Trends in Cognitive Sciences

CellPress OPEN ACCESS

Opinion

# Identifying indicators of consciousness in AI systems

Patrick Butlin [1,2,*], Robert Long [2], Tim Bayne [3,4], Yoshua Bengio [5,6], Jonathan Birch [7], David Chalmers [8], Axel Constant [9], George Deane [10], Eric Elmoznino [5,6], Stephen M. Fleming [4,11], Xu Ji [12], Ryota Kanai [13], Colin Klein [14], Grace Lindsay [15], Matthias Michel [16], Liad Mudrik [4,17], Megan A.K. Peters [4,18], Eric Schwitzgebel [19], Jonathan Simon [10], and Rufin VanRullen [20]

**Rapid progress in artificial intelligence (AI) capabilities has drawn fresh attention to the prospect of consciousness in AI. There is an urgent need for rigorous methods to assess AI systems for consciousness, but significant uncertainty about relevant issues in consciousness science. We present a method for assessing AI systems for consciousness that involves exploring what follows from existing or future neuroscientific theories of consciousness. Indicators derived from such theories can be used to inform credences about whether particular AI systems are conscious. This method allows us to make meaningful progress because some influential theories of consciousness, notably including computational functionalist theories, have implications for AI that can be investigated empirically.**

## Highlights

The prospect of consciousness in artificial intelligence (AI) systems increasingly demands attention given recent advances in AI and increasing capacity to reproduce features of the brain that are associated with consciousness.

There are risks of both under- and over-attribution of consciousness to AI systems, entailing a need for methods to assess whether current or future AI systems are likely to be conscious.

We argue that progress can be made by drawing out the implications of some neuroscientific theories of consciousness.

We outline a method that involves deriving indicators from theories and using them to assess particular AI systems.

## The problem of AI consciousness

The issue of consciousness in AI is increasingly attracting attention. There is deep uncertainty about whether AI consciousness is possible at all, as some researchers argue that only living organisms can be conscious [1–3]. However, AI capabilities are developing rapidly, and others argue that AI systems could be strong candidates for consciousness within the next decade[i]. If AI consciousness is possible at all, there is some reason to suspect that it may be realized in the near term. Researchers aiming to improve AI capabilities have proposed – and in some cases built – systems that intentionally reproduce computational features associated with human consciousness [4,5].

Furthermore, modern AI systems are likely to give users the impression that they are conscious. In a recent study, a majority of participants were willing to attribute some possibility of consciousness to ChatGPT, with more frequent users tending to say that consciousness is more likely [6]. AI companions are proliferating, and some users will likely believe that these companions are conscious [7]. We may be entering a period of considerable public disagreement and uncertainty about AI consciousness [8].

We face risks of both underattribution and overattribution of consciousness to AI systems. If we fail to identify consciousness in systems in which it is present, we risk causing avoidable harms to those systems, which may exist in large numbers [9]. Conversely, if we attribute consciousness to non-conscious systems, we may waste resources or risk lives trying to promote their welfare. If concern about consciousness in AI grows, we will need a principled basis on which to either dismiss these concerns or, potentially, take action to regulate AI development or use. We need empirically-grounded, rigorous, and reliable methods for assessing AI consciousness.

[1]Global Priorities Institute, University of Oxford, Oxford, UK
[2]Eleos AI Research, Berkeley, CA, USA
[3]Philosophy Department, Monash University, Melbourne, VIC, Australia
[4]Brain, Mind, and Consciousness Program, Canadian Institute for Advanced Research (CIFAR)
[5]Department of Computer Science and Operations Research, University of Montreal, Montreal, QC, Canada
[6]MILA Quebec AI Institute, Montreal, QC, Canada
[7]Department of Philosophy, Logic, and Scientific Method, London School of Economics and Political Science, London, UK
[8]Department of Philosophy, New York University, New York, NY, USA
[9]School of Engineering and Informatics, University of Sussex, Falmer, UK

This situation sets a challenge for consciousness science. Although some progress has been made in developing tests for consciousness, it remains unclear how they should (or even could) be validated, and tests for AI consciousness are an especially challenging case [10]. In this article we focus on how to assess AI systems for consciousness, rather than on whether AI consciousness is possible at all. We offer a guide to the theory-derived indicator method, which we believe offers a tractable way to reduce uncertainty. This method involves deriving **indicators** (see Glossary) of consciousness from neuroscientific theories and using them to assess particular AI systems. It was adopted using a cluster of **computational functionalist** theories in a recent report, 'Consciousness in artificial intelligence: insights from the science of consciousness' (henceforth 'Consciousness in AI' [11]), but can be used with other theories, including theories yet to be developed. We describe how to derive indicators from theories and apply them to AI systems, as well as explaining the rationale for the method and its relationship to computational **functionalism**.

## The theory-derived indicator method

Can we use theories of consciousness to assess AI systems for consciousness (as defined in Box 1)? A skeptic could point to major obstacles: researchers disagree about theories of consciousness [12,13], and some doubt whether conventional hardware can support consciousness at all [3,14–16][ii]. Furthermore, most theories have been developed based on evidence from humans and other mammals, leaving it unclear how to extend them to AI systems [17–21]. Despite these challenges, we can make progress in evaluating AI consciousness by investigating the implications of mainstream theories. Some mainstream theories suggest conditions for consciousness that AI systems could meet; in many cases, whether a system meets such conditions is a substantive empirical question. So we propose the following method: identify the conditions implied by suitable theories, then investigate whether AI systems meet them, construing these conditions as indicators of consciousness. This method can help us to judge how likely particular AI systems are to be conscious.

### Criteria for suitable theories

We can use this approach most productively with theories that have two properties. First, the theories must warrant sufficiently high credence that it is worthwhile to draw out their implications.

#### Box 1. Defining 'consciousness'

By 'consciousness' we mean phenomenal consciousness [85]. One way of gesturing at this concept is to say that an entity has phenomenally conscious experiences if (and only if) there is 'something it is like' for the entity to be the subject of these experiences [86]. One approach to further definition is through examples [87]. Clear examples of phenomenally conscious states include perceptual experiences, bodily sensations, and emotions. A more difficult question, which relates to the possibility of consciousness in large language models (LLMs), is whether there can be phenomenally conscious states of 'pure thought' with no sensory aspect [88]. Phenomenal consciousness does not entail a high level of intelligence or human-like experiences or concerns.

A further question is whether consciousness is determinately present or absent in all cases, with no borderline cases in between. One possibility is that any given system is either wholly conscious or wholly non-conscious [89]. However, an alternative is that it can is sometimes be indeterminate whether a system is conscious or not [90]. A distinct issue is whether consciousness comes in degrees, so that one system can be more conscious than another [91], perhaps along multiple dimensions [92].

Some theories of consciousness focus on access mechanisms rather than the phenomenal aspects of consciousness (e.g., [28]). However, some argue that these two aspects entail one another or are otherwise closely related (e.g., [93]). So these theories may still be informative about phenomenal consciousness.

Illusionists claim that there is no such thing as phenomenal consciousness, at least as it is usually understood [94]. If illusionism is correct, then rather than asking whether any AI systems could be phenomenally conscious, it would make more sense to ask what gives some entities the kinds of significance often associated with phenomenal consciousness, and whether AI systems could have this property.

[10]Department of Philosophy, University of Montreal, Montreal, QC, Canada
[11]Department of Experimental Psychology and Institute of Cognitive Neuroscience, University College London, London, UK
[12]School of Life Sciences, University of Westminster, London, UK
[13]Araya, Inc., Tokyo, Japan
[14]School of Philosophy, Australian National University, Canberra, ACT, Australia
[15]Department of Psychology, New York University, New York, NY, USA
[16]Department of Linguistics and Philosophy, Massachusetts Institute of Technology, Cambridge, MA, USA
[17]Sagol School of Neuroscience, Tel-Aviv University, Tel-Aviv, Israel
[18]Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, USA
[19]Department of Philosophy, University of California, Riverside, Riverside, CA, USA
[20]Brain and Cognition Research Center, Centre National de la Recherche Scientifique, Toulouse, France

*Correspondence:
patrick.butlin@gmail.com (P. Butlin).

Second, we will learn more from theories that imply clear and testable conditions that AI systems might meet; some theories imply conditions that AI systems evidently cannot meet, and these are less relevant. As the science of consciousness progresses, different theories will come to satisfy these criteria, so the selection of theories should change accordingly.

Computational functionalist theories propose computational properties as conditions for consciousness (Box 2). Such properties may be found in AI systems, so these theories will satisfy the second criterion provided that their conditions are clear and testable. These theories claim that certain brain states are conscious due to the roles they play in the brain's information-processing architecture. For example, global workspace theory (GWT) identifies consciousness with the global broadcast of information to many neurocognitive modules, allowing integration between them [22,23]. Integration through global broadcast is a condition that AI systems might meet; in contrast, if a theory claimed that having a cortex is necessary for consciousness, no AI system could meet this condition.

Consequently, at present, examples of theories that arguably meet the two criteria include recurrent processing theory (RPT) [24–26], GWT [22,23,27,28], higher-order theories (HOT) [29–31], and attention schema theory (AST) [32,33]. These theories are the products of a substantially shared research program in neuroscience, studying both brain activity associated with consciousness (its 'neural correlates' [34]) and the relationships between consciousness and functions such as attention, learning, memory, and decision-making. Refining these theories, which are among the most influential in the field [12,13,35], has driven significant methodological progress [36].

Theories that do not endorse computational functionalism could, in principle, also satisfy the second criterion. For example, AI systems using non-conventional hardware might meet the conditions of integrated information theory (IIT)[ii]. However, focusing on theories that can be given

## Box 2. Computational functionalism and alternative views

As we interpret them, the theories we rely on to derive indicators for consciousness share a commitment to computational functionalism. That is, they agree that implementing computations of the right kind is necessary and sufficient for consciousness. According to computational functionalism, two systems that are similar at the relevant algorithmic level of description will also be similar with respect to consciousness.

If computational functionalism is true, then consciousness in AI systems built on conventional hardware is possible in principle – assuming that conventional hardware is capable of implementing the relevant computations. One version of the method we propose adopts computational functionalism as a working assumption, and considers questions that flow from this assumption – which computational properties are necessary and sufficient for consciousness, and could they be implemented in AI systems at present or in the near future? However, many theorists favor alternative views, and each of these other views raises different questions about AI consciousness.

Biological substrate views, on which properties such as being made of living cells are necessary for consciousness, are one alternative to computational functionalism [2,3,16,18,95]. These views suggest that a biological substrate may be necessary either because it makes possible certain fine-grained, non-computational patterns of functional organization [3,16] or due to some more direct connection with consciousness [3]. On biological substrate views, relevant questions about AI consciousness include which distinctively biological properties of organisms are necessary for consciousness, and whether these could be implemented in AI, perhaps using unconventional hardware [96].

A further alternative is integrated information theory (IIT), which claims that consciousness depends on the structure of the causal relations between the physical components of a system. What matters, however, is not whether this structure of causal relations implements a certain algorithm, but whether the components that are thus related form a unified whole, according to a mathematical definition specified by the theory [97]. Proponents of IIT argue that AI systems on conventional hardware are unlikely to be conscious [98]. This again raises the question of whether unconventional hardware could make AI consciousness possible, as some proponents suggest [17].

## Glossary

**Algorithmic recurrence:** a form of processing in which the same operation is applied repeatedly, such as processing in a neural network in which information passes through layers with the same weights. This is algorithmically similar to processing in a network that has backward connections at the level of physical implementation, such as the brain, because this also entails that the same operations are applied repeatedly.
**Computational functionalism:** the thesis that implementing computations of a certain kind is necessary and sufficient for consciousness. Computational functionalism entails functionalism but not vice versa.
**Functionalism:** the thesis that having a certain kind of functional organization is necessary and sufficient for consciousness.
**Indicators:** properties that we can look for in artificial intelligence (AI) systems that indicate that they are more (or less) likely to be conscious. We do not claim that the indicators are individually necessary for consciousness or that any combination is sufficient.
**Interpretability methods:** methods to understand the workings and outputs of machine learning models, such as by investigating the algorithms they use and the internal representations they form.
**Minimal implementation problem:** the problem that some computational functionalist theories of consciousness may give conditions that would be met by very simple artificial systems. These systems are potential counterexamples to the theories.
**Negative indicators:** properties of a system that should decrease our credence that the system is conscious.
**Positive indicators:** properties of a system that should increase our credence that the system is conscious.
**Sparse and smooth coding:** a coding scheme in which properties are represented by relatively few neurons (sparseness) and by a continuous scheme, rather than one that divides them into discrete categories (smoothness). For example, red/green/blue (RGB) coding represents colors continuously, whereas color words such as 'purple' and 'yellow' divide them into categories.
**Specificity and sensitivity:** indicators can be useful in virtue of either specificity or sensitivity. An indicator property has high specificity if few non-conscious

computational functionalist interpretations makes this method tractable and relevant to current and near-future systems. While many of us are agnostic about computational functionalism, we agree that it provides a useful focus for assessments. Biological substrate views (Box 2) will not meet the second criterion because they imply conditions that AI systems straightforwardly cannot meet. But these views should still be considered in overall assessments of the likelihood of consciousness in AI.

systems have it, and high sensitivity if few conscious systems lack it.
**Valenced conscious experience:** conscious experience that feels good or bad, such as pleasure or pain.

### Deriving and interpreting indicators

A factor affecting the interpretation of inferences from neuroscientific theories is that these theories can be formulated either narrowly, as making claims about what grounds the distinction between conscious and unconscious states in humans, or broadly, as making claims about necessary and/or sufficient conditions for consciousness in systems of any kind. Narrow formulations of theories are more directly supported by evidence from human subjects, while broad formulations make more explicit claims about AI systems. Theories like the four mentioned above can be formulated in either way; advocates of AST, GWT, and HOT have sometimes formulated their theories broadly [31,37,38], despite their basis in human neuroscience. For our purposes, what matters is that theories have implications for AI when formulated in either way. If a theory says that condition C suffices for consciousness in all systems (a broad claim), then, conditional on the truth of the theory, any AI system that satisfies C must be conscious. If a theory says that condition C distinguishes conscious from unconscious states in humans (a narrow claim), we cannot infer that an AI system that meets C would be conscious because certain background conditions may also be necessary. However, we can reasonably increase our credence that the system is conscious if we have non-zero credence that the relevant background conditions are met.

Because no one theory of consciousness is currently dominant, a program to assess AI systems using our approach should draw on multiple theories. These competing theories will not collectively provide a set of necessary and sufficient conditions, so our approach is to treat the properties that they each identify as indicators of consciousness – markers that can increase or decrease one's credence that the system is conscious. We focus on **positive indicators**, which increase credences (see the section 'What does it tell us if a system possesses indicator properties?'). Using indicators has been proposed in earlier work on the distribution of consciousness [39,40], especially concerning non-human animals [41–44], but our approach is distinctive in deriving indicators from multiple theories. AI systems are better candidates for consciousness – we have more reason to believe that they are conscious – if they have more of these properties.

To the extent that one has confidence in the theories from which (positive) indicators are derived, finding that an AI system has some of the indicators should increase one's credence that it is conscious, and finding that it has none or few should decrease one's credence. Every theory of consciousness faces objections, and compelling objections should lead us to give less weight to the corresponding indicators. However, we stress that indicators are merely intended to be credence-shifting; we do not need to be certain that a theory is correct for it to provide useful indicators.

We envisage deriving indicators from theories in two ways. First, theories typically make claims about what distinguishes conscious from unconscious states; indicators can be taken from these accounts and will be attributable to particular theories. Second, the broader theoretical landscape suggests plausible background conditions. These may be necessary for consciousness but not sufficient, whereas theories may claim that sets of conditions are necessary and jointly sufficient. Background conditions might include the

presence of representational states, predictive processing, agency or embodiment [11]. Some of these properties are emphasized by many theories; for example, sensorimotor [45], active inference [46], neurorepresentationalist [47], and midbrain [48] theories all emphasize links between agency and consciousness.

### Internal and behavioral evidence

An advantage of using theories to derive indicators of consciousness is that this gives us standards by which to assess the internal processes of AI systems, rather than their behavior or capabilities. We assume that whether a system is conscious depends on features of its internal processes. This does not mean that behavioral evidence cannot be useful in some cases - indeed, it has been argued that behavioral evidence should currently be prioritized in research on the distribution of consciousness in non-human animals ([17]; *cf* [49]). But behavioral tests for consciousness in AI systems [50,51] face significant challenges. One problem is that in building AI systems we are likely to discover new ways to achieve behavioral capabilities, which may not involve consciousness, since biological constraints do not apply [21,39]. Recent large language models (LLMs) provide a dramatic illustration of this, showing that in the case of AI, inferences from behavior to features of internal processes are often unreliable. This problem is exacerbated by incentives to build AI systems that mimic aspects of human behavior [52] (Box 3). That said, carefully-designed behavioral tests could provide some evidence for the presence of our indicators and show that they support consciousness-linked capacities in particular systems.

## Identifying indicator properties

The method we propose is to derive indicators from theories of consciousness, then assess whether AI systems are likely to be conscious by determining whether they possess these properties. In this section we focus on the issue of how to derive indicators from theories.

### Box 3. The 'gaming problem' for measures of AI consciousness

Any measure or indicator that is merely correlated with, and is neither constitutive of nor sufficient for, a phenomenon of interest is vulnerable to being 'gamed' ([99]; *cf* Goodhart's law in [100]). This potentially includes some of the indicators of consciousness listed in Table 1. An indicator is gamed if its presence is better explained by the fact that it makes a system seem to possess a property of interest than by the fact that the system actually possesses the property. In AI contexts, the gaming worry arises especially for superficial behavioral indicators of consciousness, such as speech or facial expressions. Although in an ordinary human, saying 'Hello!' or smiling might indicate the presence of consciousness, AI systems can mimic these aspects of human behavior while lacking consciousness [101,102][iii]. For any purported behavioral indicator of consciousness, an engineer might attempt to design a nonconscious system that manifests that indicator. Accordingly, behavioral properties proposed as potential indicators of consciousness in animals might be too readily gamed in AI.

Although simple behavioral markers are especially vulnerable to the gaming problem, the problem can also arise for computational markers. Suppose that some computational feature N is not sufficient for consciousness but is taken to be an indicator of consciousness. It would be possible to design a nonconscious system with N, thereby gaming N and making it a less reliable indicator. Even if engineers are not explicitly seeking to mislead users about a system's consciousness, to the extent that users or others value systems because they possess what seem to be indicators of consciousness, the gaming problem arises.

To mitigate the gaming problem, we recommend (i) emphasizing, to the extent possible, indicators that are sufficient for consciousness or that cannot easily be designed without also creating consciousness, and (ii) when evaluating systems with gameable indicators, assessing whether the system lacks or possesses other secondary or supporting features that increase the likelihood that the indicator is accurate.

From the perspective of computational functionalism, these conditions are more likely to be satisfied if a system has high computational similarity to biological systems that are known to be conscious. In the absence of a complete computational theory of consciousness, what is computationally sufficient for consciousness might depend on features that are not yet known to be relevant.

We state four guidelines for this process, illustrating them with examples from the list of indicators in 'Consciousness in AI' [11] (Table 1 and Figure 1).

(i) Indicators should focus on the central explanatory posits of theories of consciousness

Theories of consciousness are often presented in detailed accounts and use concepts that may imply further commitments. However, indicators to be used to assess AI systems for consciousness should focus on theories' central explanatory posits, abstracting away from much of this detail. This focus will typically mean that indicators are conditions that theories claim are individually

Table 1. Potential indicators of consciousness [12]

| | |
|---|---|
| Recurrent processing theory (RPT) [24–26] | |
| RPT-1: Input modules using algorithmic recurrence | RPT-1 and RPT-2 are largely independent indicators<br>RPT-1 is also supported by the idea that consciousness is integrated over time [103]<br>Discussion related to RPT-2 can be found in [36,104] |
| RPT-2: Input modules generating organized, integrated perceptual representations | |
| Global workspace theory (GWT) [27–30] | |
| GWT-1: Multiple specialized systems capable of operating in parallel (modules) | GWT claims that these are necessary and jointly sufficient<br>GWT-1–GWT-4 build on one another<br>GWT-3 and GWT-4 entail RPT-1 |
| GWT-2: Limited capacity workspace, entailing a bottleneck in information flow and a selective attention mechanism | |
| GWT-3: Global broadcast: availability of information in the workspace to all modules | |
| GWT-4: State-dependent attention, giving rise to the capacity to use the workspace to query modules in succession to perform complex tasks | |
| Computational higher-order theories (HOT) [31–33] | |
| HOT-1: Generative, top-down, or noisy perception modules | Perceptual reality monitoring theory (PRM [33]) claims that these are necessary and jointly sufficient<br>HOT-1–HOT-3 build on one another, whereas HOT-4 is independent [105,106]<br>The first clause of HOT-3 is also supported by arguments concerning intentional/flexible agency and entails AE-1; HOT-3 is connected to PP |
| HOT-2: Metacognitive monitoring distinguishing reliable perceptual representations from noise | |
| HOT-3: Agency guided by a general belief-formation and action-selection system, and a strong disposition to update beliefs in accordance with the outputs of metacognitive monitoring | |
| HOT-4: **Sparse and smooth coding** generating a 'quality space' | |
| Attention schema theory (AST) [34,35] | |
| AST-1: A predictive model representing and enabling control over the current state of attention | Discussion of links between AST, GWT, and HOT can be found in [107] |
| Predictive processing (PP) [63,108,109] | |
| PP-1: Input modules using predictive coding | Entails RPT-1 and HOT-1; PP-compatible versions of GWT and HOT can be found in [110,111] |
| Agency and embodiment [47,50,112,113] | |
| AE-1: Minimal agency: Learning from feedback and selecting outputs in such a way as to pursue goals, especially where this involves flexible responsiveness to competing goals | Both indicators are supported to some extent by GWT, PRM, and PP, especially AE-1<br>Systems meeting AE-2 are likely, but not guaranteed, to also meet AE-1; on the formulation of AE-2, see [114,115] |
| AE-2: Embodiment: Modeling output-input contingencies, including some systematic effects, and using this model in perception or control | |

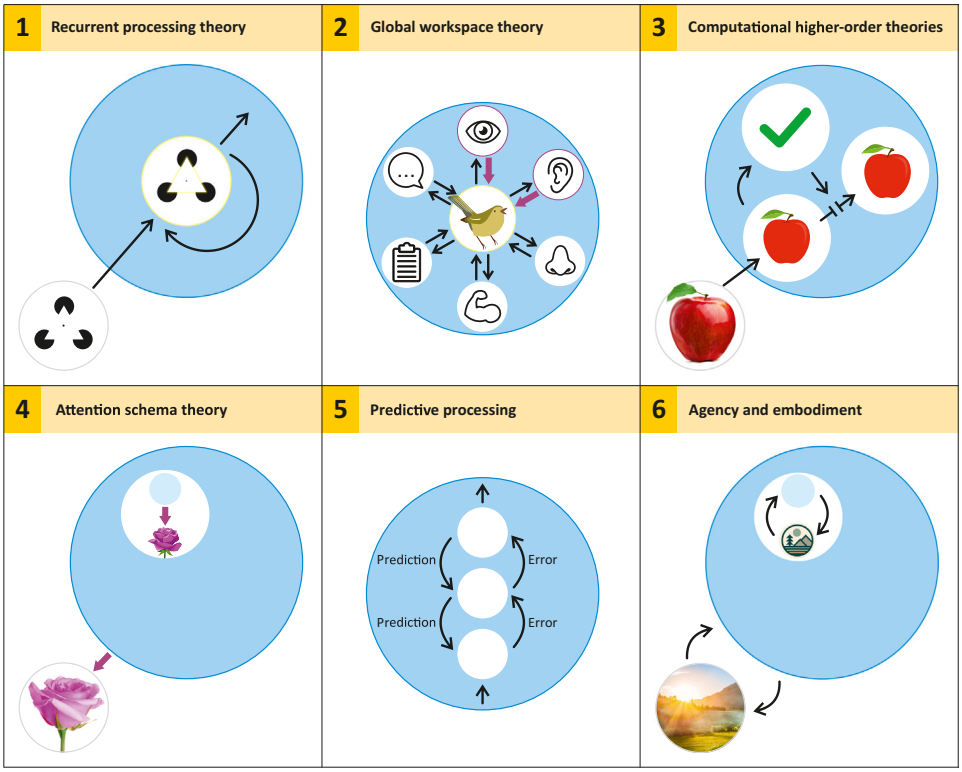


Figure 1. Theories of consciousness that inform potential indicators. Key features of the theories that inform the potential indicators listed in Table 1. Objects in the environment are shown using realistic images, and internal representations are illustrated with simplified images. (1) Recurrent processing generating a representation of an organized perceptual scene, illustrated by the Kanizsa illusion. (2) Multiple specialized subsystems integrated by a workspace, a selection mechanism determining what is represented in the workspace (here, visual and auditory information), and global broadcast back to all subsystems. (3) A metacognitive monitoring mechanism determining which perceptual representations contribute to belief-formation. (4) Attention to a visual stimulus and a model representing that state of attention. (5) Hierarchical processing in which predictions are sent down the hierarchy and error signals are sent up. (6) Interaction with an environment and the use of a model of this interaction.

necessary and jointly sufficient for consciousness. For present purposes, identifying a theory's central posits is a matter of understanding the explanation offered by the most-promising and best-supported formulation of that theory, rather than understanding the account offered by any particular theorist. This focus on central posits is necessary because theories typically aim to describe the processes underlying human consciousness, and many details may be different in other conscious beings, especially in the case of AI. These may include functional details as well as details of implementation. A restricted focus is also valuable because long lists of indicators risk redundancy or confusingly wide variation in the significance of individual indicators.

For example, a key property of the global workspace is that it can sustain representations over time, coordinating activity in modules to support complex tasks [22,38,53]. Indicator GWT-4 relates to this property (Table 1), which connects the global workspace with working memory. However, not all details of the relationship between global workspace and working memory are central explanatory posits of GWT. The current version of the theory claims that the global workspace corresponds to attended items in working memory, with other representations in

working memory being unconscious [28]. But this posit does not appear to be central to the account and therefore is not included in the indicators.

(ii) Indicator selection should maximize openness to varied forms of consciousness while avoiding the minimal implementation problem

Theories of consciousness can be formulated in more or less restrictive ways [17,18,20]; as we derive indicators from them we must avoid pitfalls on each side. For example, a restrictive formulation of GWT might specify how the workspace operates in great detail, including descriptions of exactly how information is selected, what operations are performed in the workspace, and so on. A liberal formulation might merely require a space accessible to multiple subsystems through which they can share information. The problem with very liberal formulations is that they can be satisfied by very simple artificial systems that are not plausibly conscious; many computational functionalist theories allegedly fail by giving such liberal conditions [54,55]. This is the **minimal implementation problem**: the simplest possible implementation of a theory may be a counterexample.

Consequently, some indicators should be sufficiently demanding that, if an AI system satisfies many of them, that would provide some evidence of consciousness rather than a counterexample to the theories. Examples in Table 1 arguably include GWT-4, which mentions 'complex tasks', and HOT-3, which refers to a 'general belief-formation and action selection system' [56]. However, common and simple properties of AI systems, such as RPT-1, **algorithmic recurrence**, can also be useful indicators. It may be that the absence of this property is strong evidence that a system is not conscious (see the section 'What does it tell us if a system possesses indicator properties?').

We have already argued that theories and indicators should not be excessively restrictive. However, a further consideration is that developments in AI may lead to exotic forms of consciousness [57]. To avoid false negatives in such cases, indicators should omit features such as specific sensory modalities that are unlikely to be necessary for consciousness. But indicators should still reflect theories' core commitments. For example, indicator RPT-2 specifies 'organized, integrated perceptual representations'. One might object that this is chauvinistic, invoking imagined conscious beings with radically different perceptual systems, or without perception altogether [58]. But RPT-2 should still be included because the method is intended to reflect diverse theoretical perspectives.

(iii) Indicators based on potential background conditions should be included to mitigate the narrow focus of some theories and reflect shared commitments

We propose adding indicators based on potential background conditions for consciousness for two reasons. First, because theories of consciousness tend to focus on the differences between conscious and unconscious states in humans, it is likely that they will not emphasize properties that humans always or nearly always have that may be necessary for consciousness. Predictive processing (PP-1) is an example of an indicator that may be justified in this way. The connection between predictive processing and consciousness has been widely discussed [59,60], but one perspective is that predictive processing provides a framework within which detailed theories of consciousness may be developed [61]. Given that predictive processing is also argued to be a fundamental feature of human and animal cognition, it is a plausible background condition.

Second, indicators based on background conditions may be justified when many theories, which may or may not be those from which other indicators are derived, suggest that some property of

humans and other animals is necessary for consciousness. Agency is an example of a property like this. The midbrain theory identifies consciousness with a 'unified multimodal neural model of the agent within its environment, which is weighted by the current needs and state of the agent' [62] (also see [48,63]), and neurorepresentationalism claims that consciousness subserves goal-directed behavior [47,64]. GWT can arguably also be included since Dehaene and Naccache list 'intentional behavior' together with 'durable and explicit information maintenance' and 'novel combinations of operations' as a 'type of mental activity specifically associated with consciousness' [23].

Formulating an agency indicator is challenging because accounts of agency vary widely [65]. Theorists from biology associate agency with autonomy and self-maintenance [66,67], AI researchers have recently focused on goal-directedness [68], and traditional philosophical views understand agency in terms of beliefs, desires, and intentions [69,70]. Theories of consciousness also differ in the forms of agency that they emphasize: in midbrain theory, consciousness supports a form that may be more basic than goal-directed or intentional behavior. Indicator AE-1 follows an approach that attempts to identify a key feature shared by animals and AI agents, which is that they can learn how to bring about goals more effectively through interaction with an environment [71,72]. This is a relatively novel proposal compared to other indicators, but such a proposal is needed to begin to synthesize disparate ideas about agency – and again, indicators can be revised in the light of new developments in theory.

(iv) Indicators should avoid ambiguous terms and contested concepts where possible, but without prematurely committing to precise specifications

Some concepts that are used in theories of consciousness are ambiguous in ways that are especially salient in the context of AI. For example, 'recurrence' usually refers to an algorithmic-level property in AI, as opposed to the implementation-level recurrence found in the brain in which neural connections form feedback loops. RPT-1 is formulated in terms of algorithmic recurrence (defined in 'Consciousness in AI' [11]) to avoid this ambiguity. Similarly, the AI context raises questions about the concept of embodiment, such as whether, and under what conditions, controlling an avatar in a virtual environment is sufficient for embodiment. AE-2 defines embodiment in a way that implies that this can be sufficient, motivated partly by the aim of finding a definition that is consistent with computational functionalism.

However, most current theories of consciousness remain underspecified [73] – they do not make perfectly precise claims about what it takes to be conscious – and this is rightly reflected in indicators. To make the indicators precise would involve anticipating possible uncertainties or controversies about how they should be applied then attempting to head these off in advance. But it will be more productive to work with indicators that reflect the current state of research and update them in response to future developments. Applying theories to AI through our method may help to motivate refinements to these theories (see the section 'Looking ahead').

## Finding indicator properties in AI systems

Theory-derived indicators can be used to make provisional assessments of the likelihood of consciousness in particular AI systems. However, determining whether such systems possess indicator properties will not always be straightforward. In this section we discuss two challenges that can arise in this process, again illustrated by examples from Table 1. The first challenge is that we do not have ready access to the representations and algorithms that trained deep neural networks use to perform tasks. We can make progress in uncovering these representations and algorithms using the techniques of mechanistic, or inner, interpretability [74], but these methods have significant

limitations at present. One example of an indicator that calls for the use of **interpretability methods** is RPT-2; the most direct way to determine whether a deep learning system uses organized and integrated perceptual representations would be to examine its inner workings. That said, it is also possible to imagine behavioral tests that would provide evidence of such representations. For example, susceptibility to the Kanizsa illusion (Figure 1) has been used in the research program that led to RPT and could provide evidence of integrated representations [26]. Several other indicators could potentially be probed using empirical studies – involving either mechanistic interpretability methods or behavioral tests – although it is also often possible to infer whether a system has an indicator property from knowledge of its training and architecture.

The second challenge is that, as we have mentioned, it can be a matter of interpretation whether AI systems possess indicator properties. For example, transformers are feedforward neural networks, so at first glance transformer-based LLMs lack algorithmic recurrence. However, one could argue that, when used autoregressively, they generate text using a feedback loop through the context window, with each feedforward pass adding one token. Arguably, this makes it seem that whether LLMs are recurrent depends on where we draw the boundaries of the system – should we include or exclude the context window? Various arguments could be made on this issue, but the point is that whether systems possess indicators can be debatable even if we understand their operation in detail and can turn on philosophical questions such as how to delineate the system in question.

## What does it tell us if a system possesses indicator properties?

We propose a broadly Bayesian attitude to indicators. Indicators are properties that should shift one's credence that an AI system is conscious. In addition to positive indicators, which are our focus here, **negative indicators** are also possible. Positive indicators increase the probability that the system is conscious, while negative indicators decrease it. That is, if E is the presence of the indicator and H is the system's being conscious, $p(H|E_p) > p(H)$ for positive indicators and $p(H|E_n) < p(H)$ for negative indicators.

Indicators can vary in their **specificity and sensitivity**. Focusing on positive indicators, an indicator is specific to the extent to which, in expectation, systems that have this property tend to be conscious. This is compatible with there being many conscious systems that lack it. An indicator is sensitive to the extent to which, in expectation, conscious systems tend to have this property, which is compatible with there being many non-conscious systems that also have it. The absence of a sensitive indicator tells us that a system is unlikely to be conscious; this is why indicators like RPT-1, algorithmic recurrence, may be useful. It is possible for indicators to be both highly specific and highly sensitive, but also for these attributes to come apart.

When we find evidence that a system possesses an indicator property, we should update our credence that it is conscious by conditionalizing on this evidence. The absolute amount of change will depend on one's prior credence that the system is conscious. It will also depend on credences in the theory T that links the indicator to consciousness because our indicators are, in the first instance, positive indicators relative to theories – formally, $p(H|E\&T) > p(H|T)$. One might also be uncertain about further relevant facts, such as whether the indicator is indeed present. Moreover, conditionalization is complicated by the fact that indicators need not be independent. In Table 1, some indicators entail or presuppose others, and some theories claim that sets of indicators are jointly sufficient for consciousness.

Once we have gathered all the evidence we can about a system, our credences that it is conscious should depend not only on our credences in the theories from which we derive indicators

but also on our credences in alternative theories and in possibilities that have not yet been described in theories ('unknown unknowns'). Sets of theory-derived indicators might leave out some necessary condition for consciousness – either a further computational condition or a requirement for a non-computational feature (Box 2).

Using our method makes sense if it provides indicators that can shift credences enough to have substantial practical significance. This depends on two conditions. First, one must have sufficient confidence in theories from which indicators can be derived. Second, it matters whether any of the theories' indicators are ever taken to be evidence against consciousness, perhaps by supporters of rival theories. This will not typically be the case, but if it is, how the indicators affect credences in consciousness will depend on credences in the opposing theories.

## Looking ahead

We anticipate productive interaction between research on the prospect of AI consciousness and more traditional neuroscientific consciousness research. As we have noted, progress in neuroscience should inform updated indicators. However, AI research may also contribute to stronger theories of consciousness. When researchers derive indicators from a theory and apply them to AI systems, they may reveal hidden ambiguities or unintended implications of the theory. Advocates of theories of consciousness may be especially motivated to clarify their views if they appear to imply that existing systems are conscious. For example, GWT advocates might explain whether they think that the system built to implement all four GWT indicators [5], which we mentioned above, is conscious. Moreover, AI systems that meet some indicators could be tested for capacities that are thought to be associated with consciousness, thus testing some of the predictions of the theories. For example, recent studies have used AI to test predictions of AST [75,76] and GWT [77,78]. More broadly, the mathematical precision of AI research and its approach to understanding systems through their architectures, objective functions, learning rules, and training data offer a framework that may lead to new insights in neuroscience [79], including the neuroscience of consciousness.

The use of theory-derived indicators to investigate consciousness in AI could also be one strand in a process of developing better tests for consciousness and validating their use in new populations. Developing such tests involves trialing new methods and extending existing ideas to new groups with the aim of establishing converging lines of evidence [10]. The method we propose could contribute to validating other assessment methods in the future as well as benefiting from validation itself. Validating our method in the AI case would be challenging because it would require the development of alternative assessment methods suited to AI; behavioral capacities could play a role here [80], but, the gaming problem (Box 3) makes matters more difficult. However, comparing the theory-derived indicator approach with other tests for consciousness in populations in which those tests are applicable could give evidence of its reliability.

Given that it may already be possible to build AI systems that possess many of the indicators, in looking ahead we should also contemplate the possibility that some near-future AI systems will be plausible candidates for consciousness. This would presumably have substantial ethical, legal, and social implications [81,82].

## Concluding remarks

Assessing AI systems for consciousness is challenging, but using scientific theories offers a principled, substantive method for doing so. We propose deriving indicator properties from scientific theories, then basing evaluations of the probability of consciousness in particular systems on whether they possess these indicators. The list of indicators can be revised as the

## Outstanding questions

How could the list of indicators in Table 1 be improved? This could involve adding indicators from other plausible theories of consciousness or stating the indicators in more detailed or more readily operationalizable terms, and could help to alleviate concerns about small network implementations or 'gaming' of the indicators.

Which of the indicator properties listed in Table 1 are displayed by existing AI systems, including frontier generative language or multimodal models, language agents, and deep reinforcement learning agents?

Can we develop quantitative or behavioral tests for consciousness in AI? These are challenging but would be valuable, and behavioral tests would make it possible to make assessments of consciousness in 'black-box' systems.

What are the implications of alternative approaches to consciousness, such as narrow biological views and IIT, for AI consciousness?

Can implementation of the specific features of consciousness contribute to the capabilities, reliability, or safety of AI systems?

How should research on consciousness in AI take into account the moral significance and potential social implications of this topic? In particular, how careful should researchers be in trying to avoid building systems that may be conscious?

science of consciousness progresses. As theories continue to be tested and refined, and as new theories are developed, the approach may be expected to provide increasingly plausible assessments.

Several lines of future research could provide further insights into the prospect of AI consciousness and identify complementary assessment methods (see Outstanding questions). New arguments for or against computational functionalism could help to provide clarity on whether AI consciousness is possible at all. Investigating in detail whether a representative sample of existing AI systems possess potential indicator properties – a project that has been begun [11] but is far from being completed – would both give a fuller picture of the current situation and help to refine the indicators. It is possible that interpretability methods could provide further evidence about indicators in particular systems or serve as the basis for distinct tests for consciousness. Since quantitative or behavioral tests for consciousness would have some advantages over our method if they were sufficiently reliable, investigating the prospects for such tests may be another important project. Finally, since **valenced conscious experience** is arguably especially morally significant [83,84], scientific research on these forms of experience may be crucial to understanding the moral status of some future AI systems.

## Acknowledgments

This project was supported by Effective Ventures and the EA Long-Term Future Fund. Y.B., A.C., G.D., and J.S. were supported by Open Philanthropy. J.S. was additionally supported by Fonds de Recherche du Québec (FRQ) grant 2023-NP-312582 and Conseil de Recherches en Sciences Humaines (CRSH/SSHRC) grant 430-2023-01017. D.C. was supported by Templeton World Charity Foundation grant 0561. A.C. was supported by European Research Council (ERC) grant (XSCAPE) ERC-2020-SyG 951631. E.E. was supported by a Vanier Doctoral Canada Graduate Scholarship. S.F. was supported by UK Research and Innovation (UKRI) under the UK government Horizon Europe funding guarantee (selected as ERC consolidator, grant 101043666). C.K. was supported by Templeton World Charity Foundation grant TWCF-2020-20539 and Australian Research Council grant DP240100400. T.B., L.M., M.P., and SF. were supported by CIFAR. R.V. was supported by ERC grant (GLoW) ERC-2022-ADG 101096017.

## Declaration of interests

P.B. has consulted for Anthropic and Conscium, R.L. has consulted for Anthropic, and J.B. has received research funding from Google. D.C. is a former member of the *Trends in Cognitive Sciences* advisory board and has given paid talks on consciousness to technology companies and other groups. A.C. has consulted for Verses AI. R.K. is a founder, shareholder, and the president of Araya, Inc. The other authors declare no competing interests.

## Resources

[i]https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

[ii]https://www.scientificamerican.com/article/what-does-it-feel-like-to-be-a-chatbot/

[iii]https://nautil.us/moving-beyond-mimicry-in-artificial-intelligence-238504

## References

1. Damasio, A. and Damasio, H. (2022) Homeostatic feelings and the biology of consciousness. *Brain* 145, 2231–2235
2. Aru, J. *et al.* (2023) The feasibility of artificial consciousness through the lens of neuroscience. *Trends Neurosci.* 46, 1008–1017
3. Seth, A.K. (2025) Conscious artificial intelligence and biological naturalism. *Behav. Brain Sci.* Published online April 21, 2025. https://doi.org/10.1017/S0140525X25000032
4. Goyal, A. and Bengio, Y. (2022) Inductive biases for deep learning of higher-level cognition. *Proc. R. Soc. A.* 478, 20210068
5. Dossa, R.F.J. *et al.* (2024) Design and evaluation of a global workspace agent embodied in a realistic multimodal environment. *Front. Comput. Neurosci.* 18, 1352685
6. Colombatto, C. and Fleming, S.M. (2024) Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* 2024, niae013
7. Shevlin, H. (2024) All too human? Identifying and mitigating ethical risks of social AI. *Law Ethics Technol.* 2024, 0003
8. Schwitzgebel, E. (2023) AI systems must not confuse users about their sentience or moral status. *Patterns* 4, 100818
9. Sebo, J. and Long, R. (2023) Moral consideration for AI systems by 2030. *AI Ethics* 5, 591–606
10. Bayne, T. *et al.* (2024) Tests for consciousness in humans and beyond. *Trends Cogn. Sci.* 28, 454–466
11. Butlin, P. *et al.* (2023) Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv* Published online August 17, 2023. https://doi.org/10.48550/arXiv.2308.08708
12. Seth, A.K. and Bayne, T. (2022) Theories of consciousness. *Nat. Rev. Neurosci.* 23, 439–452
13. Yaron, I. *et al.* (2022) The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nat. Hum. Behav.* 6, 593–604

14. Cao, R. (2022) Multiple realizability and the spirit of functionalism. *Synthese* 200, 506
15. Godfrey-Smith, P. (2016) Mind, matter, and metabolism. *J. Philos.* 113, 481–506
16. Seth, A. (2021) *Being You: A New Science of Consciousness*, Penguin
17. Birch, J. (2022) The search for invertebrate consciousness. *Noûs* 56, 133–153
18. Carruthers, P. (2019) *Human and Animal Minds: The Consciousness Questions Laid to Rest*, Oxford University Press
19. Mudrik, L. *et al.* (2023) Theories of consciousness and a life worth living. *Curr. Opin. Behav. Sci.* 53, 101299
20. Shevlin, H. (2021) Non-human consciousness and the specificity problem: a modest theoretical proposal. *Mind Lang.* 36, 297–314
21. Browning, H. and Veit, W. (2020) The measurement problem of consciousness. *Philos. Top.* 48, 85–108
22. Baars, B.J. (1993) *A Cognitive Theory of Consciousness*, Cambridge University Press
23. Dehaene, S. and Naccache, L. (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* 79, 1–37
24. Lamme, V. (2006) Towards a true neural stance on consciousness. *Trends Cogn. Sci.* 10, 494–501
25. Lamme, V. (2010) How neuroscience will change our view on consciousness. *Cogn. Neurosci.* 1, 204–220
26. Lamme, V. (2020) Visual functions generating conscious seeing. *Front. Psychol.* 11, 83
27. Dehaene, S. and Changeux, J.P. (2011) Experimental and theoretical approaches to conscious processing. *Neuron* 70, 200–227
28. Mashour, G.A. *et al.* (2020) Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105, 776–798
29. Lau, H. and Rosenthal, D. (2011) Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* 15, 365–373
30. Brown, R. *et al.* (2019) Understanding the higher-order approach to consciousness. *Trends Cogn. Sci.* 23, 754–768
31. Lau, H. (2022) *In Consciousness We Trust: The Cognitive Neuroscience of Subjective Experience*, Oxford University Press
32. Graziano, M.S. and Webb, T.W. (2015) The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol.* 6, 500
33. Graziano, M.S. (2019) *Rethinking Consciousness: A Scientific Theory of Subjective Experience*, WW Norton & Company
34. Crick, F. and Koch, C. (1990) Toward a neurobiological theory of consciousness. *Semin. Neurosci.* 2, 263–275
35. Francken, J.C. *et al.* (2022) An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neurosci. Conscious.* 2022, niac011
36. Block, N. *et al.* (2014) Consciousness science: real progress and lingering misconceptions. *Trends Cogn. Sci.* 18, 556–557
37. Graziano, M. (2017) The attention schema theory: a foundation for engineering artificial consciousness. *Front. Robot. AI* 4, 60
38. Dehaene, S. *et al.* (2017) What is consciousness, and could machines have it? *Science* 358, 486–492
39. Pennartz, C.M. *et al.* (2019) Indicators and criteria of consciousness in animals and intelligent machines: an inside-out approach. *Front. Syst. Neurosci.* 13, 25
40. Bayne, T. *et al.* (2023) Consciousness in the cradle: on the emergence of infant experience. *Trends Cogn. Sci.* 27, 1135–1149
41. Sneddon, L. *et al.* (2014) Defining and assessing animal pain. *Anim. Behav.* 97, 201–212
42. Crump, A. *et al.* (2022) Sentience in decapod crustaceans: a general framework and review of the evidence. *Anim. Sentience* 32, 1–35
43. Gibbons, M. *et al.* (2022) Can insects feel pain? A review of the neural and behavioural evidence. *Adv. Insect Physiol.* 63, 155–229
44. Nieder, A. (2022) In search for consciousness in animals: using working memory and voluntary attention as behavioral indicators. *Neurosci. Biobehav. Rev.* 142, 104865
45. O'Regan, J.K. and Noë, A. (2001) A sensorimotor account of vision and visual consciousness. *Behav. Brain Sci.* 24, 939–973
46. Vilas, M.G. *et al.* (2022) Active inference as a computational framework for consciousness. *Rev. Philos. Psychol.* 13, 859–878
47. Pennartz, C.M. (2022) What is neurorepresentationalism? From neural activity and predictive processing to multi-level representations and consciousness. *Behav. Brain Res.* 432, 113969
48. Merker, B. (2007) Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behav. Brain Sci.* 30, 63–81
49. Key, B. and Brown, D. (2018) Designing brains for pain: human to mollusc. *Front. Physiol.* 9, 1027
50. Elamrani, A. and Yampolskiy, R.V. (2019) Reviewing tests for machine consciousness. *J. Conscious. Stud.* 26, 35–64
51. Schneider, S. (2019) *Artificial You: AI and the Future of Your Mind*, Princeton University Press
52. Johnson, L.S.M. (2024) Entities, uncertainties, and behavioral indicators of consciousness. *J. Cogn. Neurosci.* 36, 1675–1682
53. Juliani, A. *et al.* (2022) The perceiver architecture is a functional global workspace. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society* (44), pp. 955–961
54. Herzog, M.H. *et al.* (2007) Consciousness & the small network argument. *Neural Netw.* 20, 1054–1056
55. Doerig, A. *et al.* (2021) Hard criteria for empirical theories of consciousness. *Cogn. Neurosci.* 12, 41–62
56. Michel, M. and Lau, H. (2021) Higher-order theories do just fine. *Cogn. Neurosci.* 12, 77–78
57. Shanahan, M. (2024) Simulacra as conscious exotica. *Inquiry* Published online December 1, 2024. https://doi.org/10.1080/0020174X.2024.2434860
58. Chalmers, D.J. (2024) Does thought require sensory grounding? From pure thinkers to large language models. *arXiv* Published online 18 August 2024. https://doi.org/10.48550/arXiv.2408.09605
59. Hohwy, J. and Seth, A. (2020) Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philos. Mind Sci.* 1, 3
60. Miller, M. *et al.* (2022) Predictive processing and consciousness. *Rev. Philos. Psychol.* 13, 797–808
61. Seth, A.K. and Hohwy, J. (2021) Predictive processing as an empirical theory for consciousness science. *Cogn. Neurosci.* 12, 89–90
62. Klein, C. and Barron, A.B. (2016) Insects have the capacity for subjective experience. *Anim. Sentience* 1, 1
63. Merker, B. (2005) The liabilities of mobility: a selection pressure for the transition to consciousness in animal evolution. *Conscious. Cogn.* 14, 89–114
64. Pennartz, C.M. (2018) Consciousness, representation, action: the importance of being goal-directed. *Trends Cogn. Sci.* 22, 137–153
65. Dung, L. (2025) Understanding artificial agency. *Philos. Q.* 75, 450–472
66. Barandiaran, X.E. *et al.* (2009) Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adapt. Behav.* 17, 367–386
67. Virenque, L. and Mossio, M. (2024) What is agency? A view from autonomy theory. *Biol. Theory* 19, 11–15
68. Kenton, Z. *et al.* (2023) Discovering agents. *Artif. Intell.* 322, 103963
69. Davidson, D. (2001) *Essays on Actions and Events*, Oxford University Press
70. Bratman, M. (1987) *Intention, Plans and Practical Reason*, CSLI Publications
71. Dretske, F. (1988) *Explaining Behavior: Reasons in a World of Causes*, MIT Press
72. Butlin, P. (2024) Reinforcement learning and artificial agency. *Mind Lang.* 39, 22–38
73. Mudrik, L. *et al.* (2025) Unpacking the complexities of consciousness: theories and reflections. *Neurosci. Biobehav. Rev.* 170, 106053
74. Rauker, T. *et al.* (2023) Toward transparent AI: a survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 464–483, IEEE Computer Society

75. Liu, D. *et al.* (2023) Attention schema in neural agents. *arXiv* Published online May 27, 2023. https://doi.org/10.48550/arXiv.2305.17375
76. Piefke, L. *et al.* (2024) Computational characterization of the role of an attention schema in controlling visuospatial attention. *arXiv* Published online May 8, 2024. https://doi.org/10.48550/arXiv.2402.01056
77. Devillers, B. *et al.* (2024) Semi-supervised multimodal representation learning through a global workspace. *IEEE Trans. Neural Netw. Learn. Syst.* 36, 7843–7857
78. Maytié, L. *et al.* (2024) Zero-shot cross-modal transfer of reinforcement learning policies through a global workspace. *arXiv* Published online March 7, 2024. https://doi.org/10.48550/arXiv.2403.04588
79. Richards, B.A. *et al.* (2019) A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770
80. Palminteri, S. and Wu, C.M. (2025) Beyond computational functionalism: the behavioral inference principle for machine consciousness. *psyarXiv* Published online February 6, 2025. https://doi.org/10.31234/osf.io/s7ptu_v2
81. Metzinger, T. (2021) Artificial suffering: an argument for a global moratorium on synthetic phenomenology. *J. Artif. Intell. Conscious.* 8, 43–66
82. Long, R. *et al.* (2024) Taking AI welfare seriously. *arXiv* Published online November 4, 2024. https://doi.org/10.48550/arXiv.2411.00986
83. Bentham, J. (1789) *An Introduction to the Principles of Morals and Legislation*, Payne & Son
84. Shepherd, J. (2018) *Consciousness and Moral Status*, Taylor & Francis
85. Block, N. (1995) On a confusion about a function of consciousness. *Behav. Brain Sci.* 18, 227–247
86. Nagel, T. (1974) What is it like to be a bat? *Philos. Rev.* 83, 435–450
87. Schwitzgebel, E. (2016) Phenomenal consciousness, defined and defended as innocently as I can manage. *J. Conscious. Stud.* 23, 224–235
88. Bayne, T. and Montague, M. (2011) *Cognitive Phenomenology*, Oxford University Press
89. Simon, J.A. (2017) Vagueness and zombies: why 'phenomenally conscious' has no borderline cases. *Philos. Stud.* 174, 2105–2123
90. Schwitzgebel, E. (2023) Borderline consciousness, when it's neither determinately true nor determinately false that experience is present. *Philos. Stud.* 180, 3415–3439
91. Lee, A.Y. (2023) Degrees of Consciousness. *Noûs* 57, 553–575
92. Birch, J. *et al.* (2020) Dimensions of animal consciousness. *Trends Cogn. Sci.* 24, 789–801
93. Naccache, L. (2018) Why and how access consciousness can account for phenomenal consciousness. *Philos. Transac. R. Soc. B Biol. Sci.* 373, 20170357
94. Frankish, K. (2016) Illusionism as a theory of consciousness. *J. Conscious. Stud.* 23, 11–39
95. Searle, J. (2017) Biological naturalism. In *The Blackwell Companion to Consciousness* (Schneider, and Velmans, , eds), pp. 327–336, Blackwell
96. Schuman, C.D. *et al.* (2017) A survey of neuromorphic computing and neural networks in hardware. *arXiv* Published online May 19, 2017. https://doi.org/10.48550/arXiv.1705.06963
97. Albantakis, L. *et al.* (2023) Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. *PLoS Comput. Biol.* 19, e1011465
98. Tononi, G. and Koch, C. (2015) Consciousness: here, there and everywhere? *Philos. Transac. R. Soc. B Biol. Sci.* 370, 20140167
99. Birch, J. (2024) *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*, Oxford University Press
100. Strathern, M. (1997) 'Improving ratings': audit in the British university system. *Eur. Rev.* 5, 305–321
101. Bender, E.M. and Koller, A. (2020) Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (58), pp. 5185–5198
102. Schwitzgebel, E. and Pober, J. (2024) The Copernican argument for alien consciousness; the mimicry argument against robot consciousness. *arXiv* Published online November 12, 2024. https://doi.org/10.48550/arXiv.2412.00008
103. Dainton, B. (2024) Temporal consciousness. In *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition) (Zalta, E.N. and Nodelman, U., eds)
104. Treisman, A. (2003) Consciousness and perceptual binding. In *The Unity of Consciousness: Binding, Integration, and Dissociation* (Cleeremans, A. and Frith, C., eds), pp. 95–113, Oxford University Press
105. Dennett, D.C. (1988) Quining qualia. In *Consciousness in Contemporary Science* (Marcel, and Bisiach, , eds), pp. 42–77, Oxford University Press
106. Lau, H. *et al.* (2022) The mnemonic basis of subjective experience. *Nat. Rev. Psych.* 1, 479–488
107. Graziano, M.S. *et al.* (2020) Toward a standard model of consciousness: reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognit. Neuropsychol.* 37, 155–172
108. Hohwy, J. (2013) *The Predictive Mind*, Oxford University Press
109. Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36, 181–204
110. Whyte, C.J. (2019) Integrating the global neuronal workspace into the framework of predictive processing: towards a working hypothesis. *Conscious. Cogn.* 73, 102763
111. Fleming, S.M. (2020) Awareness as inference in a higher-order state space. *Neurosci. Conscious.* 2020, niz020
112. Hurley, S.L. (1998) *Consciousness in Action*, Harvard University Press
113. Ginsburg, S. and Jablonka, E. (2019) *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*, MIT Press
114. Clark, A. (2008) *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press
115. McNamee, D. and Wolpert, D.M. (2019) Internal models in biological control. *Annu. Rev. Control Robot. Auton. Syst.* 2, 339–364