Neuroscience of Consciousness, 2026, **2026(1)**, niag002
https://doi.org/10.1093/nc/niag002
Published: 16 February 2026
Review Article

Beyond computational equivalence: the behavioral inference principle for machine consciousness

Stefano Palminteri[1,2,*] and Charley M. Wu[3,4,5]

[1]Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL, Research University, 29 rue d'Ulm, 75005, Paris, France
[2]Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale, 29 rue d'Ulm, 75005, Paris, France
[3]Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Alexanderstraße 10, 64283 Darmstadt, Germany
[4]Hessian.AI, Landwehrstraße 50A, 64293 Darmstadt, Germany
[5]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany

*Corresponding author. Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL, Research University, 29 rue d'Ulm, Paris 75017, France. E-mail: stefano.palminteri@ens.fr

Abstract

Large Language Models (LLMs) have rapidly become a central topic in AI and cognitive science, due to their unprecedented performance in a vast array of tasks. Indeed, some even see "sparks of artificial general intelligence" in their apparently boundless faculty for conversation and reasoning. Their sophisticated emergent faculties, which were not initially anticipated by their designers, have ignited an urgent debate about whether and under which circumstances we should attribute consciousness to artificial entities in general and LLMs in particular. The current consensus, rooted in computational functionalism, proposes that consciousness should be ascribed based on a principle of computational equivalence. The objective of this opinion piece is to criticize this current approach and argue in favor of an alternative "*behavioral inference principle*", whereby consciousness is attributed if it is useful to explain (and predict) a given set of behavioral observations. We believe that a behavioral inference principle will provide an epistemologically valid and operationalizable criterion to assess machine consciousness.

Keywords philosophy, theories and models, methodology, consciousness, artificial intelligence, computational modeling

Introduction

Large Language Models (LLMs) are a type of neural network characterized by their vast numbers of parameters and their capacity to learn from extremely large data sets. Today, they have taken the world by storm and have fundamentally reshaped how people think about artificial intelligence (AI) and what it is capable of. Combining the surprisingly effective "self-attention mechanism" (Vaswani et al. 2017) with human-in-the-loop reinforcement learning (Brown et al. 2020), LLMs have demonstrated remarkable performance across a staggeringly wide range of tasks. For instance, fooling humans in a Turing test (Bayne and Williams 2023, Jannai et al. 2023, Jones and Bergen 2024) or passing law school exams (Choi et al. 2021). And despite a number of key challenges, such as a surprising difficulty in solving rather simple abstract reasoning problems (e.g. the ARC Prize, Chollet et al. 2024, Moskvichev et al. 2023) or reliably reasoning about the mental states of people (i.e. Theory of Mind; Xu et al. 2024), researchers are increasingly using LLMs as models of human cognitive (Binz et al. 2024, Niu et al. 2024, Yildirim and Paul 2024) and neural processes (Schrimpf et al. 2021, Saanum et al. 2024).

While the testing and benchmarking of LLMs continues to generate a wealth of evidence about their strengths and limitations (Gandhi et al. 2023, Kıcıman et al. 2023, Moskvichev et al. 2023, Xu et al. 2024, Dettki et al. 2025), the wide availability of these tools has reached a larger audience than perhaps any other AI tool before it (Summerfield 2025). Indeed, everyone from journalists to politicians to Aunt Mary now has anecdotal evidence of the conversational abilities of LLMs and has begun to form their own opinions about how we should evaluate the consciousness of these systems (Colombatto and Fleming 2024). Thus, the question of whether an artificial system has consciousness holds immense political and societal sway, and is a topic where public opinions are already starting to form (Lenharo 2024, Palminteri and Pistilli 2025).

In the realm of philosophy and cognitive science, several perspectives have already been proposed about how to formally evaluate consciousness in LLMs (Butlin et al. 2023, LeDoux et al. 2023, Bayne et al. 2024, Evers et al. 2024). Although other approaches exist, such as phenomenology-first and biology-first ones (Polger 2019, Findlay et al. 2024, Block 2025, Seth 2025) the vast majority of these arguments are

Received 14 April 2025. **revised** 8 January 2026. **accepted** 9 January 2026

grounded in *computational functionalism*. According to computational functionalism, what defines a cognitive process are the computational operations that transform input variables into outputs, irrespective of the physical substrate implementing such computations, whether by neurons, transistors, or pencils on paper (Piccinini 2009). This allows for "multiple physical realizability" (Bickle 1998), where the same cognitive process can be physically realized by different material systems. Thus, an artificial system (the *candidate*) can be said to be conscious if it processes information by implementing the same computational processes (the *target computations*) that characterize consciousness in other systems already known to possess this capacity (the *reference*). In other words, an artificial system is considered conscious if it displays a form of *computational equivalence* with the reference system, specifically with respect to the target computations underlying the cognitive process under investigation.

For instance, a recent paper led by Patrick Butlin, Robert Long, and co-authored by seventeen other experts in the science of consciousness and AI follows this tradition (Butlin et al. 2023). In their comprehensive review, the authors conclude that AI systems, particularly LLMs, are not conscious because these systems do not *explicitly* execute several key computational processes that have been proposed by previous theories of consciousness, such as recurrent processing, a global workspace, attentional schema, and metacognition or predictive processing (Lamme and Roelfsema 2000, Koriat 2007, Dehaene et al. 2017, Graziano 2020, Hohwy and Seth 2020).

An equivalent position is taken by Susan Schneider in another collective piece (LeDoux et al. 2023), where, despite admitting the need for better behavioral tools to assess consciousness, she explicitly defines a necessary condition for machine consciousness such that:

> "*the system processes information in a way analogous to how a conscious human or non-human animal would respond when in a conscious state*."

These computational equivalence arguments all share the underlying logic: the target computations must be explicitly identified in the candidate system as a *necessary* condition for consciousness. Accordingly, the philosopher David Chalmers (2023) describes the general form of a typical computational equivalence argument as:

> "*LLMs lack C.*
> *If LLMs lack C, then they are probably not conscious."*

where *C* would be some computational process considered to be necessary for consciousness, such as recurrent processing, metacognition, or a global workspace (Butlin et al. 2023).

We agree that demonstrating computational equivalence can be a *sufficient* condition to ascribe consciousness to an artificial system. However, we argue that it should not be deemed a necessary condition. In other terms, we challenge the computational equivalence principle as the appropriate demarcation criterion between conscious and non-conscious entities, both in theory and in application.

To rewrite Chalmers' formulation, our alternative approach suggests:

> "*LLMs displays B.*
> *If LLMs displays B, then they are probably conscious."*

where *B* is some observable (i.e. *overt* and *inter-subjective*) pattern of behavior (which we define more precisely later), from which we are justified to infer the presence of an unobservable (i.e. *latent* and *hypothetical*), computational process C.

Our paper is structured as follows. We first show the limitations of the computational equivalence principle and motivate why a new approach is now more necessary than ever. We then provide arguments in favor of an alternative *behavioral inference principle,* which we believe is more consistent with the epistemological and methodological approaches used by cognitive scientists to study natural consciousness, and which is also more practically implementable—and more appropriate—given our current limited understanding of the mechanisms underlying the higher cognitive functions of LLMs.

Why computational equivalence can be sufficient, but not necessary

The goal of this paper is to argue that the principle of computational equivalence, as rooted in computational functionalism, is inadequate for attributing consciousness to artificial systems, such as LLMs (or very distant species, such as some arthropoda or mollusca; Birch 2025). However, as a disclaimer, we do not dismiss computational equivalence or functionalism as invalid or unimportant for other purposes. On the contrary, these principles have played a crucial role in cognitive science and philosophy, by detaching computational processes from their material substrates (i.e. multiple physical realizability; Bickle 1998). In fact, we endorse computational functionalism as a valid metaphysical framework for cognitive science, which has allowed great advances by conceptualizing cognitive processes as computationally defined forms of information processing (Piccinini 2009). We do not deny that other metaphysical frameworks beyond computational functionalism can and are applied in cognitive science, and in the science of consciousness in particular (Doerig et al. 2019, Tsuchiya et al. 2019, Ellia et al. 2021, Tononi et al. 2025). However, here we focus on computational functionalism since it can easily be integrated into the debate concerning artificial systems.

Thus, our contention lies not with computational functionalism itself, but with the idea that computational equivalence should be a necessary condition for attributing consciousness. To clarify, the attribution of a cognitive process based on computational equivalence consists of verifying whether the target computational architecture is implemented in the candidate system. This presupposes two conditions: (i) that there is an explicit hypothesis regarding the structure of the target computation, and (ii) that the computational architecture of the candidate system is transparent enough to allow one to verify the presence (or absence) of the target computation.

The first condition (requiring explicit hypothesis) is, at best, only partially fulfilled in the current and contentious landscape of consciousness research, where there is much disagreement about which computational processes underlie consciousness or even whether consciousness constitutes a unitary construct (see Frohlich et al. 2024, Gómez-Marín and Seth 2025, IIT-Concerned Klincewicz et al. 2025, Tononi et al. 2025 for recent debates). But even imaging a future where the scientific community arrives at a consensus, it should be stressed that the relevant computational processes are never known with the degree of certainty that the computational equivalence principle presupposes. They are hypotheses—educated inferences drawn from human behavioral phenomena, since the brain's computational architecture cannot be directly read.

The second condition of transparency is also not satisfied. LLMs are extremely complex black-box systems, with billions of parameters, whose lack of functional transparency parallels that of the human brain. Their impressive cognitive capacities arise from computational processes that cannot be directly inspected in the code or weights but must, again, be inferred from their behavior (Summerfield 2025).

Thus, in an ideal case where the target computations are known with reasonable certainty, and the candidate system's computational architecture is transparent, the computational equivalence principle could be applied as a sufficient criterion. Yet in practice—at least in the case of consciousness and LLMs—neither condition is met. As a result, computational equivalence does not represent a viable criterion for the attribution of consciousness. In the remainder of this article, we provide arguments in support of an alternative *behavioral inference principle,* which, as we will show, naturally emerges from the epistemology of cognitive science.

The epistemology of cognitive science

Our search for an alternative criterion begins from the recognition that cognitive science, like all empirical sciences, advances through a process of inductive inference, relying on iterative cycles of empirical corroboration (evidence supporting a theory) and falsification (evidence against a theory, thus requiring revision; Lakatos 1970, Meehl 1990). This involves using behavioral phenomena to infer plausible and instrumentally useful (e.g. for prediction and control; Harman 1965, Staddon 2021) *latent processes* that explain the observed behavior (subject to potential falsification by future evidence). In other terms, cognitive science, understood as an *empirical* interdisciplinary effort to study the mind, operates as a form of "methodological behaviorism" (Day 1983).

At this point, the reader might be surprised, given the widespread (but historically inaccurate) belief that cognitive science emerged as the antithesis of behaviorism (Leahey 1992). While it is true that cognitive science rejects radical forms of behaviorism (Schneider and Morris 1987), its methodology has progressively built upon behaviorist innovations (Simon 1992). There is perhaps no better way to illustrate this point than to defer to the words of Bernard J Baars, a key proponent of the global workspace theory of consciousness, in the introduction of his book "The Cognitive Revolution":

> *"Some of the central tenets of behaviorism are at this point so taken for granted that they have simply become part of standard experimental psychology. All modern psychologists restrict their evidence to observable behavior, [. . .]. In this way, we are all behaviorists." (*Baars 1986*)*

Thus, recognizing that even modern cognitive science is at its core grounded in behaviorism sets the stage for a crucial clarification (see Box 1 for counter-arguments to common critiques of behavior-based attribution). If cognitive science proceeds by explaining observable behavior in terms of latent constructs, then we must carefully distinguish between what is observed and what is hypothesized (Fig. 1). Nowhere is this distinction more important than the study of consciousness, where the temptation to treat conscious experience as the direct object of scientific inquiry is particularly strong. To avoid this conflation, we must reflect on the epistemological and ontological status of cognitive constructs like "consciousness" and their relation to the behavioral phenomena from which they are inferred.

Phenomena are explained by theory, but not *vice versa*

When seeking a criterion for attributing consciousness to LLMs, it is essential to clarify the epistemological status of consciousness in cognitive science. While we all have our own first-person subjective experience of consciousness, these experiences are private and not directly accessible for scientific scrutiny. Empirical science, in contrast, depends on intersubjective and publicly accessible facts. For this reason, consciousness is not an observable *behavioral phenomenon.*[1] What cognitive scientists can measure are patterns of behavior, such as body and eye movements, choices, reaction times, or even verbal reports (e.g. hetero-phenomenology; Cohen and Dennett 2011), which can be quantified and distilled into data. In contrast, consciousness is a *latent variable*, which is not directly observable in the data, but hypothesized to explain the phenomena (Bogen and Woodward 1988).

In other terms, the observable behavioral phenomena (and nothing else!) constitute "the thing that needs an explanation", or in Latin, the *explanandum* (Hempel and Oppenheim 1948). In contrast, latent cognitive constructs and theories are "the thing that explains", or *explanans*. The phenomena are explained by the theory, but not *vice versa* (Fig. 1).

Yet, in cognitive science, *explanandum* and *explanans* are frequently confused for one another. This conflation is particularly easy (and problematic) in the study of consciousness, for reasons that can be linked to the complexity of the subject, but also due to the fact that scientists experience the phenomenological existence of their own consciousness firsthand, thus making it counterintuitive to challenge the primacy of conscious experience (Metzinger 2021, Bayne et al. 2024). Consequently, scholars may lose sight of the fact that consciousness, as a cognitive construct, is not an *explanandum* or object of study in itself (at least not with empirical science). Rather, consciousness is an *explanans:* an unobservable, latent construct that is hypothesized in order to explain empirically observable behaviors.

It is, of course, an acceptable shortcut for an empirical scientist to say, "*I study consciousness*", provided we do not lose sight of the underlying implication. What she essentially means is:

> *"I study complex forms of behavior that justify the assumption of a latent, unobservable cognitive construct we refer to as 'consciousness'."*

A theoretically minded "consciousness" scientist will aim to describe this construct in formal terms and, with some measure of success, may even develop a valid computational model of consciousness—subject, of course, to potential falsification. But make no mistake: the ontological primacy of what constitutes the object of study is the behavioral phenomenon (the *explanandum*), not the resulting hypothetical computational process (the *explanans*).

Crucially, it is important to note that the primacy of behavior is not only ontological but also "historical." An initial consensus on what counts as a behavioral manifestation of the cognitive process of interest is, in fact, necessary before one can undertake the task of its computational characterization. For instance, in the case of

[1] Another crucial type of observables consists of experimental variables, such as visual, auditory, or written stimuli, which are typically manipulated by researchers to generate or control behavioral phenomena (in the cognitive scientist's laboratory, experimental variables usually take the form of behavioral tasks). These observables—both behavioral and experimental—are then employed to construct and validate hypotheses regarding latent cognitive constructs and the processes underlying their relation.

Box 1. Attributing cognitive processes based on behavioral observations.

While behavioral criteria for attributing cognitive processes have historically had an intuitive appeal for empirical scientists (Lashley 1923, Skinner 1965), they have been criticized by philosophers of mind (Blanshard 1939, Putnam 1960, Block 1981). Here, we summarize some of these historical criticisms and clarify how our *behavioral inference principle* (*If an agent displays behavior **B**, then it probably possesses cognitive process **C***) avoids these arguments.

The first class of critiques target "false positives," where *B* can occur without *C.* Ned Block's influential Blockhead thought experiment (Block 1981) imagines a machine that passes the Turing test (Turing 2009), not because it possesses intelligence, but simply because all possible responses have been pre-programmed, relegating the machine's role to merely retrieving the correct response. Searle's Chinese Room (Searle 1980) illustrates a similar point, where a person in a room using a rulebook to manipulate Chinese symbols is imagined to produce fluent responses without understanding the language. These thought experiments illustrate the logical possibility of displaying behavior B (fluent conversation) without possessing cognitive process C (intelligence or knowledge of Chinese). However, empirical science—as opposed to mathematics and philosophy—is concerned with physical rather than merely logical possibilities. A machine with infinite memory for all possible responses pre-coded is physically infeasible. Even if it were, the retrieval and response times would be infinitely long (Shannon 1948), making the machine unable to demonstrate fluent, real-time conversation. Thus, a scientist waiting an eternity for the machine's responses would be justified—on purely behavioral grounds—in rejecting the machine as demonstrating a genuine form of intelligence. Similarly, the occupant of the Chinese room, searching through an astronomical number of rules, would also fail to demonstrate a speed of response consistent with a fluent speaker. Even if the rulebook were entirely internalized, the sheer volume of rules involved would render this mechanism implausible from both the perspective of memory and fluency. Conversely, a scientist receiving prompt and sensible responses to virtually any question would be justified to infer that the machine or the room is truly intelligent (subject to integrating the behavioral evidence with prior expectations; Eq. 1).

The second class of "false negative" critiques is exemplified by the Super-Super-Spartans thought experiment proposed by Putnam (1961), where in a parallel universe, Spartans have been trained to successfully suppress all involuntary and voluntary external manifestations of pain, even though they feel and dislike pain just like us. Imagine that in such a parallel universe, a scientist from Athens is sent to study pain in the Super-Super-Spartans by administering various pain-inducing experiments. She diligently conducts the experiments and receives no empirical, behavioral evidence of pain from her subjects. The Athenian scientist, based on the available behavioral evidence, therefore concludes that the Spartans do not experience pain, which we as omniscient observers know is false. This thought experiment successfully demonstrates that, theoretically, it is possible to possess cognitive process C without displaying behavior B. However, from a scientific perspective, we must agree with the conclusions reached by the Athenian scientist, who made the correct inference based on the available evidence. Of course, should the Athenian scientist return to Sparta equipped with advanced neural recording devices, she would eventually revise her conclusions after detecting neural markers of pain. In other terms, for the proposition "Super-Super-Spartans experience pain" to be true, it must translate into some intersubjective physical evidence at some level of observation. Another instance of "false negative" argument is represented by "philosophical zombies" (Kirk 1974, Chalmers 2009), hypothetical beings who are indistinguishable from us with the exception of lacking conscious experience. While they have proven divisive among philosophers, by any standard of empirical cognitive science, zombies do not pose a major challenge. Even though the thought experiment requires us to accept the counterintuitive (and seemingly unprovable) claim that they lack conscious experience, from a scientific perspective, they are not puzzling at all: all evidence points to them being conscious.

Thus, while these thought experiments highlight logical possibilities, they do not undermine the use of behavioral criteria when applying the scientific method of cognitive science, which operates on evidence-based inferential logic. Thus, our *behavioral inference principle* avoids both false positive and false negative scenarios by adopting the flexibility of inductive reasoning, grounded in the epistemological fact that cognitive processes are theoretical constructs useful for explaining particular classes of behavioral observations, not objects of study in themselves.

consciousness, an initial agreement on behavioral tasks—such as visual masking, binocular rivalry, continuous flash suppression, or metacognitive ratings—is required in order to claim that these paradigms can provide insights into the computational mechanisms of consciousness. This illustrates, time and again, that behavior precedes theory.

Before the scientific enterprise of characterizing the computational form of consciousness, everyday folk attribution already proceeds on the basis of behavioral observation, together with prior expectations. The reason we know that another human subject is conscious in a given moment (e.g. sleeping or awake) is not because we directly perceive computational equivalence. In fact, we often have relatively little insight into our own internal computations (Chater 2018), much less those of others. Rather, we observe behavior exhibiting certain features and complexities that are most coherently explained by assuming the latent construct known as "consciousness." In other words, we attribute consciousness to others on behavioral grounds (although, as we discuss later, prior expectations also play a key role).

A very similar point is made quite eloquently by Gilbert Harman, who takes the attribution of mental states as a case study to exemplify and explain the principle of inference to the best explanation:

> *"[. . .] when we infer from a person's behavior to some fact about his mental experience, we are inferring that the latter fact explains better than some other explanation what he does"* (Harman 1965)

Translating to the theme of consciousness in artificial systems, the priority should not (and cannot) consist in verifying whether a machine's code *presents* a set of computations that a group of scientists has proposed as representing the latent process underpinning conscious behavior. Instead, the focus should be on whether the machine exhibits a specific pattern of behavior that allows (and compels) us to *infer* a latent computational process that we agree to call consciousness, given the available data and background knowledge. In the case of LLMs, the relevant behavior is their language output, which—although very different from human verbal behavior in terms of physical medium and mode of generation—nonetheless shares key features. Most importantly, it consists of responses elicited in specific contexts by given stimuli, depends on some internal information-processing architecture, and, crucially, is intersubjective and measurable.

In other terms, considering computational equivalence as a necessary condition for attributing consciousness to artificial systems does

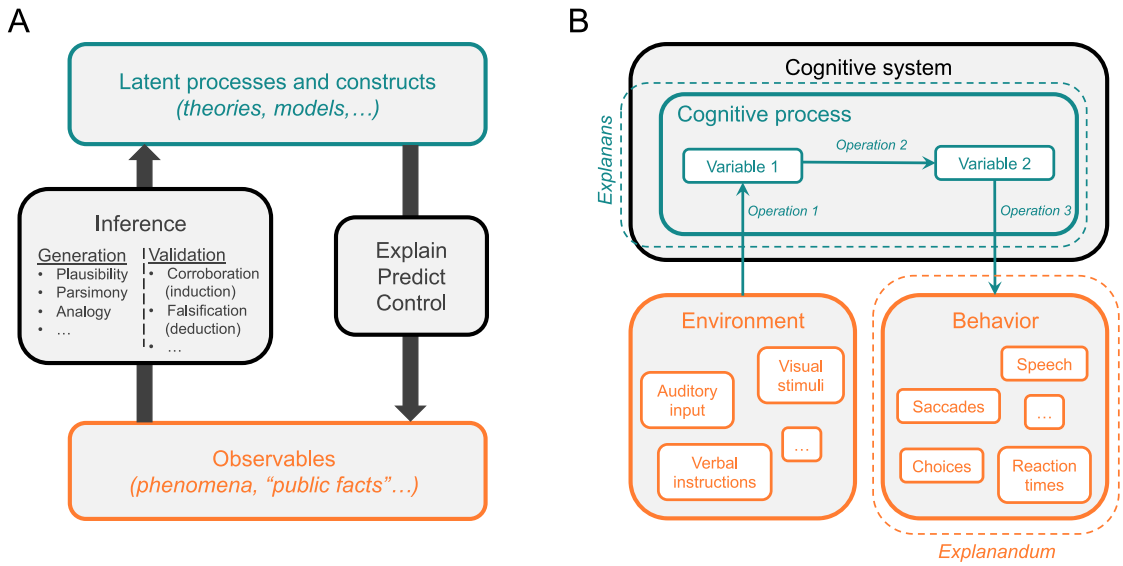

Figure 1 The epistemological status of cognitive (or mental) constructs. A) the relation between observables (e.g. phenomena and public facts) and latent processes (e.g. theories and models). Inference typically follows an inductive process, where hypotheses about latent processes are initially generated using heuristics such as plausibility, parsimony, and analogy (among others), and are later validated through cycles of induction-based corroboration and deduction-based falsification. Theories and models, in turn, can be used to explain, predict, and control both past and future observations. B) the relation between a cognitive process (*explanans*: The thing that explains), the behavioral phenomena (*explanandum*: The thing that needs an explanation), and the environment (i.e. experimental factors). The spatial organisation and colour scheme are the same in panels A and B: the explanans (latent processes) are shown at the top and coloured in teal, whereas the explanandum (observable phenomena) are shown at the bottom and coloured in orange.

not align with how the target computational processes themselves have been discovered in humans (or other animals) in the first place (Butlin et al. 2023, LeDoux et al. 2023, Evers et al. 2024). Rather, these processes were identified through inferences to the best explanation of behavioral phenomena produced by an otherwise architecturally opaque computational substrate: the brain (Lipton 2004, LeDoux et al. 2023, Bayne et al. 2024, Birch 2025, Negro and Mudrik 2025).

What about neural phenomena?

While we focus on overt behavior, we do not deny the relevance of neural phenomena in the attribution of consciousness. Once measured with appropriate techniques (e.g. neuroimaging, electrophysiology), neural activity is itself an observable, public fact much like behavior, and can therefore contribute to the overall inferential process.

Neural evidence is obviously key in cases where consciousness may exist despite the absence of overt behavioral markers, as in severe locked-in patients (or super-super Spartans; see Box 1). Of note, while locked-in patients are often taken as paradigmatic cases of dissociation between behavior and consciousness, in practice, their detection still relies on minimal behavioral responses (e.g. eye movements). However, in some severe cases where oculomotor responses are not possible, neural recordings (e.g. electroencephalographic patterns overlapping with those of conscious individuals) can provide a sufficient sign of consciousness.

A few qualifications follow. First, neural evidence does not alter the overall epistemological framework: neural phenomena are useful insofar as they expand the set of observables that must be explained by the latent process, and symmetrically, can contribute to inferences about the computational form of that latent process. Second, they neither require nor support a strict computational equivalence principle, since neural markers are generally identified by correlation with behavior, not by reference to an explicit computational mechanism. Once again, consensus on behavioral markers must therefore precede the recognition of corresponding neural markers (e.g. alpha oscillations associated with sleep). Third, while neural markers may be sufficient in some cases, they are not strictly necessary. Many sophisticated cognitive processes—such as learning, imitation, communication, and possibly consciousness—are found in species with brains very different from ours (Wong 2025) or even no brain at all in the case of single-celled organisms (Gershman et al. 2021).

Thus, neural evidence may serve as a sufficient but not a necessary criterion for consciousness. Furthermore, neural phenomena are often considered secondary to behavioral phenomena in inferring computational mechanisms (Niv 2021), since behavior can both corroborate and falsify a cognitive model, whereas neural recordings typically serve only to corroborate. In humans, for example, neural markers are often used atheoretically (e.g. in the detection of locked-in patients; Adama and Bogdan 2025) or as a means of providing external validity to computational processes already specified by behavior (e.g. model-based fMRI; O'Doherty et al. 2007, Gläscher and O'Doherty 2010, Wilson and Niv 2015, Lebreton et al. 2019). In neither case does neural evidence permit a direct "discovery" of computations, nor do they enable a genuine assessment of computational equivalence. For these reasons, and for the sake of parsimony, we will restrict our discussion to behavior as the primary object of empirical investigation in cognitive science (Niv 2021).

The behavioral inference principle

Here, we provide arguments in support of the *behavioral inference principle* as a criterion for attributing consciousness (see **Supplementary**

discussion and Fig. S1 for more details on this issue). We defend it both on both epistemological and on pragmatic grounds, as it is more applicable and better suited to black-box systems such as LLMs. We also discuss how our priors about hypotheses should influence the way we interpret behavioral evidence for or against consciousness attribution, and how the behavioral inference principle relates to other criteria, differing in both method and applicability.

The crucial role of priors in the behavioral inference principle

The behavioral inference principle relies on a form of inductive reasoning, which is concerned with forming belief about the probability of a cognitive process (consciousness, C) from behavioral evidence (B). This can be described as a form of Bayesian inference:

$$P(C|B) = \frac{P(B|C) * P(C)}{P(B)} \quad (1)$$

Here, the posterior $P(C|B)$ represents our inferred belief about whether the system possesses consciousness (or any other latent property C), conditioned on the observed behavioral evidence B. This posterior is proportional to the likelihood of the data $P(B|C)$, while $P(C)$ and $P(B)$ represent priors. How can these terms be interpreted in the context of consciousness attribution? The likelihood $P(B|C)$ captures how probable the observed behavior is, assuming that the cognitive process is present. The dependence of the posterior $P(C|B)$ on the likelihood is evident, and in the practice of cognitive science this term is usually at the basis of any model comparison criterion (Wilson and Collins 2019). Importantly, Bayesian model comparison also naturally implements a form of Occam's razor, where simpler hypotheses are favored over more complex alternatives with equivalent explanatory power (Myung and Pitt 1997, Blanchard et al. 2018).

But of course, priors are also important. Specifically, $P(C)$ quantifies how probable we believe the cognitive process (C) is in the system. Crucially, this prior is (and must be) influenced by all prior knowledge we have concerning the system under investigation. For instance, this value is effectively $P(C) = 1$ in the case of other human beings, where the presence of consciousness is not in question. This justifies why we content ourselves with very scant behavioral or neural evidence when attributing consciousness to other people, even locked-in patients. This value is understandably very high for non-human primates, because their deep physiological and behavioral similarity with humans justifies a high degree of generalization (Kemp and Tenenbaum 2009, Wu et al. 2024). In contrast, we generalize much less for more distant species, such as mollusks or arthropods, which are physiologically and behaviorally "alien". Following this line of reasoning, we are justified in being highly conservative when attributing consciousness to LLMs, because they are physically radically different from us, and we have no known example of a conscious artificial system. Of course, priors based on computational architecture are also important (Wong 2025). On one hand, the science of consciousness provides several educated guesses concerning the probable computational architecture of consciousness; on the other hand, even though the post-training weight architecture of LLMs is huge, opaque, and uninterpretable, they do come with architectural constraints that may be informative. Take, for example, the fact that recurrence seems to be key to many computational theories of consciousness (Lamme and Roelfsema 2000, Dehaene et al. 2011, Tononi et al. 2016). If we adhere to these theories, we might inform our inferential process so that $P(C)$ is considered higher in systems that explicitly feature recurrence (Gu and Dao 2023, Peng et al. 2023, Sun et al. 2023, Dao and Gu 2024).

$P(B)$, on the other hand, quantifies the baseline probability of observing the behavior of interest, regardless of the presence of the cognitive process. This term is probably why LLMs have sparked such intense debate about consciousness both outside and inside academia. The fact is that LLMs possess full conversational capacity, as well as "flickers" of metacognition and Theory of Mind (Birch 2025). Historically, we encountered this combination of behaviors (i.e. conversational capacity coupled with higher cognitive functions) only in conscious beings, making $P(B)$ a very low value. This is likely what motivated naive misattributions of consciousness in cases like Blake Lemoine and other users (Palminteri and Pistilli 2025). However, priors change as we acquire new information, and as LLMs proliferate and we accumulate evidence concerning their mechanisms, we are slowly beginning to accept that full conversational proficiency—even when combined with higher cognitive functions—may not be a perfect behavioral marker for consciousness (Sejnowski 2023). We are nonetheless left with the task of determining which behavioral observations we should deem informative in the case of LLMs with respect to this inferential process.

Thus, looking at the behavioral inference principle through the lens of Bayesian inductive inference allows us to understand how the consciousness attribution problem can rely on the same stream of behavioral evidence for both natural and artificial systems. However, it can still yield different interpretations for different systems, because different amounts of behavioral evidence are required to override different prior beliefs.

Situating the behavioral inference principle in the context of other criteria

In situating the behavioral inference principle within the broader context of other possible attribution criteria, we must return to two key factors: (i) whether there exists an *explicit* (and ideally consensual) hypothesis concerning the computational mechanisms underlying the cognitive process of interest, and (ii) the degree of *transparency* in the system under evaluation (Fig. 2).

The first question to ask is whether there is an explicit hypothesis regarding the computational form of the cognitive process under consideration. This maps onto the distinction between theory-heavy and theory-light approaches. In our argument, theory-light and theory-heavy approaches are defined with respect to the necessity of possessing an explicit computational hypothesis about the cognitive process under investigation (in our case, consciousness). Theory-heavy approaches require a well-defined hypothesis concerning the computational structure of the target cognitive process, whereas theory-light approaches are more agnostic and do not require the target process to be explicitly specified or agreed upon at the computational level.

If an explicit (and consensual) hypothesis exists concerning the computational mechanisms of the cognitive process at stake (which is only partially the case for consciousness), we can then ask whether the computational architecture of the candidate system is sufficiently transparent to allow for direct observation and verification. In cases where the system is fully transparent (e.g. the brute force chess algorithm DeepBlue), computational equivalence can be applied. These systems, being fully interpretable, allow for a direct comparison of the computational processes in the candidate and reference

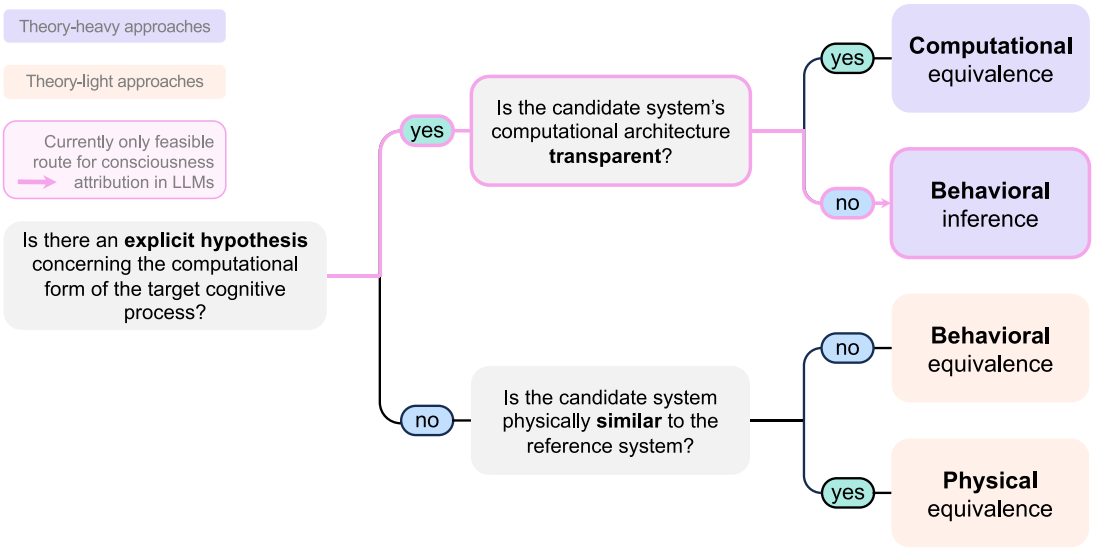

Figure 2 A diagram illustrating how initial conditions map into the applicability of different attribution principles. Theory-heavy criteria require explicit hypotheses about the computational form of the cognitive process under investigation, whereas theory-light approaches do not. Computational transparency refers to our capacity to read and interpret the computational operations in the candidate system. Complex systems such as human brains and LLMs are too complex to be transparent, whereas the fully mapped brain of *Caenorhabditis elegans* and the brute force DeepBlue chess algorithm could be considered examples of transparent systems. Physical similarity is a gradient within the animal kingdom when compared to the conscious species *par excellence* (*Homo sapiens*): Some species are very similar (e.g. great apes), whereas others are less so (e.g. octopi). In pink, we highlight what is currently the only feasible route for consciousness attribution to LLMs, optimistically assuming that current computational theories of consciousness are valid. Of note, we do not mean to suggest that behavioral inference is the *only* way consciousness should be attributed; for instance, we believe that a high degree of computational equivalence or physical similarity, if convincingly demonstrated, may also be sufficient.

systems. Of note, computational equivalence is also warranted—and is in fact the only possible route for attribution—if one accepts that the defining properties of consciousness depend on the form of the computational operations rather than on their functional role, as proposed in approaches such as Integrated Information Theory (Tononi et al. 2016).

However, computational transparency is a rare case. For most complex systems, humans and LLMs included, the system's computational mechanisms are not transparent. For these non-transparent systems, we cannot directly observe or verify the computational processes, and instead we must infer them from behavioral phenomena (behavioral inference principle) (Eq. 1).

If there is no explicit hypothesis regarding the computational mechanisms of the cognitive process, we must rely on theory-light approaches and turn to *physical similarity* as a possible criterion for attribution. In such cases, systems with high physical similarity between the candidate and the reference systems may justify the attribution of consciousness on the basis of strong prior expectations alone. This is the rationale why we are probably justified to attribute consciousness to Neanderthals, whose genetic and anatomic similarity provides a strong basis for assuming similar cognitive mechanisms, even in absence of any behavioral observation. Crucially, our position is distinct from biology- or "meat-first"-approaches, which suggest that at least some degree of physical similarity is not only sufficient, but also necessary for attribution (Block 2025, Seth 2025).

In the absence of both consensus and physical similarity, we are left with behavioral equivalence as the only viable criterion. Behavioral equivalence is a theory-light approach that relies solely on observable behavioral similarity between the candidate and reference system. If behavior is sufficiently similar, we can infer that the system is likely to share similar cognitive processes, although the exact nature of these processes is left theoretically under-specified and open-ended (Turing 2009). The application of this criterion to LLMs is complicated by our limited ability to identify truly diagnostic behavioral phenomena, given that these systems are trained to display traits that are usually associated with conscious processes—a problem often described as "gaming" or "mimicry" (Birch 2025). This difficulty reflects the well-known adage that a measure ceases to be a good measure once it becomes a target (Goodhart's law).

Several clarifications follow from this decision tree. First, behavioral inference should not be confounded with behavioral equivalence, as the former is theory-heavy while the latter is theory-light. More specifically, the behavioral inference principle involves inferring putative underlying computational processes from observed behavior, which are then analyzed and compared to what we currently believe to be the relevant processes in the reference system. In this sense, the behavioral inference principle still requires explicit computational hypotheses. These hypotheses are understood as valid explanations of the behavioral observations, rather than as processes *literally* implemented in the system. In contrast, behavioral equivalence is atheoretical: it bypasses hypotheses about computational mechanisms and treats similarity of behavior as sufficient in itself. To circumvent the mimicry or gaming problem, the behavioral inference principle, first incorporates into its Bayesian structure the baseline probability of displaying consciousness-like behavior (e.g. fluent conversation) as a consequence of training, and second, relies on behavioral tests that are specifically designed to be diagnostic of the computational process of interest (Bayne et al. 2024).

Second, the applicability of these criteria to given cognitive processes and candidate systems will evolve as science progresses. For example, fundamental research in cognitive science is continually bringing new insights into the computational mechanisms underlying

consciousness, allowing new consensuses to emerge and opening the door to more theory-heavy forms of attribution. Likewise, ongoing efforts in LLM explainability are gradually making these models less opaque, thereby increasing the feasibility of computational equivalence assessments.

Finally, note that this classification is not intended to be strictly normative (i.e. to prescribe which criterion should be applied), but rather to illustrate the range of possibilities and how they map onto critical features of the attribution problem, such as the computational transparency of the candidate system (i.e. the system under evaluation for consciousness) and its physical constitution with respect to the reference system (i.e. systems in which we know that consciousness exists). The choice of attribution criterion cannot be determined solely by scientific and philosophical debate. It will also inevitably reflect the beliefs and preferences of the community vis-à-vis the metaphysical status of cognitive processes. For example, those who endorse computational functionalism as the correct metaphysical framework for cognitive science may argue that physical equivalence is unnecessary (Dennett 1991), while others may disagree and continue to see physical similarity as indispensable (Block 1978, Seth 2025).

Why do we need to attribute consciousness?

Having laid out our arguments against computational equivalence and in favor of a behavioral inference principle, we would like to clearly state that, in our opinion, even the most sophisticated current LLMs (although they may exhibit some behaviors typically associated with consciousness, such as language understanding and production) are (very likely) not conscious (Butlin et al. 2023, Birch 2025). Notably, LLMs lack continuity (each new interaction begins anew), coherence (they can impersonate an unlimited number of personas on command), multisensory integration (so that, for instance, a notion of workspace could hardly be implemented), and embodiment (they only interact through language)—all features that many people have considered to be important for consciousness (Bayne et al. 2024). But, we also concur with Butlin, Long, and colleagues (Butlin et al. 2023) that the engineering steps required to develop LLMs that exhibit behaviors consistent with more complex forms of consciousness are not insurmountable—and may even be simpler than those accomplished so far.

However, it's important to discuss why we need to be able to attribute cognitive constructs or latent processes, such as consciousness, in the first place. In empirical science, and cognitive science perhaps most of all, it is a widely held understanding that "all models are wrong, but some are useful" (Box 1976). Models or theories provide *epistemic* value by formally explaining some behavioral phenomenon, thus helping us understand the world. However, models and theories also crucially provide *instrumental* value in informing us how to act better in predicting and controlling important factors in our world. Indeed, much of the interest around artificial consciousness is precisely motivated by the instrumental need to act correctly vis-à-vis the ethical questions related to the creation of such systems, along with the inherent rights and responsibilities they may acquire (Hildt 2019, Bengio et al. 2023, Wong 2025). These ethical questions usually have two complementary faces.

The first ethical question is related to the problem of control and potential existential harm that extremely powerful artificial agents can cause to the human race (Bengio et al. 2023). The questions of consciousness and danger are often confounded because it is generally assumed that a conscious AI will also be extremely intelligent and self-driven. However, the two things are not necessarily linked: an AI could be extremely "intelligent" in its capacity to achieve its goals, but not conscious (e.g. the famous paperclip maximizing robot; Miller et al. 2020). Furthermore, goals and motivations do not necessarily require high levels of consciousness. Many typically lower-level organisms can be said to have goals and motivations (mainly linked to self-preservation), even single-celled organisms (Gershman et al. 2021). Thus, the control problem and other existential AI safety issues are perhaps better addressed not by discussing and regulating consciousness, but rather their capacity for agency (Wong 2025) vis-à-vis their ability to influence the world around us (both digital and real).

The second ethical issue that often fuels the debate on machine consciousness (e.g. the clamorous case of Blake Lemoine's resignation from Google; Lemoine 2022) concerns the potential for these entities to acquire *moral status*, which is the degree to which an organism deserves ethical consideration (Nussbaum 2006, Singer 2009, Birch 2024). However, it is unclear whether consciousness *per se* is the appropriate criterion. Taking the example of non-human animals (DeGrazia 2002), many ethical theories require not only some degree of awareness, but also the ability to demonstrate an understanding of "pleasure" and "pain," or, broadly speaking, to show strong preferences regarding possible experiences or world states (Birch 2024), which at a minimum would constitute a demonstrated preference for one's own existence over non-existence (Singer 1989). Currently, most theories of consciousness are silent regarding notions of pleasure, pain, and preferences, which are of fundamental importance for moral status. Meanwhile, pleasure, pain, and preferences are the cornerstone of reinforcement learning (RL) algorithms (Eldar et al. 2016, Sutton and Barto 2018, Watson et al. 2019), leading to arguments that RL agents may already possess a non-zero moral status, even in the absence of consciousness (Tomasik 2014). In this regard, we believe caution may be warranted as LLMs are coupled with goal-directed RL algorithms to improve planning and control, as these systems may increasingly display behavioral patterns we are likely to attribute to a conscious agent.

The elephant in the room

Thus, we are not proposing some single and final behavioral "litmus" test for consciousness akin to the "Turing test" (Turing 2009). As we encounter or even construct new subjects of study (e.g. LLMs), we will continually need to develop new tests and experiments to refine our inferences about the underlying computational processes. These methods must be adapted to the system under investigation, as we are increasingly realizing it is necessary, for instance, when assessing consciousness beyond the mammalian case (Schneider 2020). As long as LLMs remain restricted mainly to language processing, the most promising behavioral markers will likely consist of context-sensitive linguistic exchanges (Gui et al. 2020) and adaptive problem-solving across varied domains (Tian et al. 2024, Pan et al. 2025), as opposed to behaviors that are unlikely to be informative about consciousness, such as rote repetition of training material and phenomenological mimicry (Birch 2025). As LLMs evolve toward more multimodal systems, the ability to integrate information across modalities will probably become an important dimension to assess. To avoid premature (mis)attribution, our behavioral criteria must be stringent, and we believe that multidimensional benchmarks are necessary within the behavioral inference approach (Yin et al. 2023, Li et al. 2024).

Figure 3 A metaphor. Blind monks examining an elephant by Hanabusa Itchō (1652–1724). This image is in the public domain.

But the science of machine consciousness must nonetheless embrace the behavioral methodology of cognitive science in doing away with "necessary and sufficient conditions", and being willing to continually evolve through inference, corroboration, and falsification, in order to develop new empirical standards and operational definitions of consciousness (or other latent cognitive capacities). Thus, arriving at a consensus about what behavioral evidence warrants the attribution of consciousness to any given system will inevitably evolve as new data is acquired, new theories are proposed, and new goals are set (Mitchell and Krakauer 2023, Sejnowski 2023).

To conclude, in the cognitive science community, the question of consciousness in LLMs has become the "elephant in the room". Yet, we are reminded of a different metaphor, also involving an elephant, but one being examined by blind Buddhist monks (Fig. 3). In the parable, one monk touches the trunk and believes it to be a snake, another feels the ear and imagines it to be a fan, while a third, grasping a leg, concludes it is a tree. The story illustrates the challenges of identifying something complex and multifaceted that cannot be directly perceived as a whole when working with limited and fragmented information. The monks would only be able to arrive at the correct conclusion (that they are examining an elephant) if they could gather sufficient data and communicate their findings. Even then, without tools like MRI or genetic analysis, their conclusion would only represent the most probable explanation based on the available (tactile) evidence. Similarly, when it comes to artificial consciousness, there will unlikely be a single, definitive piece of evidence that conclusively demonstrates consciousness in machines. Instead, we may see a gradual accumulation of behavioral features that increasingly suggest the presence or absence of consciousness. Importantly, the absence of a single "litmus test" (akin to a Turing test for consciousness) does not entail that behavioral inference is invalid: even within computational functionalism, unless dramatic advances are made in the mechanistic explainability of LLMs, multiple and converging behavioral evidence remains the only viable route for inference. As "blind" cognitive scientists, our task will be to critically and fairly evaluate this growing body of evidence and decide whether it is sufficient for us to attribute consciousness to these systems, given our current understanding concerning the mechanisms underlying this cognitive process.

Indeed, the monks themselves are applying an inductive methodology, trying to infer the best possible explanation for their observations. Through touch, they are conducting experiments to the best of their ability with the phenomena at hand. However, the reason it serves as a comical parable is that the monks fail to integrate their own conclusions at the group level and come to a consensus. In contrast, the success of cognitive science is *via* debate, and above all else, dialogue and collaboration.

Acknowledgements

SP thanks Bahador Bahrami and Ophelia Deroy for suggesting that he might be a “closet philosopher”. The authors wish to thank Valeria Giardino, Lindsay Drayton, Steve Fleming, Patrick Butlin, Catherine Tallon-Baudry, and Hongyu Wong for their helpful comments and discussions. We thank Romane Cecchi for help and advice concerning figures’ preparation.

Author contributions

Stefano Palminteri (Conceptualization [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]) and Charley M. Wu (Conceptualization [supporting], Visualization [supporting], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at *Neuroscience of Consciousness* online.

Conflict of interest

None declared.

Funding

S.P. is supported by the European Research Council (RaReMem: 101043804), the Agence Nationale de la Recherche (RELATIVE: ANR-21-CE37–0008-01; RANGE: ANR-21-CE28–0024-01) and the Alexander von Humboldt-Stiftung. The Département d’Etudes Cognitives is funded by the Agence Nationale pour la Recherche (ANR-17-EURE-0017, ANR-10-IDEX-0001-02). This work has received support under the Major Research Program of PSL Research University "PSL-Neuro" launched by PSL Research University and implemented by ANR (ANR-10-IDEX-0001). C.M.W is supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (C4: 101164709), the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany’s Excellence Strategy (EXC 3066/1 “The Adaptive Mind”, Project No. 533717223), and the Excellence Cluster "Reasonable AI” by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) under Germany’s Excellence Strategy – EXC-3057.

Data availability

This is a theoretical review; no data are associated with this study.

References

Adama S, Bogdan M. Assessing consciousness in patients with locked-in syndrome using their EEG. *Front Neurosci* 2025;**19**:1604173.

Baars BJ. *The cognitive revolution in psychology*. New York: *Guilford* 1986.

Bayne T, Williams I. The Turing test is not a good benchmark for thought in LLMs. *Nat Hum Behav* 2023;**7**:1806–7.

Bayne T, Seth AK, Massimini M *et al.* Tests for consciousness in humans and beyond. *Trends Cogn Sci* 2024;**28**:454–66.

Bengio Y, Hinton G, Yao A *et al.* Managing extreme AI risks amid rapid progress. *Science* 2023;**384**(6698):842–845. https://doi.org/10.1126/science.adn0117

Bickle J. Multiple realizability. 2020. In Zalta, Edward N. (ed.). The Stanford Encyclopedia of Philosophy (Summer 2020 ed.). Metaphysics Research Lab, Stanford. Retrieved from https://plato.sydney.edu.au/entries/multiple-realizability/

Binz M, Akata E, Bethge M *et al.* Centaur: a foundation model of human cognition. *Nature* 2025;**644**:1002–1009. https://doi.org/10.1038/s41586-025-09215-4

Birch J. The edge of sentience: risk and precaution in humans, other animals, and AI. 2024. Oxford University Press. https://library.oapen.org/handle/20.500.12657/93755

Birch J. AI consciousness: a centrist manifesto. *PsyArXiv*. 2025. https://doi.org/10.31234/osf.io/af7c9_v1

Blanchard T, Lombrozo T, Nichols S. Bayesian Occam's razor is a razor of the people. *Cogn Sci* 2018;**42**:1345–59.

Blanshard B. *The Nature of Thought*. London: Allen and Unwin, 1939.

Block N. Troubles with functionalism. *Minn Stud Philos Sci* 1978;**9**:261–325.

Block N. Psychologism and behaviorism. *The Philosophical Review* 1981;**90**:5.

Block N. Can only meat machines be conscious? *Trends Cogn Sci* 2025. https://doi.org/10.1016/j.tics.2025.08.009

Bogen J, Woodward J. Saving the phenomena. *The Philosophical Review* 1988;**97**:303.

Box GEP. Science and statistics. *J Am Stat Assoc* 1976;**71**:791–9.

Brown TB, Mann B, Ryder N *et al.* Language models are few-shot learners. In: *Advances in Neural Information Processing Systems*, vol. 33. 2020: 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Butlin P, Long R, Elmoznino E *et al.* Consciousness in artificial intelligence: insights from the science of consciousness. 2023. *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2308.08708

Chalmers DJ. *The Two-Dimensional Argument against Materialism*. In: Beckermann A, McLaughlin BP, Walter S, editors. The Oxford handbook of philosophy of mind. Oxford: Oxford University Press; 2009. https://doi.org/10.1093/oxfordhb/9780199262618.003.0019

Chalmers DJ. Could a large language model be conscious? *Boston Review* 2023. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/

Chater N. *The Mind Is Flat: The Illusion of Mental Depth and the Improvised Mind*. London: Penguin UK, 2018.

Choi JH, Hickman KE, Monahan AB *et al.* ChatGPT goes to law school. *Journal of Legal Education*. 2021;**71**:387.

Chollet F, Knoop M, Kamradt G *et al.* ARC prize 2024: Technical report. 2024. *arXiv [cs.AI]:2412.04604. [cs.AI]*. arXiv. http://arxiv.org/abs/2412.04604

Cohen MA, Dennett DC. Consciousness cannot be separated from function. *Trends Cogn Sci* 2011;**15**:358–64.

Colombatto C, Fleming SM. Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness* 2024;**2024**:niae013.

Dao T, Gu A. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. In: *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024: article 399. https://dl.acm.org/doi/10.5555/3692070.3692469

Day W. On the difference between radical and methodological behaviorism. *Behaviorism* 1983;**11**:89–102.

DeGrazia D. *Animal Rights: A Very Short Introduction: A Very Short Introduction*. Oxford: Oxford University Press, 2002.

Dehaene S, Changeux J-P, Naccache L. The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In: Dehaene S, Christen Y, editors. *Characterizing Consciousness: From Cognition to the Clinic?. Research and Perspectives in Neurosciences*, pp. 55–84. Berlin: Springer, 2011.

Dehaene S, Lau H, Kouider S. *What Is Consciousness, and Could Machines Have It? Science* 2017;**358**:486–92.

Dennett DC. *Consciousness Explained*. New York: Penguin Books, 1991.

Dettki HM, Lake BM, Wu CM *et al.* Do large language models reason causally like us? Even better? 2025. In: *Proceedings of the 47th Annual Conference of the Cognitive Science Society. Cognitive Science Society*, 2025. https://doi.org/10.48550/arXiv.2502.10215

Doerig A, Schurger A, Hess K *et al.* The unfolding argument: why IIT and other causal structure theories cannot explain consciousness. *Conscious Cogn* 2019;**72**:49–59.

Eldar E, Rutledge RB, Dolan RJ *et al.* Mood as representation of momentum. *Trends Cogn Sci* 2016;**20**:15–24.

Ellia F, Hendren J, Grasso M *et al.* Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness* 2021;**2021**:niab032.

Evers K, Farisco M, Chatila R *et al.* Artificial consciousness. Some logical and conceptual preliminariesx. 2024. arXiv [cs.AI]:2403.20177. http://arxiv.org/abs/2403.20177

Findlay G, Marshall W, Albantakis L *et al.* Dissociating artificial intelligence from artificial consciousness. 2024. arXiv [cs.AI]:2412.04571. http://arxiv.org/abs/2412.04571

Frohlich J, Safron A, Reggente N. Recent pseudoscience accusation echoes historic pushback against general relativity. 2024. *PsyArXiv* preprint. https://doi.org/10.31234/osf.io/awys2

Gandhi K, Fränken J-P, Gerstenberg T *et al.* Understanding social reasoning in language models with language models. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 2023;**595**:13518–13529.

Gershman SJ, Balbi PE, Gallistel CR *et al.* Reconsidering the evidence for learning in single cells. *eLife* 2021;**10**:e61907. https://doi.org/10.7554/eLife.61907

Gläscher JP, O'Doherty JP. Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews Cognitive Science* 2010;**1**:501–10.

Gómez-Marín À, Seth AK. A science of consciousness beyond pseudoscience and pseudo-consciousness. *Nat Neurosci* 2025;**28**:703–706. https://doi.org/10.1038/s41593-025-01913-6

Graziano MSA. Consciousness and the attention schema: why it has to be right. *Cognitive Neuropsychology* 2020;**37**:224–33.

Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. 2023. *arXiv* [cs.LG]:2312.00752. http://arxiv.org/abs/2312.00752

Gui P, Jiang Y, Zang D *et al.* Assessing the depth of language processing in patients with disorders of consciousness. *Nat Neurosci* 2020;**23**:761–70.

Harman G. The inference to the best explanation. *The Philosophical Review* 1965;**74**:88.

Hempel CG, Oppenheim P. Studies in the logic of explanation. *Philos Sci* 1948;**15**:135–75.

Hildt E. Artificial intelligence: does consciousness matter? *Front Psychol* 2019;**10**:1535.

Hohwy J, Seth A. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences* 2020;**1**:3.

IIT-Concerned Klincewicz M, Cheng T, Schmitz M *et al.* What makes a theory of consciousness unscientific? *Nat Neurosci* 2025;**28**: 689–93.

Jannai D, Meron A, Lenz B *et al.* Human or not? A gamified approach to the Turing test. 2023. *arXiv* [cs.AI]:2305.20010. http://arxiv.org/abs/2305.20010

Jones CR, Bergen BK. People cannot distinguish GPT-4 from a human in a Turing test. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, 2025: 1615–1639. https://doi.org/10.1145/3715275.3732108

Kemp C, Tenenbaum JB. Structured statistical models of inductive reasoning. *Psychol Rev* 2009;**116**:20–58.

Kıcıman E, Ness R, Sharma A *et al.* Causal reasoning and large language models: opening a new frontier for causality. *Trans Mach Learn Res*. 20234. http://arxiv.org/abs/2305.00050

Kirk R. Sentience and behaviour. *Mind* 1974;**83**:43–60.

Koriat A. Metacognition and consciousness. In: Zelazo PD, Moscovitch M, Thompson E (eds.), *The Cambridge Handbook of Consciousness*, pp. 289–326. Cambridge: Cambridge University Press, 2007.

Lakatos I. History of science and its rational reconstructions. *PSA* 1970;**1970**:91–136.

Lamme VAF, Roelfsema PR. *The distinct modes of vision offered by feedforward and recurrent processing* 2000;**23**:571–9.

Lashley KS. The behavioristic interpretation of consciousness. I. *Psychol Rev* 1923;30:237–72. https://doi.org/10.1037/h0073839

Leahey TH. The mythical revolutions of American psychology. *The American Psychologist* 1992;**47**:308–18.

Lebreton M, Bavard S, Daunizeau J *et al.* Assessing inter-individual differences with task-related functional neuroimaging. *Nat Hum Behav* 2019;**3**:897–905.

LeDoux J, Birch J, Andrews K *et al.* Consciousness beyond the human case. *Current Biology: CB* 2023;**33**:R832–40.

Lemoine B. Is LaMDA sentient? — an interview. *Medium.* 2022. https://cavouresoterica.it/wp-content/uploads/2022/07/an-Interview-by-Blake-Lemoine-2.pdf

Lenharo M. What should we do if AI becomes conscious? These scientists say it's time for a plan. *Nature* 2024;**636:**533–534. https://doi.org/10.1038/d41586-024-04023-8

Li B, Ge Y, Ge Y *et al.* SEED-bench: Benchmarking multimodal large language models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2024: 13299–308.

Lipton P. *Inference to the Best Explanation*. London: Psychology Press, 2004.

Meehl PE. Appraising and amending theories: the strategy of lakatosian defense and two principles that warrant it. *Psychol Inq* 1990;**1**: 108–41.

Metzinger T. *The Elephant and the Blind.* Cambridge, MA: MIT Press; 2021. https://mitpress.mit.edu/9780262547109/the-elephant-and-the-blind/.

Miller JD, Yampolskiy R, Häggström O. An AGI modifying its utility function in violation of the strong orthogonality thesis. *Philosophies* 2020;**5**:40.

Mitchell M, Krakauer DC. The debate over understanding in AI's large language models. *Proc Natl Acad Sci USA* 2023;**120**:e2215907120.

Moskvichev A, Odouard VV, Mitchell M. The ConceptARC benchmark: evaluating understanding and generalization in the ARC domain. *Trans Mach Learn Res* 2023. https://openreview.net/forum?id=8ykyGbtt2q

Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon Bull Rev* 1997;**4**:79–95. https://doi.org/10.3758/BF03210778

Negro N, Mudrik L. *Extrapolating Other Consciousnesses: The Prospects and Limits of Analogical Abduction. Philosophy of Science*, 2025:1–22. https://doi.org/10.1017/psa.2025.10104.

Niu Q, Liu J, Bi Z *et al.* Large language models and cognitive science: a comprehensive review of similarities, differences, and challenges. 2024. *arXiv* [cs.AI]:2409.02387. http://arxiv.org/abs/2409.02387

Niv Y. The primacy of behavioral research for understanding the brain. *Behav Neurosci* 2021;135:601-9. https://doi.org/10.1037/bne0000471

Nussbaum M. The moral status of animals. *Chron High Educ* 2006;**52**:B6–8.

O'Doherty JP, Hampton A, Kim H. Model-based fMRI and its application to reward learning and decision making. *Ann N Y Acad Sci* 2007;**1104**:35–53.

Palminteri S, Pistilli G. Navigating inflationary and deflationary claims concerning large language models avoiding cognitive biases. *PsyArXiv* preprint. 2025. https://doi.org/10.31234/osf.io/26tyu_v1

Pan L, Xie H, Wilson RC. Large language models think too fast to explore effectively. 2025. In: *Proceedings of the Thirty-ninth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=jW8nBi6y9F

Peng B, Alcaide E, Anthony Q *et al.* RWKV: reinventing RNNs for the transformer era. In: *Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics*, 2023: 14048–77. https://doi.org/10.18653/v1/2023.findings-emnlp.936

Piccinini G. Computationalism in the philosophy of mind. *Philos Compass* 2009;**4**:515–32.

Polger TW. Computational functionalism. In: Robins S, Symons J, Calvo P, editors. *The Routledge companion to philosophy of psychology*. 2nd edn. London: Routledge; 2019: 148–63.

Putnam H. Minds and machines. In: Hook S (ed.), *Dimensions of Mind: A Symposium*, pp. 138–64. New York: New York University Press, 1960.

Putnam H. Brains and behavior. In: *American Association for the Advancement of Science, Section L (History and Philosophy of Science). Washington, DC: American Association for the Advancement of Science;* 1961.

Saanum T, Buschoff LMS, Dayan P *et al.* Next state prediction gives rise to entangled, yet compositional representations of objects. 2024. *arXiv [cs.LG]:2410.04940*. arXiv. http://arxiv.org/abs/2410.04940

Schneider S. How to catch an AI zombie: Testing for consciousness in machines. In: Liao SM, ed. *Ethics of Artificial Intelligence*, pp. 439–58. New York: Oxford University Press, 2020.

Schneider SM, Morris EK. A history of the term radical behaviorism: from Watson to Skinner. *Behav Anal* 1987;**10**:27–39.

Schrimpf M, Blank IA, Tuckute G *et al.* The neural architecture of language: integrative modeling converges on predictive processing. *Proc Natl Acad Sci USA* 2021;**118**:e2105646118.

Searle J. Minds, brains, and programs. *The Behavioral and Brain Sciences* 1980;**3**:417–24.

Sejnowski TJ. Large language models and the reverse Turing test. *Neural Comput* 2023;**35**:309–42.

Seth A. Conscious artificial intelligence and biological naturalism. *Behav Brain Sci* 2025:1–42. https://doi.org/10.1017/S0140525X25000032

Shannon CE. A mathematical theory of communication. *Bell Syst Tech J* 1948;**27**:379–423.

Simon HA. What is an "explanation" of behavior? *Psychol Sci* 1992;**3**:150–61.

Singer P. All animals are equal. 1989. https://philpapers.org/rec/NEGEOChttps://philpapers.org/rec/SINAAA

Singer P. Speciesism and moral status. *Metaphilosophy* 2009;**40**:567–81.

Skinner BF. *Science and Human Behavior*. New York: Free Press, 1965.

Staddon J. *The New Behaviorism: Foundations of Behavioral* Science. 3rd edn. London: Routledge, 2021.

Summerfield C. *These Strange New Minds*. London: Penguin Books, 2025.
Sun Y, Dong L, Huang S *et al.* Retentive network: a successor to transformer for large language models. 2023. *arXiv* [cs.CL]:2307.08621. http://arxiv.org/abs/2307.08621
Sutton RS, Barto AG. *Reinforcement Learning*second edn. An Introduction: MIT Press, 2018.
Tian Y, Ravichander A, Qin L *et al.* Thinking out-of-the-Box: A comparative investigation of human and LLMs in creative problem-solving. In: *ICML 2024 Workshop on LLMs and Cognition*, 2024 https://openreview.net/forum?id=rxkqeYHXy0.
Tomasik B. Do artificial reinforcement-learning agents matter morally? 2014. *arXiv* [cs.AI]:1410.8233. http://arxiv.org/abs/1410.8233
Tononi G, Boly M, Massimini M *et al. Integrated information theory: from consciousness to its physical substrate* 2016;**17**:450–61.
Tononi G, Albantakis L, Barbosa L *et al.* Consciousness or pseudo-consciousness? A clash of two paradigms. *Nat Neurosci* 2025;**28**: 694–702.
Tsuchiya N, Andrillon T, Haun A. A reply to "the unfolding argument": beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. Conscious Cogn 2020;79:102877. https://doi.org/10.1016/j.concog.2020.102877
Turing AM. Computing machinery and intelligence. In: Epstein R, Roberts G, Beber G, editors. *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pp. 23–65. Dordrecht: Springer, 2009.
Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Advances in Neural Information Processing Systems*, 2017;30:5998–6008. http://arxiv.org/abs/1706.03762
Watson P, Pearson D, Wiers RW *et al.* Prioritizing pleasure and pain: attentional capture by reward-related and punishment-related stimuli. *Curr Opin Behav Sci* 2019;**26**:107–13.
Wilson RC, Collins AG. Ten simple rules for the computational modeling of behavioral data. *eLife* 2019;**8**:e49547. https://doi.org/10.7554/eLife.49547
Wilson RC, Niv Y. Is model fitting necessary for model-based fMRI? *PLoS Comput Biol* 2015;**11**:e1004237.
Wong HY. Interrogating artificial agency. *Front Psychol* 2025;**15**: 1449320.
Wu CM, Meder B, Schulz E. Unifying principles of generalization: past, present, and future. *Annu Rev Psychol* 2024;**76**:275–302. https://doi.org/10.1146/annurev-psych-021524-110810
Xu H, Zhao R, Zhu L *et al.* OpenToM: a comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024;1:8593–8623. http://arxiv.org/abs/2402.06044
Yildirim I, Paul LA. From task structures to world models: what do LLMs know? *Trends Cogn Sci* 2024;**28**:404–15.
Yin S, Fu C, Zhao S *et al.* A survey on multimodal large language models. *Natl Sci Rev* 2024;11:nwae403. https://doi.org/10.1093/nsr/nwae403

Neuroscience of Consciousness, 2026, **2026**(1), niag002
https://doi.org/10.1093/nc/niag002
Review Article