AI consciousness do not need to prevent ad hoc solutions, I propose two tests of machine consciousness which are taken directly from the literature on indicators of animal consciousness. I claim that, given that ad hoc AI cannot undermine tests of AI consciousness, the most reliable tests of animal consciousness are valuable tests of machine consciousness as well.

4 What is ad hoc AI?

In this section, we will explore when a machine is an ad hoc construction. Let's consider our analogy to scientific theorizing once again. According to Zahar (1973), a fact is novel (and thus predicted, rather than accommodated) "if it did not belong to the problem-situation which governed the construction of the hypothesis". However, the context of known facts during the construction of the hypothesis does not seem intrinsically relevant for the evidential value of a prediction (Gardner, 1982). Instead, what matters is whether the theory was built to fit particular facts, i.e., whether knowledge of these facts was used during the construction of the theory or not.[8] Applied to our topic, we can say that an AI is natural with respect to a consciousness test if it wasn't built to fit this test, i.e. if knowledge about the test did not influence the design of the AI.[9]

While this characterization is a useful starting point, it does not suffice to guide attributions of naturalness to specific, concrete AI system. How do we determine whether a specific AI is built to fit a particular consciousness test? I will not provide an exact analysis as a list of precise necessary and sufficient conditions for a

[8] Arguably, one might even go further and claim that what matters is not the actual process of theory-construction, but whether the theory could have been constructed without knowing the relevant facts (Worrall, 1989). Applied to machine consciousness, this amounts to asking whether one could have identified something as a promising consciousness candidate without appealing to the test for machine consciousness. In the case of ad hoc AIs, so the thought, this would not have been possible.

[9] In the philosophy of science literature, there is a distinction between saying that a hypothesis is ad hoc and saying that it accommodates evidence (rather than predicting novel evidence). Most notably, calling a hypothesis "ad hoc" implies a negative value judgement while accommodations of evidence, even if they lack evidential value, are legitimate. For an accommodation to be problematic and thus ad hoc, further criteria have to be satisfied (Leplin, 1975). I slide over this distinction here because, when discussing machine consciousness, the relevant accommodation is one where an AI is built solely to pass a specific test of machine consciousness and no other means of ascertaining consciousness are available. This form of accommodation qualifies for the negative verdict of being ad hoc.

machine's being ad hoc. As argued by recent as well as historical authors (Margolis & Laurence, 2003; Wittgenstein, 1953), I do not think precise definitions of this sort are typically available. Moreover, ad-hoc-ness is a rich and possibly context-dependent notion, such that the prospects of a proper analysis are dim.

However, in general and in this particular case, the absence of an explicit definition does not prevent a concept from being useful. While we will not be able to specify exact, unambiguous criteria for when an AI counts as ad hoc, the analogy to science inspires optimism. There are no exact criteria for classifying a scientific hypothesis as ad hoc either. Nevertheless, in practice, scientists often reach a consensus that a particular hypothesis has ad hoc character. Analogously, I take it, experts on AI have sufficient implicit knowledge of the criteria for an AI's being ad hoc to fruitfully apply this concept.

In an effort to make some of this implicit conceptual knowledge explicit, I will mention some features which tend to indicate that an AI is designed ad hoc (with respect to a particular test). First, in the paradigmatic case, an AI was built with the explicit intention to pass a particular test of consciousness. Thus, its design is such that it is particularly well-equipped to pass the specific test.

Note, however, that the mere intention to build the AI such that it passes the test is neither necessary nor sufficient for ad-hoc-ness. It is not sufficient because what ultimately matters for naturalness is the relation between the features of the AI and the particular test in question, not the psychology of the designer. Arguably, naturalness supervenes on the combination of the intrinsic properties of the AI, the consciousness test in question and the relation between them. If so, when two AIs are identical in functionality and physical structure, they are equally natural (with respect to a particular test), even if their designers had different reasons for building them.

Similarly, an AI can end up narrowly tailored to a specific consciousness test, even if that was not the intention of its designer. In principle, this could happen due to random chance, but it will normally have more systematic reasons. For instance, an AI might be trained to do a task which very well correlates with a particular consciousness test. A further limitation is that the intentions of the designer of an AI are not always accessible to us retrospectively. Hence, complementarily, we have to look out for other features of naturalness.

A further sign that a machine is constructed in an ad hoc manner is that the machine would not be a candidate for attributions of consciousness worthy of consideration, if it did not pass the test in question. That is, with ad hoc AI, there often is not even a reason to think they might be conscious or should be subjected to a consciousness test, if they were not designed such that they might pass a particular test. Ad hoc AIs tend to lack independent evidence or an independent rationale for attributions of consciousness.

By contrast, our confidence that a machine is natural should increase when it is designed to satisfy a particular theory of the mechanisms underlying consciousness (such as global-workspace or integrated-information theory).[10] For if the design of an AI is motivated by a theory of consciousness, then the AI is not designed to possess any specific behavioral capacity or pass a particular behavior-based consciousness

[10] I owe this observation to Wanja Wiese.

test. Moreover, if an AI aims to emulate the computational mechanisms (Wiese & Friston, 2021) which might underlie consciousness and as a result exhibits behavioral capacities which seem like a plausible test of consciousness, then convergent evidence for attributions of consciousness is beginning to accumulate.

There are further signs of illegitimate ad hoc construction. In particular, ad hoc AIs tend to have features specifically relevant for solving a particular test, but which do not seem to have a larger relevance for consciousness. First, this might mean that those machines are equipped with some specific behavioral programs or pieces of innate knowledge which help to pass the test. For instance, in language-based tests of consciousness, the AI might be pre-equipped with the tendency to produce certain statements which the test deems evidence of consciousness. Second, machines might be trained (via machine learning techniques) to optimize performance at a particular test, or a task closely related to it. For instance, producing the right kind of answers in a language-based test of consciousness might serve as the goal during training.

While not being sufficient, there are some heuristically relevant signs that a machine may be natural. First, the capacity of an AI to pass a test may be *emergent* in the sense that it was surprising to its designers and could not have been predicted in advance. This points to naturalness. Second, the AI may have been built before the test of consciousness existed. This rules out that the AI was designed with the explicit intention to fit the test. Note, however, that we discovered earlier that explicit intentions of AI designers to game a test are not necessary for ad-hoc-ness.

In practice, as with scientific theories, there may be borderline cases when we have to determine whether hard-wired knowledge or training procedures are so specific and so intimately related to a particular test that they count as ad hoc solutions of it.[11] In those cases, we have to rely on experts' implicit knowledge of when a machine is natural and when not. Moreover, we can say that the degree of confirmation conferred by the possession of an indicator of consciousness depends on the degree of naturalness with which the AI produced the indicator. The more natural an AI is, the more strongly supports passing the test attributions of consciousness.

With these clarifications, we gain a toolkit for evaluating the naturalness of various AI systems with respect to particular consciousness tests. Then, using the naturalness constraint, we can use tests of consciousness even if they can be passed via ad hoc means.

5 The AI Consciousness Test

Before setting out to motivate my own proposal for tests of machine consciousness, I will first discuss the AI Consciousness Test (ACT) suggested by Schneider (2019). Considering this suggestion is instructive, since Schneider – while she does not pro-

[11] While specialization constitutes a hint of ad hoc construction, it is possible that a system is designed such that it possesses a single mechanism carefully selected and tailored such that it enables passing different tests of consciousness. Thus, an AI can be specialized, e.g. trained, to pass multiple different tests of consciousness, and so still count as ad hoc.

pose a naturalness constraint – implicitly grapples with the same issue of ad hoc constructions and looks for ways to exclude them from consideration.

In the ACT, the AI is challenged "with a series of increasingly demanding natural language interactions to see how readily it can grasp and use concepts based on the internal experience we associate with consciousness" (Schneider, 2019, p. 51). That is, we ask the AI about scenarios such as reincarnation or out-of-body experience and mention philosophical issues such as the hard problem of consciousness. We might even see whether the machine comes up with consciousness-based concepts and utters intuitions about the subjectivity of consciousness by itself, without human prompts. If the conversational skills of the AI reveal an understanding of consciousness, then it is deemed conscious.[12]

The ACT presupposes that an intelligent machine without consciousness lacks the concepts to properly describe subjective experience, i.e., that fluently and appropriately answering questions regarding consciousness requires introspective familiarity with consciousness. In a sense, this assumption can easily be refuted. For if the training data of the AI are not restricted, then many non-conscious machines could easily pass the ACT by learning from human conversations about consciousness what the appropriate, human-like responses to questions about consciousness are. For this reason, Schneider requires that the AI be severed from the internet and prevented from gaining too much knowledge about consciousness and neuroscience. Hence, the ACT is only a valid test of consciousness if the learning opportunities of the test subject are constrained.

The constraint on learning Schneider proposes for machines undergoing the ACT can partially be regarded as an instance of the general naturalness constraint. If an AI were trained on explicit information about human responses to questions about consciousness or if explicit rules for responding to such questions were hard-coded, then the AI would be ad hoc in respect to the ACT. The AI would be directly trained to, among many other things, give correct answers to questions about consciousness. Admittedly, the constraint proposed by Schneider is not equivalent to (a specific form of) the naturalness constraint. For Schneider needs to restrict the learning opportunities of the AI not just to data directly relevant to the ACT, but more severely. Much knowledge about the world and the mind could – by an advanced AI – potentially be used for inferring claims about consciousness without being conscious.

At the same time, the AI needs to have sufficient access to a broad range of training data to have a fair chance of attaining intelligence, language skills and perhaps consciousness. At a minimum, the AI needs to be allowed to learn so much that it can interpret the questions that are posed to it and has some knowledge base to resort to when answering them. Thus, the ACT needs to aim for an "epistemic sweet spot" (Udell, 2021) where the AI is allowed to learn enough about consciousness-relevant features non-introspectively that it can interpret the questions of the ACT and communicate about the topic verbally, but not so much that it can infer appropriate answers to the questions (without having to rely on introspection of its own conscious experience). It is not clear whether such a balance can actually be achieved

[12] This test is asymmetrical, like all consciousness tests considered in this paper. A machine may be conscious even if it fails the ACT, for instance because it lacks linguistic skills.

and whether we could know that we achieved it, if we had done so. This constitutes a difficulty for the ACT.

In light of this deficit, I will now unveil my proposal for two suitable tests of machine consciousness. Fortunately, they don't require us to pick an epistemic sweet spot, but only to exclude ad hoc AIs.

6 The Theory-light Strategy

Earlier, we have rejected the suggestion that tests of animal consciousness can also serve as tests of machine consciousness. I hold that this rejection was premature. It stemmed from two sources: First, it underestimates how powerful some proposed tests of animal consciousness which have been conceptualized in the literature may be. Second, animal consciousness tests can be passed in an ad hoc manner. However, given the naturalness constraint, this does not prevent these tests from being valid tests of AI consciousness.

In short, my proposal is that the strongest behavioral tests of animal consciousness can be applied to machines as well, once we have established that it is not problematic that they could also be passed in an ad hoc manner. In contrast to the ACT, the following tests do not rely on language processing or other high-level mental capacities (e.g., mental simulation (Halina, 2021). This is an asset since it makes these tests potentially more widely applicable, even to less sophisticated AIs. On the other hand, these tests presuppose sensory processing or a functionally analogous capacity. Thus, the AI needs to be either embodied and equipped with sensors or accept proxy inputs which can functionally replace actual sensory input. While the technical details relevant to the application of these tests need to be left for other occasions, the rest of the paper does outline a viable strategy for approaching investigations of AI consciousness.

The first test of machine consciousness I have in mind was put forward by Birch (2022) as the 'theory-light' strategy for investigating animal consciousness. The theory-light approach adopts a minimal theoretical commitment, the *facilitation hypothesis*: "*Phenomenally conscious perception of a stimulus facilitates, relative to unconscious perception, a cluster of cognitive abilities in relation to that stimulus*" (ibd.). This hypothesis comprises two claims: First, it assumes that consciousness causally contributes to cognition. Second, it assumes that consciousness facilitates not just one cognitive capacity but several which cluster together in the sense that their presence and absence covaries and correlates mutually and with whether the stimulus is consciously perceived.

The facilitation hypothesis satisfies the theory-neutrality desideratum. The causal efficacy of consciousness is a legitimate assumption since it may even be a presupposition for the empirical study of consciousness. From the third-person stance, we can only access consciousness in virtue of its effects. The assumption that consciousness facilitates different types of cognitive capacities which cluster together

is independently plausible and consistent with every prominent scientific theory of consciousness.[13]

Birch's proposed methodology for animal consciousness research proceeds in several steps. Initially, we have to identify cognitive capacities which are plausibly facilitated by consciousness. Those we can ascertain through research on human consciousness, namely by looking for capacities which are caused by consciousness in humans. However, Birch allows that each capacity might sometimes occur without consciousness, as long as it is *facilitated* when it occurs alongside consciousness. This facilitation might be indicated, for example, by an increase in speed or reliability. Since we need to look for clusters, Birch's methodology demands to identify several of such cognitive capacities. Birch's three examples of such capacities are trace conditioning, cross-modal learning and rapid reversal learning. Those are all forms of learning which demand more cognitive sophistication than standard associative learning and have been previously linked to consciousness in humans (Bellebaum & Daum, 2004; Clark & Squire, 1998, 1999; Mudrik et al., 2014; Palmer & Ramsey, 2012; Travers et al., 2018).[14] Since the identity of the specific cognitive capacities doesn't matter for illustrating the general approach to testing consciousness, let's call them C_1, C_2 and C_3.

Next, we look for this cluster – C_1, C_2 and C_3 – in the animal in question. Finally, we need to choose experimental protocols which can selectively switch consciousness of a stimulus on or off. These methods have been tested and refined in human consciousness research. Backward masking is a technique which is ideal for this purpose. In backward masking, a second stimulus (the 'mask') is presented shortly after presenting the first stimulus. Because of this, the first stimulus does not reach consciousness (although it would have been seen consciously, if the second stimulus hadn't occurred).

If the animal in question is conscious, we would predict that a significant share of the cluster is selectively switched on or off by masking (and that capacities which aren't facilitated by consciousness are not sensitive to masking). In other words, we would expect the animal to be able to perform some of C_1, C_2 and C_3 in respect to unmasked, but not to masked stimuli (or to perform them faster or more successfully in respect to unmasked stimuli). This prediction relies on the assumptions that the masking procedure effectively suppresses consciousness and that C_1, C_2 and C_3 are indeed facilitated by consciousness. Thus, the correctness of the prediction supports both assumptions simultaneously.

[13] According to anti-functionalist theories of consciousness like the integrated-information theory (IIT), the causal role of conscious experience is not essential to it. Nevertheless, IIT allows (and implies) that consciousness has (causal) effects.

[14] In associative learning, an animal learns to associate distinct kinds of stimuli. For instance, in classical conditioning, the animal learns to respond to a neutral (conditioned) stimulus in the same way as to an evolutionarily potent (unconditioned) stimulus. This conditioning effect is achieved by letting both types of stimuli cooccur repeatedly. In trace conditioning, the stimuli are not presented simultaneously, but separated by a temporal interval. In cross-modal learning, two stimuli from different sensory modalities (e.g., vision and audition) are associated. In rapid reversal learning, the animal quickly learns when relations between two stimuli, which have been learned, are reversed.

A pattern of dependence between the cluster of cognitive capacities and masking strongly supports the view that the animal is conscious. It seems that denying consciousness of an animal which responds in this way to masking would require us to accept that the capacities which cluster in humans because its components are facilitated by consciousness form a cluster in the animal for a different reason. Furthermore, we would need to believe that masking – while it switches off the cluster in humans because it extinguishes consciousness – switches off the cluster in animals for a different reason. The assumption that the animal is conscious explains the presence of the cognitive capacities, the sensitivity of those capacities to masking and the fact that the capacities correlate. Thus, the theory-light approach supplies a methodology which allows for particularly trustworthy inferences to animal consciousness as the best explanation of cognition and behavior. It results a test for consciousness: An animal is conscious if it possesses a cluster of cognitive capacities which are each facilitated by consciousness in humans and which can be selectively switched off by masking.

The same test can be applied to an AI system, if it has a functional analogue to visual perception. We can look in AI systems for capacities which are facilitated by consciousness in humans; they don't even need to be the three we have mentioned. In addition, we can investigate the machine's threshold for supraliminal as opposed to subliminal[15] perception and the timing which leads to masking (if any). This can be done in one of two ways: If the machine is able to report its experiences – verbally or otherwise (e.g. via button presses) – then these reports can indicate which signal strength suffices for processing which may be conscious. Second, we can see at which threshold the relevant cluster of cognitive capacities becomes available.

Based on this, the test works analogously to the animal case. We have good grounds for believing that an AI is conscious if it possesses a cluster of cognitive capacities which are each facilitated by consciousness in humans and which can be selectively switched off by masking. This validates attributions of consciousness and the masking procedure at the same time. Note that this test works only in conjunction with the naturalness constraint. The clustering of cognitive capacities and the sensitivity to masking only provide good evidence for consciousness because it would be a strange *coincidence* if those patterns occurred without having conscious experience as a common cause. However, if an AI were constructed ad hoc to have a set of cognitive capacities that clusters and to make the occurrence of this cluster dependent on the absence of masking, this would not be a coincidence. Thus, in the case of ad hoc AIs, the theory-light test does not have much evidential strength.

Nevertheless, one can still coherently be skeptical of consciousness in an AI that passes this test. Perhaps the machine's cognition functions in ways that are – even in many details – very similar to the aspects of human cognition which are facilitated by consciousness, but it nonetheless lacks one mechanism necessary for consciousness. In this case, the machine might pass the theory-light test and satisfy the naturalness constraint without being conscious.

[15] A stimulus is perceived *subliminally* if the stimulus signal is so weak that it is not experienced consciously (e.g., because the stimulus is shown very briefly), but it still has significant cognitive effects.

While I do not deny this possibility, it is important to see that the same skeptical worry can be raised in the case of animals. In the animal as well as the AI case, the theory-light test does not provide definitive proof since alternative explanations of the indicators of consciousness can be thought of. Yet, it is questionable whether any set of tests of consciousness – let alone a single test – could provide definitive proof of consciousness.[16] I submit that the theory-light test, coupled with the naturalness constraint, does provide good, although not infallible, grounds for attributions of consciousness to machines. It enables attributions of consciousness to AIs which are no less justified than attributions of consciousness to animal species like bees or octopodes, where we cannot support claims of consciousness in terms of neurophysiological similarity to humans. In the next section, we will look at the second test of machine consciousness which can be adapted from animal research.

7 Double Dissociation Paradigms

Ben-Haim et al. (2021) recently conducted an experiment on Rhesus monkeys which can serve as a blueprint for further research on animal consciousness more generally and – I claim – machine consciousness. This paradigm exploits double dissociations of visual consciousness, i.e., conditions in which processing of consciously accessible stimuli (supraliminal stimuli) and of stimuli just below the threshold for conscious perception (subliminal stimuli) have opposite effects on performance. To elicit this double dissociation, they used a spatial cueing task. In this task, a target appears in one of two possible locations on a screen and the subjects have to identify the target location as quickly as possible (the subjects eye gaze is used as the mode of response). A cue precedes the target. This cue is either presented subliminally (17/33 ms) or supraliminally (250 ms). Crucially, the cue is presented in the opposite location of the target. That is, the cue predicts the appearance of the target, but it is incongruent, i.e., its location is always the opposite from the target's location.

Thus, in the standard setup, we have three types of conditions: A condition with supraliminal cues, one with subliminal cues and a control condition in which the cues do not predict the target's location. The question is in which conditions humans and Rhesus monkeys learn the cue-target relationship and can thus use the cue to identify the target more quickly. The results are the same for humans and Rhesus monkeys: When the cue is presented supraliminally, humans and monkeys identify the target faster than in the control condition. When the cue is presented subliminally, they perform *worse* than in the control condition. Based on their subsequent verbal reports, Ben-Haim et al. verified that humans indeed were conscious of the supraliminally presented cues and did not perceive the subliminal ones.

This result is more revealing than the kinds of cases considered by Birch (2022) since it suggests the existence of two distinct processing modes. It is not only the case that conscious perception improves performance in humans, but non-conscious processing can have characteristic effects as well: it impairs performance. Since Rhesus monkeys show the same double dissociation in the same task, it seems very likely

[16] After all, do we have conclusive proof that our fellow humans are conscious (Avramides, 2020)?

that the dissociation signifies the distinction between conscious and non-conscious processing in them as well.

Crump and Birch (2021) discuss the worry that the difference in performance between supra- and subliminally presented cues may be due not to a difference between conscious and non-conscious processing but to a difference in signal strength. The predictive relationship between cue and stimulus may be easier to learn in the supraliminal condition, not because a consciousness difference, but because supraliminal signals are stronger and therefore easier to learn about. Ben-Haim et al. can rule out this possibility in the human case by performing a modified version of the experiment, in which they informed the human subjects that there are subliminal cues. In this variation, humans perform better in identifying the target location when presented the subliminal cue than in the original experiment. Since the signal strength did not change, this performance enhancement must be due to the fact that subjects manage to become conscious of more of the subliminal stimuli. This is confirmed by the subjects' verbal report.

While telling subjects about the subliminal cues and collecting their verbal report later on is not possible when dealing with non-human animals, it may be possible regarding some AIs. In cases where it is not, one may, however, worry that differences in signal strength, not consciousness, explain the double dissociation when it is found in machines. Based on the theory-light strategy, there is a method to alleviate this worry:

Crump & Birch suggest identifying a putative threshold for supraliminal perception by varying the stimulus duration continuously and confirming at which point the double dissociation appears. Then, one can test whether the threshold is the same for other cognitive tasks which are facilitated by consciousness in humans (like the ones mentioned last section). If this is the case, then this strongly suggests that the threshold is indeed the threshold for conscious perception and that the distinction between conscious and non-conscious processing, rather than differences in signal strength, explains the double dissociation. For there is no reason why the same signal strength should be necessary for performing a diverse range of cognitive tasks. In contrast, the hypothesis that consciousness is necessary for these tasks explains why a range of cognitive capacities depends on similar stimulus thresholds.

In this case, one enhances the evidential weight of double dissociation experiments by allying them with the theory-light strategy. It is important to see that the contribution of double dissociations is not superfluous. The tools of the theory-light strategy – masking and the search for clusters of cognitive capacities – support the contention that the differential performance in response to supra- versus subliminal cues is caused by the distinction between conscious and non-conscious processing. However, if an AI shows these two distinctive and opposing performance patterns in response to different kinds of cues, which mirror what we find in humans, then this is a striking contribution to the evidence provided by the theory-light approach. Left by itself, the theory-light tests do not directly indicate a distinction between conscious and non-conscious processing of a stimulus, but only between consciously processing the stimulus and not processing it at all.

While double dissociation tasks provide solid evidence of consciousness, they again are susceptible to ad hoc constructions. There is no reason why it should not

be possible to build a machine which exhibits opposing responses to cues below and above a certain threshold. For this reason, we need to add the naturalness constraint to this test of consciousness as well.

Given the naturalness constraint, the rationale for classifying this test as a valid test of animal consciousness transfers to AI systems. Finding behavioral similarities between AI behavior and human behavior characteristic of consciousness down to such a fine level of grain is best explained by postulating that the AI is conscious. While non-conscious causes are not definitively ruled out, passing this test would license the believe that an AI is conscious, if further evidence is absent.

I conclude that we have good grounds for believing that a machine is conscious if it behaves like humans and Rhesus monkeys in a double dissociation task, the dissociation cannot be explained in terms of differences in signal strength because the alleged threshold of conscious perception is constant among several different cognitive tasks and the AI was not designed ad hoc.[17]

8 Conclusion

This paper has argued for three distinctive theses. First, a proposed test of machine consciousness can be legitimate, even if AIs can be gerrymandered specifically to pass this test. This is underscored by the observation that many, if not all, putative tests of machine consciousness fail to rule out that they can be passed by non-conscious machines via gerrymandering. Second, we can identify such gerrymandered AIs by taking inspiration from the criteria philosophers of science have developed to detect ad hoc hypotheses. These criteria are far removed from delivering exact scores for determining degrees of ad-hoc-ness; nevertheless, they suffice for practical purposes. Third, given the first and the second claim, the currently most reliable tests of animal consciousness turn out to be valid and useful tests of machine consciousness as well. If a natural AI exhibits clusters of cognitive capacities facilitated by consciousness in humans which can be selectively switched off by masking and if it reproduces human behavior in suitably designed double dissociation tasks, we should treat the AI as conscious.

I will close this paper by pointing to the main limitation of the tests proposed here and to avenues for further research. This limitation is that these tests are only intended to indicate the presence, not the absence, of consciousness in a machine. There are different ways in which conscious machines may fail the test or even not be a candidate for being tested.

First, the tests I proposed here depend on specific knowledge about the way the machine processes sensory stimuli: in particular its thresholds for supraliminal versus subliminal perception and for masking. Since we need to ascertain these properties for each AI anew, the specifics of the test need to be tailored to every AI that takes it.

[17] To make our attributions of AI consciousness more reliable, a next step might consist in relaxing the theory-neutrality desideratum. We might integrate the knowledge we gained by employing relatively theory-neutral tests with our theoretical conception of consciousness, as suggested by Shevlin (2021) in the context of animal consciousness.

Second, human perception can be exteroceptive, i.e. based on processing stimuli from outside the body, and interoceptive, i.e. the perception of bodily states. Human conscious experiences might be valenced, i.e. feel good or bad. Pain, joy, fear and relief are examples of valenced experiences. These distinctions are important because they entail that AIs might be conscious even if they lack exteroceptive perceptual states and non-valenced conscious experience. Tests for interoceptive and valenced conscious experiences in AI need to be designed. Also, there may be AIs which possess a purely cognitive, non-perceptual kind of conscious experience. A suitably specified version of the ACT might turn out to be complementary to the tests proposed in this paper, since the ACT does not require its subjects to have perceptual states.

Third, there may be "alien" forms of consciousness which do not manifest in the ways typical for human consciousness. Some beings may be conscious without having any cognitive capacities which are facilitated by consciousness in humans. This challenge already arises for animal consciousness research. When applied to AI, the challenge magnifies since the space of possible artificial minds is so much vaster and more heterogenous than the set of all actual animal minds. If a form of consciousness is sufficiently alien such that it has a vastly different function from human consciousness, then the methods proposed here won't help to detect it.

Acknowledgements I thank Wanja Wiese for helpful comments.

Funding Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number GRK-2185/2 (DFG Research Training Group Situated Cognition).
Open Access funding enabled and organized by Projekt DEAL.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest No potential conflict of interest.

References

Avramides, A. (2020). Other Minds. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2020/entries/other-minds/.

Barnes, E. C. (2021). Prediction versus Accommodation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/prediction-accommodation/.

Bellebaum, C., & Daum, I. (2004). Effects of Age and Awareness on Eyeblink conditional discrimination learning. *Behavioral Neuroscience*, *118*(6), 1157–1165. https://doi.org/10.1037/0735-7044.118.6.1157.

Ben-Haim, M. S., Monte, D., Fagan, O., Dunham, N. A., Hassin, Y., Chang, R. R., S. W. C., & Santos, L. R. (2021). Disentangling perceptual awareness from nonconscious processing in rhesus monkeys (Macaca mulatta). *Proceedings of the National Academy of Sciences*, *118*(15). https://doi.org/10.1073/pnas.2017543118.

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

Birch, J., & Andrews, K. (2023). *What has feelings?* Aeon. https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals.

Birch, J., Schnell, A. K., & Clayton, N. S. (2020). Dimensions of animal consciousness. *Trends in Cognitive Sciences*, *24*(10), 789–801. https://doi.org/10.1016/j.tics.2020.07.007.

Brush, S. G. (1994). Dynamics of Theory Change: The role of predictions. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1994*(2), 133–145. https://doi.org/10.1086/psaprocbienmeetp.1994.2.192924.

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and Brain systems: The role of awareness. *Science*, *280*(5360), 77–81. https://doi.org/10.1126/science.280.5360.77.

Clark, R. E., & Squire, L. R. (1999). Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science*, *10*(1), 14–18. https://doi.org/10.1111/1467-9280.00099.

Crump, A., & Birch, J. (2021). Separating conscious and unconscious perception in animals. *Learning and Behavior*, *49*(4).

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871.

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience*, *12*(2), 41–62. https://doi.org/10.1080/17588928.2020.1772214.

Droege, P., Weiss, D. J., Schwob, N., & Braithwaite, V. (2021). Trace conditioning as a test for animal consciousness: A new approach. *Animal Cognition*, *24*(6), 1299–1304. https://doi.org/10.1007/s10071-021-01522-3.

Dung, L. (2022a). Assessing tests of animal consciousness. *Consciousness and Cognition*, *105*, 103410. https://doi.org/10.1016/j.concog.2022.103410.

Dung, L. (2022b). Why the Epistemic Objection against using sentience as Criterion of Moral Status is flawed. *Science and Engineering Ethics*, *28*(6), 51. https://doi.org/10.1007/s11948-022-00408-y.

Dung, L. (2023). How to deal with risks of AI suffering. *Inquiry*, 1–29. https://doi.org/10.1080/0020174X.2023.2238287.

Dung, L., & Newen, A. (2023). Profiles of animal consciousness: A species-sensitive, two-tier account to quality and distribution. *Cognition*, *235*, 105409. https://doi.org/10.1016/j.cognition.2023.105409.

Elamrani, A., & Yampolskiy, R. V. (2019). Reviewing tests for machine consciousness. *Journal of Consciousness Studies*, *26*(5–6), 35–64.

Gardner, M. R. (1982). Predicting Novel facts. *The British Journal for the Philosophy of Science*, *33*(1), 1–15. https://doi.org/10.1093/bjps/33.1.1.

Ginsburg, S., & Jablonka, E. (2019). *The evolution of the sensitive soul: Learning and the origins of consciousness*. The MIT Press. https://doi.org/10.7551/mitpress/11006.001.0001.

Halina, M. (2021). Insightful artificial intelligence. *Mind & Language*, *36*(2), 315–329. https://doi.org/10.1111/mila.12321.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1–34.

Howson, C. (1988). Accommodation, prediction and bayesian confirmation theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1988*(2), 381–392. https://doi.org/10.1086/psaprocbienmeetp.1988.2.192899.

Irvine, E. (2012). *Consciousness as a Scientific Concept: A philosophy of Science Perspective*. Springer.

Jaworska, A., & Tannenbaum, J. (2021). The Grounds of Moral Status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/.

Leplin, J. (1975). The Concept of an ad Hoc Hypothesis. *Studies in History and Philosophy of Science Part A*, *5*(4), 309. https://doi.org/10.1016/0039-3681(75)90006-0.

Lipton, P. (1990). Prediction and prejudice. *International Studies in the Philosophy of Science*, *4*(1), 51–65. https://doi.org/10.1080/02698599008573345.

Maher, P. (1993). Howson and Franklin on Prediction. *Philosophy of Science*, *60*(2), 329–340. https://doi.org/10.1086/289736.

Margolis, E., & Laurence, S. (2003). Concepts. In S. P. Stich, & T. A. Warfield (Eds.), *Blackwell guide to philosophy of mind* (pp. 190–213). Blackwell.

Mason, G. J., & Lavery, J. M. (2022). What Is It Like to Be a Bass? Red Herrings, Fish Pain and the Study of Animal Sentience. *Frontiers in Veterinary Science*, *9*. https://www.frontiersin.org/articles/https://doi.org/10.3389/fvets.2022.788289.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *08*(01), 43–66. https://doi.org/10.1142/S270507852150003X.

Mudrik, L., Faivre, N., & Koch, C. (2014). Information integration without awareness. *Trends in Cognitive Sciences*, *18*(9), 488–496. https://doi.org/10.1016/j.tics.2014.04.009.

Nagel, T. (1974). What is it like to be a Bat? *Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914.

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

Saad, B., & Bradley, A. (2022). Digital suffering: Why it's a problem and how to prevent it. *Inquiry*, 1–36. https://doi.org/10.1080/0020174X.2022.2144442.

Schneider, S. (2019). *Artificial You: AI and the future of your mind*. Princeton University Press. https://doi.org/10.1515/9780691197777.

Schwitzgebel, E. (2020). Is there something it's like to be a Garden snail. *Philosophical Topics*, *48*(1), 39–63. https://doi.org/10.5840/philtopics20204813.

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*, *23*(7), https://doi.org/10.1038/s41583-022-00587-4. Article 7.

Shevlin, H. (2020a). General intelligence: An ecumenical heuristic for artificial consciousness research? *Journal of Artificial Intelligence and Consciousness*. https://doi.org/10.17863/CAM.52059.

Shevlin, H. (2020b). Which animals Matter? Comparing approaches to Psychological Moral Status in Non-human systems. *Philosophical Topics*, *48*(1), 177–200. https://doi.org/10.5840/philtopics20204819.

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

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour*, *97*, 201–212. https://doi.org/10.1016/j.anbehav.2014.09.007.

Sober, E. (2015). *Ockham's razors: A user's Manual*. Cambridge University Press. https://doi.org/10.1017/CBO9781107705937.

Tomasik, B. (2014). Do Artificial Reinforcement-Learning Agents Matter Morally? *ArXiv:1410.8233 [Cs]*. http://arxiv.org/abs/1410.8233.

Travers, E., Frith, C. D., & Shea, N. (2018). Learning rapidly about the relevance of visual cues requires conscious awareness. *Quarterly Journal of Experimental Psychology*, *71*(8), 1698–1713. https://doi.org/10.1080/17470218.2017.1373834.

Tye, M. (2017). *Tense bees and Shell-shocked crabs: Are animals conscious?* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190278014.001.0001.

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

White, R. (2003). The epistemic advantage of prediction over accommodation. *Mind*, *112*(448), 653–683. https://doi.org/10.1093/mind/112.448.653.

Wiese, W., & Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences*, *2*. https://doi.org/10.33735/phimisci.2021.81.

Wittgenstein, L. (1953). *Philosophical investigations*. Wiley-Blackwell.

Worrall, J. (1989). Fresnel, Poisson and the White Spot: The role of successful predictions in the Acceptance of Scientific theories. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the Natural sciences* (pp. 135–157). Cambridge University Press.

Zahar, E. (1973). Why did Einstein's Programme supersede Lorentz's? (I). *The British Journal for the Philosophy of Science*, *24*(2), 95–123. https://doi.org/10.1093/bjps/24.2.95.