Beyond Computational Equivalence: The Behavioral Inference Principle for Machine Consciousness

Stefano Palminteri [1,2,*] & Charley M. Wu [3,4,5,6]

[1] Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL, Research University, Paris, France
[2] Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale, Paris, France
[3] Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany
[4] Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Tübingen, Germany
[5] Centre for Cognitive Science, Institute of Psychology, Technical University of Darmstadt, Darmstadt, Germany
[6] Hessian.AI, Darmstadt, Germany
*corresponding author: stefano.palminteri@gmail.com

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, has 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 criteria to assess machine consciousness.

Acknowledgements

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. 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. C.M.W is supported by the European Research Council (C4: 101164709), the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

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 & Williams, 2023; Jannai et al., 2023; Jones & 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 & Paul, 2024) and neural processes (Saanum et al., 2024; Schrimpf et al., 2021).

While the testing and benchmarking of LLMs continues to generate a wealth of evidence about their strengths and limitations (Dettki et al., 2025; Gandhi et al., 2023; Kıcıman et al., 2023; Moskvichev et al., 2023; Xu et al., 2024) , the wide availability of these tools have 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 have begun to form their own opinions about how we should evaluate the consciousness of these systems (Colombatto & 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 & Pistilli, 2025).

In the realm of philosophy and cognitive science, several perspectives have already been proposed about how to formally evaluate consciousness in LLMs (Bayne et al., 2024; Butlin et al., 2023; Evers et al., 2024; LeDoux et al., 2023). Although other approaches exist, the vast majority of these arguments are grounded in *computational functionalism* (Polger, 2019). 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 realizability" (i.e., multiple physical realizability; Bickle, 1998), where the same cognitive process can be physically realized by different physical 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*) that characterise 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 artificial intelligence 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 (Dehaene et al., 2017; Graziano, 2020; Hohwy & Seth, 2020; Koriat, 2007).

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 Z
If LLMs lack Z, then they are probably not conscious"

where *Z* 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 Chalmer's formulation, our alternative approach suggests:

"*LLMs displays X*
If LLMs displays X, then they are probably conscious"

where *X* 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 Z.

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 (REF).

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 an important 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; Ellia et al., 2021; Tononi et al., 2025; Tsuchiya et al., 2019). However, we will not engage further with these alternatives, since non-computational definitions of consciousness can hardly be integrated into the debate concerning artificial systems, which are quintessentially computational entities.

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: 1) that there is an explicit hypothesis regarding the structure of the target computation, and 2) 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; IIT-Concerned 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 data, 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.

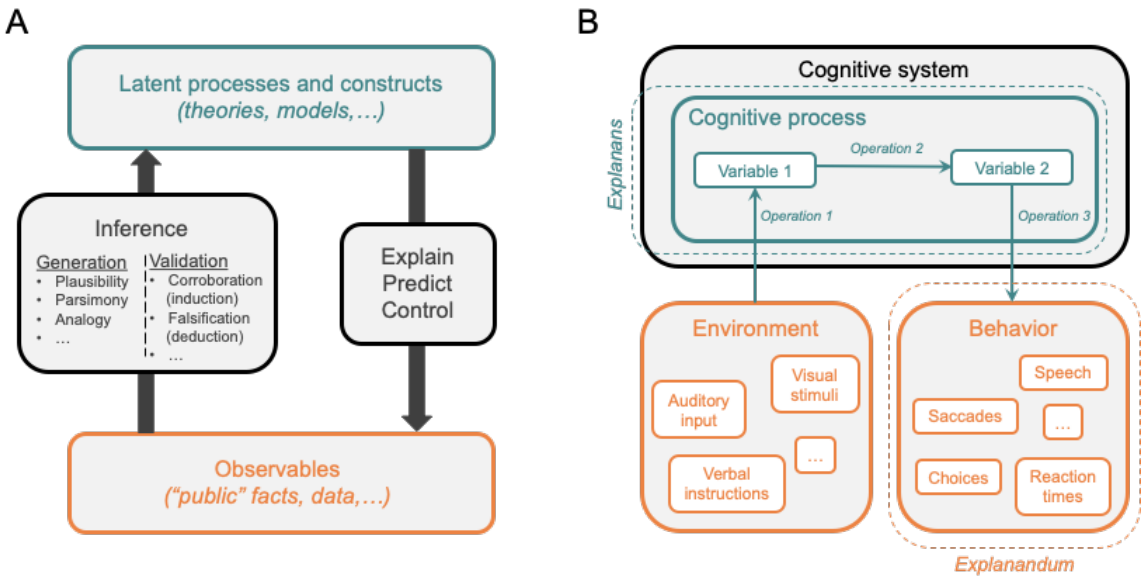

Figure 1: the epistemological status of cognitive (or mental) constructs. **A)** The relation between observables (e.g., public facts and data) and latent processes (e.g., theories and models). Inference typically follows an abductive process (or inference to the best explanation), where hypotheses about 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 color scheme is the same in both A and B, where explanans as latent processes are colored **teal** and explanandum as observables are colored **orange**.

The Epistemology of Cognitive Science

Our search for a an alternative criterion begins from the recognition that like all empirical sciences, including the cognitive science of consciousness, 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 data 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 (S. M. Schneider & 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, verbal reports, or other task performance, 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.

In other terms, the observable behavioral phenomena (and nothing else!) constitutes "the thing that needs an explanation", or in Latin, the *explanandum* (Hempel & 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 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 first hand, thus making it counterintuitive to challenge the primacy of conscious experience (Bayne et al., 2024; Metzinger, 2021). 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*).

---

[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.

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 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.

Beyond the normative, scientific task of characterizing cognitive processes, our everyday folk attribution is also 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) (see **Supplementary discussion** for more details on this issue).

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 have 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 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; Evers et al., 2024; LeDoux et al., 2023). Namely, as inferences to the best explanation for behavioral phenomena generated by an otherwise architecturally opaque computational substrate: the brain (LeDoux et al., 2023; Birch; Bayne et al., 2024).

What about neural data?

While we focus on overt behavior, we do not deny the relevance of neural data 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 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., EEG patterns overlapping with those of conscious individuals) can provide a sufficient sign of consciousness.

A few qualifications follow. First, neural data do not alter the overall epistemological framework: they are useful insofar as they provide the best explanation for observable (neural) evidence. 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 data are often considered secondary to behavioral data in inferring computational mechanisms (Niv, 2021), since behavior can both corroborate and falsify a 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 & Bogdan, 2025) or as a means of providing external validity to computational processes already specified by behavior (e.g., model-based fMRI; Gläscher & O'Doherty, 2010; Lebreton et al., 2019; O'Doherty et al., 2007; Wilson & Niv, 2015). In neither case does neural data 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. 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 attributing 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 & Collins, 2019).

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 & Tenenbaum, 2009; Wu et al., 2024). In contrast, we generalize much less for more distant species, such as mollusks or arthropods, who 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 (AF Lamme & 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 (Dao & Gu, 2024; Gu & Dao, 2023; Peng et al., 2023; Sun et al., 2023).

$P(B)$, 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 & 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 will—and must—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: 1) whether there exists an *explicit* (and ideally consensual) hypothesis concerning the computational mechanisms underlying the cognitive process of interest, and 2) the degree of *transparency* in the system under evaluation (**Fig. 2**).

The first question to ask is whether there is an explicit hypotheses regarding the computational form of the cognitive process under consideration. This maps onto the distinction between theory-heavy (explicit computational hypothesis) and theory-light approaches (no explicit computational hypothesis). Theory-heavy approaches require a well-defined understanding of the target cognitive process, whereas theory-light approaches are more agnostic and don't require the target process to be explicitly defined or agreed upon.

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 systems.

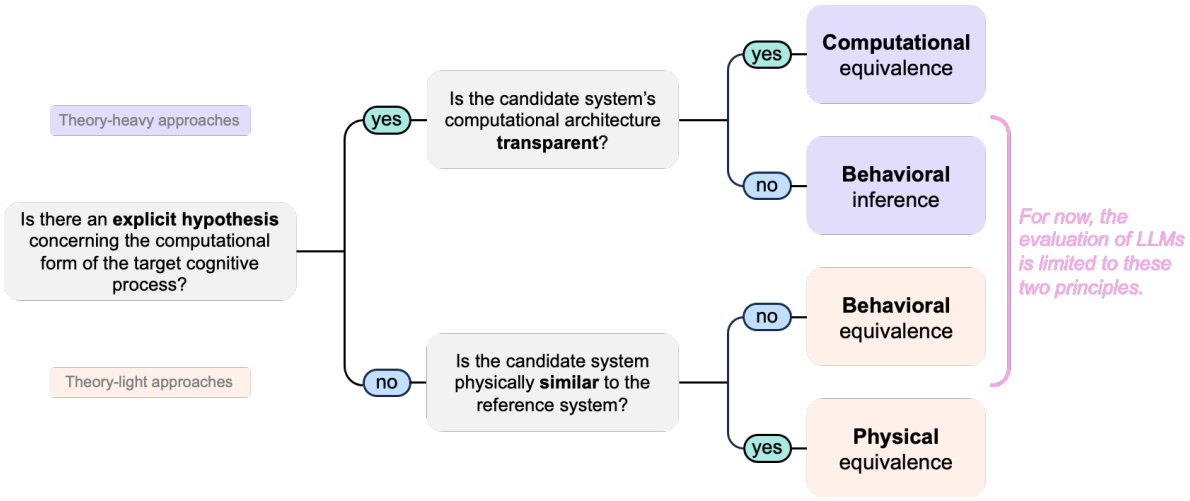

Figure 2: a diagram illustrating how initial conditions map into the applicability of different attribution principles. Theory-heavy approaches require explicit hypotheses about the computational form of the cognitive process under investigation, whereas theory-light approaches do not. Computational 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 C. Elegans and the brute force DeepBlue chess algorithm could be considered examples of transparent systems.

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

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.

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 basic functions typically associated with consciousness, such as language understanding, processing, and reactivity) are (very likely) not conscious (Birch, 2025; Butlin et al., 2023). 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 are considered important for consciousness. 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 (Bengio et al., 2023; Hildt, 2019; 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., a paperclip maximizing agent; 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). 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 world states (i.e., at the very minimum, 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 & 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-direct RL algorithms to improve planning and control, as these systems may increasingly display behavioral patterns we are likely to attribute to a conscious agent.

Figure 3: a metaphor. Blind monks examining an elephant by Hanabusa Itchō (1652 – 1724). This image is in the Public Domain.

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 (S. 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 (Pan et al., 2025; Tian et al., 2024), 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 (Li et al., 2024; Yin et al., 2023).

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 & Krakauer, 2023; Sejnowski, 2023).

To conclude, in the cognitive science community, the question of consciousness in Large Language Models (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. 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.

Box 1: Attributing cognitive processes based on behavioral observations

Behavioral criteria for attributing cognitive processes has an intuitive appeal for empirical scientists (Lashley, 1923; Skinner, 1965) historically been criticized by philosophers of mind (Blanshard, 1939; Block, 1981; Putnam, 1960). Here, we summarize 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. Otherwise, the assumption that "Super-Super-Spartans experience pain" must translate into some observable and intersubjective physical evidence at some observable level. Another instance of "false negative" argument is represented by "philosophical zombies" (Chalmers, 2009; Kirk, 1974), 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.

References

Adama, S., & Bogdan, M. (2025). Assessing consciousness in patients with locked-in syndrome using their EEG. *Frontiers in Neuroscience*, *19*(1604173), 1604173.

AF Lamme, V., & Roelfsema, P. R. (2000). *The distinct modes of vision offered by feedforward and recurrent processing*. *23*(11), 571–579.

Baars, B. J. (1986). *The Cognitive Revolution in Psychology*. Guilford Publications.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences*, *28*(5), 454–466.

Bayne, T., & Williams, I. (2023). The Turing test is not a good benchmark for thought in LLMs. *Nature Human Behaviour*, *7*(11), 1806–1807.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., … Mindermann, S. (2023). Managing extreme AI risks amid rapid progress. In *arXiv [cs.CY]*. arXiv. https://doi.org/10.1126/science.adn0117

Bickle, J. (1998). *Multiple Realizability*. https://plato.sydney.edu.au/entries/multiple-realizability/

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., Griffiths, T. L., Haridi, S., Jagadish, A. K., Ji-An, L., Kipnis, A., Kumar, S., Ludwig, T., Mathony, M., Mattar, M., … Schulz, E. (2024). Centaur: a foundation model of human cognition. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2410.20268

Birch, J. (2025). AI consciousness: A centrist manifesto. In *PsyArXiv*. https://doi.org/10.31234/osf.io/af7c9_v1

Blanshard, B. (1939). *The Nature Of Thought*. Allen and Unwin.

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, *9*, 261–325.

Block, N. (1981). Psychologism and Behaviorism. *The Philosophical Review*, *90*(1), 5.

Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2308.08708

Chalmers, D. J. (2009). *The two-dimensional argument against materialism*. Oxford University Press.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2303.07103

Chater, N. (2018). *The Mind is Flat: The illusion of mental depth and the improvised mind*. Penguin UK.

Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2021). ChatGPT goes to law school. *Journal of Legal Education*. https://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/jled71&section=34&casa_token=jpcbMeYwPEcAAAAA:dqayhyE-vXeq4u61dINqDn-KH-twQ8nvwWDDOWzEmwN26KygtzOAwmbGL_dEQXdQsaCQCsAroA

Chollet, F., Knoop, M., Kamradt, G., & Landers, B. (2024). ARC Prize 2024: Technical Report. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2412.04604

Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, *2024*(1), niae013.

Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2405.21060

Day, W. (1983). On the difference between radical and methodological behaviorism. *Behaviorism*, *11*(1), 89–102.

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

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. In *Research and Perspectives in Neurosciences* (pp. 55–84). Springer Berlin Heidelberg.

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

Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.

Dettki, H. M., Lake, B. M., Wu, C. M., & Rehder, B. (2025). Do large language models reason causally like us? Even better? In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2502.10215

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition*, *72*, 49–59.

Eldar, E., Rutledge, R. B., Dolan, R. J., & Niv, Y. (2016). Mood as representation of momentum. *Trends in Cognitive Sciences*, *20*(1), 15–24.

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., P Lang, J., M Haun, A., Albantakis, L., Boly, M., & Tononi, G. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness*, *2021*(2), niab032.

Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P. F. M. J., & Khamassi, M. (2024). Artificial consciousness. Some logical and conceptual preliminaries. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2403.20177

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

Gandhi, K., Fränken, J.-P., Gerstenberg, T., & Goodman, N. D. (2023). Understanding social reasoning in language models with language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2306.15448

Gershman, S. J., Balbi, P. E., Gallistel, C. R., & Gunawardena, J. (2021). Reconsidering the evidence for learning in single cells. *eLife*, *10*. https://doi.org/10.7554/eLife.61907

Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews. Cognitive Science*, *1*(4), 501–510.

Graziano, M. S. A. (2020). Consciousness and the attention schema: Why it has to be right. *Cognitive Neuropsychology*, *37*(3-4), 224–233.

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

Gui, P., Jiang, Y., Zang, D., Qi, Z., Tan, J., Tanigawa, H., Jiang, J., Wen, Y., Xu, L., Zhao, J., Mao, Y., Poo, M.-M., Ding, N., Dehaene, S., Wu, X., & Wang, L. (2020). Assessing the depth of language processing in patients with disorders of consciousness. *Nature Neuroscience*, *23*(6), 761–770.

Harman, G. (1965). The inference to the best explanation. *The Philosophical Review*, *74*, 88.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Hildt, E. (2019). Artificial intelligence: Does consciousness matter? *Frontiers in Psychology*, *10*, 1535.

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

IIT-Concerned, Klincewicz, M., Cheng, T., Schmitz, M., Sebastián, M. Á., & Snyder, J. S. (2025). What makes a theory of consciousness unscientific? *Nature Neuroscience*, *28*(4), 689–693.

Jannai, D., Meron, A., Lenz, B., Levine, Y., & Shoham, Y. (2023). Human or not? A gamified approach to the Turing test. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2305.20010

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. In *arXiv [cs.HC]*. arXiv. http://arxiv.org/abs/2405.08007

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Kirk, R. (1974). Sentience and Behaviour. *Mind*, *83*(329), 43–60.

Kıcıman, E., Ness, R., Sharma, A., & Tan, C. (2023). Causal reasoning and large language models: Opening a new frontier for causality. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2305.00050

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

Lakatos, I. (1970). History of science and its rational reconstructions. *PSA*, *1970*, 91–136.

Lashley, K. S. (1923). The behavioristic interpretation of consciousness. I. *Psychological Review*. https://psycnet.apa.org/record/1926-07288-001?casa_token=_FQsYCFcI3QAAAAA:WqSeU8s2rNrT5dfpW1W9xo0xf7g7PDWFBa2eJCl63QD8lm_VhpDHX6rLvwvUBxayXaNE1M7Ho32x2Skyre1JN3jL

Leahey, T. H. (1992). The mythical revolutions of American psychology. *The American Psychologist*, *47*(2), 308–318.

Lebreton, M., Bavard, S., Daunizeau, J., & Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. *Nature Human Behaviour*, *3*(9), 897–905.

LeDoux, J., Birch, J., Andrews, K., Clayton, N. S., Daw, N. D., Frith, C., Lau, H., Peters, M. A. K., Schneider, S., Seth, A., Suddendorf, T., & Vandekerckhove, M. M. P. (2023). Consciousness beyond the human case. *Current Biology: CB*, *33*(16), R832–R840.

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

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

Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., & Shan, Y. (2024). SEED-Bench: Benchmarking Multimodal Large Language Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13299–13308.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Metzinger, T. (2021, December 1). *The Elephant and the Blind*. MIT Press; The MIT Press, Massachusetts Institute of Technology. https://mitpress.mit.edu/9780262547109/the-elephant-and-the-blind/

Miller, J. D., Yampolskiy, R., & Häggström, O. (2020). An AGI modifying its utility function in violation of the strong orthogonality thesis. *Philosophies*, *5*(4), 40.

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(13), e2215907120.

Moskvichev, A., Odouard, V. V., & Mitchell, M. (2023). The ConceptARC benchmark: Evaluating understanding and generalization in the ARC domain. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2305.07141

Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., Li, M., Yan, L. K. Q., Zhang, Y., Yin, C. H., Fei, C., Wang, T., Wang, Y., & Chen, S. (2024). Large Language Models and cognitive science: A comprehensive review of similarities, differences, and challenges. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2409.02387

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*. https://doi.org/10.31234/osf.io/y8mxe

Nussbaum, M. (2006). The moral status of animals. *The Chronicle of Higher Education*, *52*(22), B6–B8.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*(1), 35–53.

Palminteri, S., & Pistilli, G. (2025). Navigating inflationary and deflationary claims concerning Large Language Models avoiding cognitive biases. In *PsyArXiv*. https://doi.org/10.31234/osf.io/26tyu_v1

Pan, L., Xie, H., & Wilson, R. C. (2025). Large Language Models think too fast to explore effectively. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2501.18009

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., Gv, K. K., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., Lau, H., … Zhu, R.-J. (2023). RWKV: Reinventing RNNs for the Transformer era. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2305.13048

Piccinini, G. (2009). Computationalism in the philosophy of mind. *Philosophy Compass*, *4*(3), 515–532.

Polger, T. W. (2019). Computational Functionalism. In *The Routledge Companion to Philosophy of Psychology* (2nd Edition, pp. 148–163). Routledge.

Putnam, H. (1960). Minds and Machines. In S. Hook (Ed.), *Dimensions Of Mind: A Symposium.* (pp. 138–164). NEW YORK University Press.

Putnam, H. (1961). Brains and Behavior. *American Association for the Advancement of Science, Section L (History and Philosophy of Science)*.

Saanum, T., Buschoff, L. M. S., Dayan, P., & Schulz, E. (2024). Next state prediction gives rise to entangled, yet compositional representations of objects. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2410.04940

Schneider, S. (2020). How to catch an AI zombie: Testing for consciousness in machines. In *Ethics of Artificial Intelligence* (pp. 439–458). Oxford University PressNew York.

Schneider, S. M., & Morris, E. K. (1987). A history of the term radical behaviorism: From Watson to Skinner. *The Behavior Analyst*, *10*(1), 27–39.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on

predictive processing. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(45), e2105646118.

Searle, J. (1980). Minds, brains, and programs. *The Behavioral and Brain Sciences*, *3*, 417–424.

Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural Computation*, *35*(3), 309–342.

Seth, A. (2025). Conscious artificial intelligence and biological naturalism. In *PsyArXiv*. https://doi.org/10.31234/osf.io/tz6an_v2

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science*, *3*(3), 150–161.

Singer, P. (1989). *All Animals Are Equal*. https://philpapers.org/rec/SINAAA

Singer, P. (2009). Speciesism and moral status. *Metaphilosophy*, *40*(3-4), 567–581.

Skinner, B. F. (1965). *Science And Human Behavior*. Free Press.

Staddon, J. (2021). *The new behaviorism: Foundations of behavioral science* (3rd ed.). Routledge.

Summerfield, C. (2025). *These strange new minds*. Penguin Books.

Sun, Y., Dong, L., Huang, S., Ma, S., Xia, Y., Xue, J., Wang, J., & Wei, F. (2023). Retentive Network: A successor to Transformer for large language models. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2307.08621

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.

Tian, Y., Ravichander, A., Qin, L., Le Bras, R., Marjieh, R., Peng, N., Choi, Y., Griffiths, T. L., & Brahman, F. (2024, July 26). Thinking Out-of-the-Box: A Comparative Investigation of Human and LLMs in Creative Problem-Solving. *ICML 2024 Workshop on LLMs and Cognition*. https://openreview.net/forum?id=rxkqeYHXy0

Tomasik, B. (2014). Do artificial reinforcement-learning agents matter morally? In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/1410.8233

Tononi, G., Albantakis, L., Barbosa, L., Boly, M., Cirelli, C., Comolatti, R., Ellia, F., Findlay, G., Casali, A. G., Grasso, M., Haun, A. M., Hendren, J., Hoel, E., Koch, C., Maier, A., Marshall, W., Massimini, M., Mayner, W. G., Oizumi, M., … Zaeemzadeh, A. (2025). Consciousness or pseudo-consciousness? A clash of two paradigms. *Nature Neuroscience*, *28*(4), 694–702.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). *Integrated information theory: from consciousness to its physical substrate. 17*(7), 450–461.

Tsuchiya, N., Andrillon, T., & Haun, A. (2019). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. In *PsyArXiv*. https://doi.org/10.31234/osf.io/a2ms9

Turing, A. M. (2009). Computing Machinery and Intelligence. In *Parsing the Turing Test* (pp. 23–65). Springer Netherlands.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762

Watson, P., Pearson, D., Wiers, R. W., & Le Pelley, M. E. (2019). Prioritizing pleasure and pain: attentional capture by reward-related and punishment-related stimuli. *Current Opinion in Behavioral Sciences*, *26*, 107–113.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *eLife*, *8*. https://doi.org/10.7554/eLife.49547

Wilson, R. C., & Niv, Y. (2015). Is model fitting necessary for model-based fMRI? *PLoS Computational Biology*, *11*(6), e1004237.

Wong, H. Y. (2025). Interrogating artificial agency. *Frontiers in Psychology*, *15*, 1449320.

Wu, C. M., Meder, B., & Schulz, E. (2024). Unifying principles of generalization: Past, present, and future. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-021524-110810

Xu, H., Zhao, R., Zhu, L., Du, J., & He, Y. (2024). OpenToM: A comprehensive benchmark for

evaluating Theory-of-Mind reasoning capabilities of large language models. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2402.06044

Yildirim, I., & Paul, L. A. (2024). From task structures to world models: what do LLMs know? *Trends in Cognitive Sciences*, *28*(5), 404–415.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2023). A survey on multimodal Large Language Models. In *arXiv [cs.CV]*. arXiv. https://doi.org/10.1093/nsr/nwae403