Beyond Computational Functionalism: The Behavioral Inference Principle for Machine Consciousness

Stefano Palminteri (1,2) & Charley M. Wu (3,4)

(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

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 unbiased and operationalizable criterion to assess machine consciousness.

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"[1] with human-in-the-loop reinforcement learning[2], LLMs have demonstrated remarkable performance across a staggeringly wide range of tasks. For instance, fooling humans in a Turing test[3–5] or passing law school exams[6]. And despite a number of key challenges, such as a surprising difficulty in solving rather simple abstract reasoning problems (e.g., the ARC Prize)[7,8] or reliably reason about the mental states of people (i.e., Theory of Mind)[9], researchers are increasingly using LLMs as models of human cognitive[10–12] and neural processes[13,14].

While the testing and benchmarking of LLMs[7,9,15,16] continues to generate a wealth of evidence about their strengths and limitations, the wide availability of these tools have reached a larger audience than perhaps any other AI tool before it. 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[17]. 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[18].

In the realm of philosophy and cognitive science, several perspectives have already been proposed about how to formally evaluate consciousness in LLMs[19–22]. The vast majority of these arguments are grounded in *computational functionalism*[23]. 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. This allows for "multiple realizability"[24], where the same computational process can be physically realized by different physical systems. Thus, according to computational functionalism, an artificial system achieves consciousness when it attains *computational equivalence*, meaning it replicates the computational processes that characterize consciousness as identified by these theories.

For instance, a recent paper led by Patrick Butlin, Robert Long, and co-signed by seventeen other experts in the science of consciousness and artificial intelligence follows this tradition[19]. 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 metacognitive/predictive processing[25–28].

An equivalent position is taken by Susan Schneider in another collective piece[22], 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 same logical form of requiring an explicit computational process as a *necessary* condition for consciousness. As framed by the philosopher David Chalmers[29], a typical argument has the following format:

"*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 or a global workspace.

Here, 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 utility of computational equivalence as a demarcation 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, *behavioral* pattern rather than an unobservable, computational process.

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 to be more consistent with the epistemological and methodological approach used by cognitive scientists to study natural consciousness. As in all empirical sciences, this approach relies on inductive inference from experimental observations, where observable behavior is the key form of evidence, rather than deductive logic.

Why Computational Equivalence is Insufficient

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. However, as a disclaimer, we are not dismissing 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—a necessary step for considering computers as relevant metaphors for the human brain, and vice versa (i.e., multiple physical realizability[30]). Thus, our contention lies not with

computational functionalism, but with their application as criteria for attributing consciousness from a scientific and empirical point of view.

Our primary argument is grounded in the understanding that like all sciences, the study of consciousness, advances through a process of inductive inference, relying on cycles of corroboration (empirical evidence supporting a theory) and falsification (evidence against a theory that requires revisiting it)[31,32]. Specifically, cognitive science as an interdisciplinary field for studying the mind operates as a form of methodological behaviorism[33]. This involves using behavioral data to infer plausible latent computational processes that explain the observed behavior (subject to potential falsification by future evidence) and are instrumentally useful to predict and control future behavior by the same system[34].

In contrast, using computational equivalence to attribute consciousness takes the reverse approach: it begins with the presumption of knowing the "correct" computational model (but see REFS[35,36] for lively debates on this question) and seeks to verify whether that computation exists in the system under study using deductive logic. In the subsequent section we detail further how the cognitive science inferential process works and how awareness of its epistemological mechanisms should affect the machine consciousness debate.

The Objects of Cognitive Science

To understand our argument, it is important to reflect upon the epistemological and ontological status of cognitive processes, such as “consciousness”, and their relation to behavioral phenomena. Our premise rests on the widely held assumption that the objects of scientific enquiry are public, interpersonal facts — or simply, observable *behavioral phenomena*. In contrast, theoretical concepts such as consciousness are *latent variables*, which are not directly observable in the data, but are hypothesised to explain the phenomena. Thus, our argument can be summarized into three main parts: (i) phenomena are explained by theory, but not vice versa, (ii) cognitive science uses inductive inference rather than deductive logic, and (iii) there can be no singular “Litmus test” for consciousness.

Phenomena are explained by theory, but not vice versa

In cognitive science, the relevant phenomena are observable *behavioral phenomena*, encompassing body, eye movements, and language production, but perhaps also measurable neural data in a broader sense. These observable phenomena are quantified and distilled into data—choices, reaction times, eye saccades, verbal reports, etc.—which cognitive scientists analyze. 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. These observables—both behavioral and experimental—are then employed to construct and validate hypotheses regarding latent cognitive constructs and the processes underlying them.

In other terms, the observable behavioral phenomena (and nothing else!) is "the thing that needs an explanation", or in Latin, the *explanandum*[37]. In contrast, latent cognitive constructs and theories are "the thing that explains", or *explanans* (**Fig. 1**). The phenomena are explained by the theory, but not vice versa.

Yet, *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[21]. 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 a 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*).

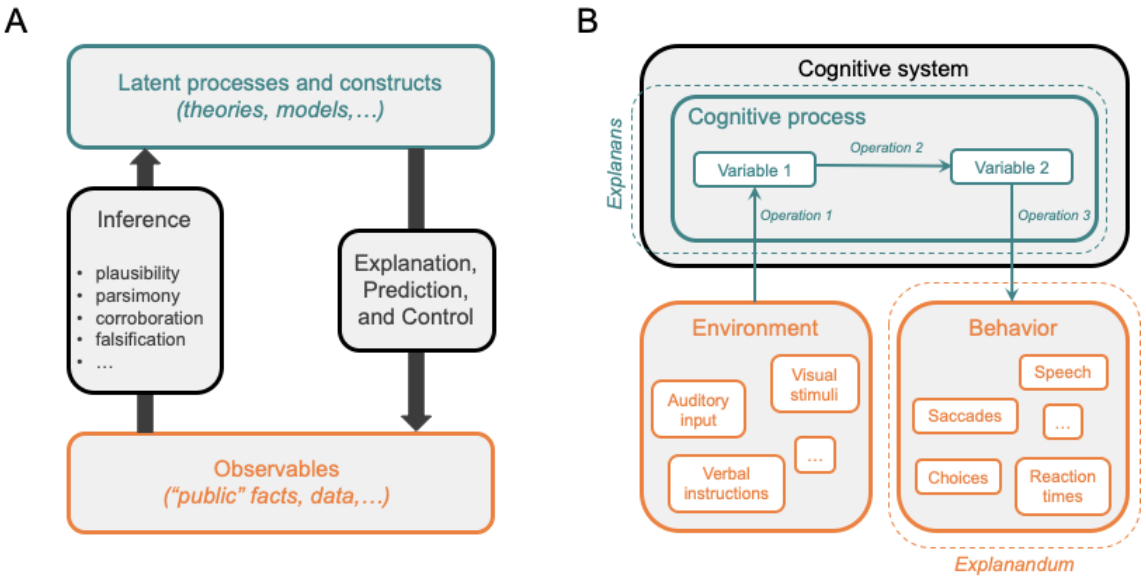

Figure 1: the epistemological status of cognitive (or mental) constructs. **A** The relation between observables (e.g., public facts and data) and latent processes (e.g., theories and models). **B** The relation between a cognitive process (*explanans*: the thing that explains), the behavioral phenomena (*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**.

Translating to the theme of consciousness in artificial systems, the emphasis should not be on whether a machine implements the 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 compells) us to infer a latent process that we agree to call consciousness, given the available data.

This perspective holds true for humans as well. The reason we know (in the scientific sense) that other humans possess consciousness is not because we directly perceive computational equivalence. In fact, we have very little insight into our own internal computations [38], 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. This is also why philosophical zombies [39,40] — hypothetical beings who are indistinguishable from us with the exception of lacking conscious experience — proved so divisive among consciousness researchers (see **Box 1**). By any standard of empirical science, we must consider philosophical zombies to be conscious, even though the thought experiment demands that we accept the counterintuitive (and seemingly unprovable) fact that they lack conscious experience.

Thus, computational equivalence as the sole necessary and sufficient condition for attributing consciousness to artificial systems [19–22] imposes a double standard between humans and machines. Since we infer consciousness in humans (or other animals; [22]) from behavior and through rigorous scientific experiments, the same standard ought to apply to machines.

Cognitive science uses inductive inference rather than deductive logic

Although computational equivalence relies on *deductive* logic, cognitive science uses *inductive* reasoning to form beliefs about the existence of some latent property Z as a function of the behavioral evidence X. This can be described as a form of Bayesian inference:

$$P(Z|X) \propto P(X|Z)P(Z) \quad (1)$$

Here, the posterior $P(Z|X)$ represents our inferred beliefs about whether the system has consciousness (or any other latent property Z), conditioned on the observed behavioral evidence X. This posterior is proportional to the likelihood of the data $P(X|Z)$ (i.e., if Z is present then likely is the system to produce behavior X?) multiplied by the prior $P(Z)$ (i.e., how likely is Z independent of data?).

The Bayesian formalism of inductive inference may help explain some of our counter-intuitive gut reactions to common philosophical thought experiments. For instance, the influence of the prior

$P(Z)$ can explain why our gut may react differently in assigning human-like properties to AI or other non-human animals based on the same face-value behavioral evidence. We may simply have stronger priors that certain properties, such as consciousness, are unlikely to appear in other systems the less similar they are to humans.

However, priors change as we acquire new information, and what drives changes to these prior beliefs is the same stream of behavioral evidence X. Thus, the Bayesian formalism of inductive inference allows us to treat biological organisms and machines as both the same (i.e., consciousness should be attributed based on behavioral evidence) but also different (i.e., requiring different amounts of behavioral evidence to overrule prior beliefs).

There can be no final "Litmus test" for consciousness

Here, we have argued for how the science of consciousness can be informed by cognitive science, with (i) a reminder about the ontological primacy of phenomena (*explanandum*) over models and theories (*explanans*), and (ii) using the inductive logic of inferring latent properties from observable behavior. This primacy does not imply the behavior is "more important" or "more interesting" than the latent cognitive constructs used to explain it. To the contrary, behavioral observations do not possess any epistemic or instrumental value, since only theories provide explanation, prediction and control over past and future observations. Yet, observable behavior must be the driving force behind how our theories are updated and developed.

This brings us to an important fact about cognitive science, which like all empirical sciences, is not static. Indeed, the recent reproducibility crisis in behavioral science has shone a much-needed spotlight on important shortcomings of established experimental and statistical methods[41], where the predictions of many foundational theories have failed to replicate. While there still exist many open challenges, the strength of empirical science is precisely in how it evolves in response to new findings, conceptual challenges, and methodological crises. This constant evolution of methods yields a dynamic evolution of theories, which are amended and expanded through empirical falsification and corroboration.

Thus, applied to the question of consciousness, this constant evolvability of empirical science is necessary to allow us to continually refine our explanations of behavioral phenomena. As we encounter or even construct new subjects of study (e.g., LLMs), we will continually need to develop new tests, experiments, and methods in order to refine our understanding about the best explanations. Therefore, we are not proposing some new (or final!) behavioral test for consciousness akin to the "Turing test"[42]. Rather, we believe there can be no singular and eternally valid behavioral test for any cognitive phenomena, most of all not consciousness[43].

Instead, the science of machine consciousness must embrace the behavioral methodology of cognitive science in doing away with "necessary and sufficient conditions", and being willing to continually evolve through corroboration, falsification, and the development of new empirical standards and operational definitions. 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. In the next section, we focus on what is at stake in the machine consciousness debate, and which instrumental goals shape the debate.

Why do we need 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 are (probably) not conscious in the richer sense of the term. Current LLMs may exhibit some basic functions typically associated with "poorer" forms of consciousness, particularly those linked to language understanding, processing, and reactivity[21]. But, we also concur with Butlin, Long, and colleagues[19] 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"[44]. 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[45–47]. 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[45]. 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[48]). 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[49]. 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[47] 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[50]) concerns the potential for these entities to acquire *moral status,* which is the degree to which an organism deserves ethical consideration[51,52]. However, it is unclear whether consciousness *per se* is the appropriate criterion. Taking the example of non-human animals[53], 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[54]). 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[55–57], leading to arguments that RL agents may already possess a non-zero moral status, even in the absence of consciousness[58]. 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.

The elephant in the room and the elephant in the brain

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. 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.

Figure 2. Blind monks examining an elephant by Hanabusa Itchō (1652 – 1724). This image is in the Public

Domain.

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 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.

Indeed, the monks themselves are applying an empirical methodology. 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, falsification, replication, and above all else, collaboration.
Boxes

Box 1: Attributing mental processes based on behavioral observations

A reliance on behavioral criteria to attribute mental or cognitive processes has an intuitive appeal for those coming from an empirical science background. However, this approach[59,60] has historically been criticized by philosophers of mind[24,61,62]. Thus, the aim of this section is to summarize some of the historical criticisms against methodological behaviorism in cognitive science and clarify how our *behavioral inference principle* avoids these arguments, based on the inductive logic of cognitive science.

Consider the following formulation of the behavioral inference principle:
*If an agent displays behavior **B**, then it probably possesses cognitive process **C**.*

In the case of the original Turing test[42], B would correspond to "full conversational proficiency" and C to "intelligence", which are admittedly, very vague constructs. However, this framework can be applied broadly to any mental construct or cognitive process, such as consciousness[4]. Historically, criticisms of behavior-centric theories of cognitive processes generally focus on demonstrating that it is possible to exhibit behavior B without possessing cognitive process C (i.e., the attribution of C is a "false positive") or that it is possible to possess cognitive process C without displaying behavior B (i.e., the non-attribution of B is a "false negative").

The first “false positive” argument can be exemplified by Ned Block’s influential Blockhead thought experiment[62]. Block suggests that it is conceivable to design a machine that passes the Turing test, 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. We can acknowledge that this thought experiment illustrates the logical possibility of displaying behavior B without possessing cognitive process C. 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 not physically feasible — and even if it were, the retrieval and response times would be infinitely long[63]. 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. Conversely, a scientist receiving prompt and sensible responses to virtually any

question, would be justified to infer that the machine is truly intelligent.

The second "false negative" result is exemplified by the Super-Super-Spartans thought experiment proposed by Putnam[64]. It is possible to conceive a parallel universe where 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 .

Thus, while thought experiments have been devised to challenge the use of behavioral criteria to infer cognitive processes, they bear little consequence when applying the scientific method of cognitive science, which operates on inductive logic. In the first case, the validity of the Turing test is negated by appealing to a physically impossible device (incidentally with infinitely long reaction times), which in practice, would fail to demonstrate convincing behavioral evidence for the cognitive process in question (i.e., intelligence). In the second case, we might reach an incorrect yet scientifically valid conclusion ("Super-Super Spartans do not feel pain"). Thus, our *behavioral inference principle* avoids both false positive and false negative scenarios by adopting the flexibility of inductive reasoning used in empirical sciences, 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.

Acknowledgements

The authors wish to thank Valeria Giardino, Lindsay Drayton, Steve Fleming, Patrick Butlin, and Hongyu Wong for their helpful comments and discussions.

References

1. Vaswani, A. *et al.* Attention is all you need. *arXiv [cs.CL]* (2017).
2. Brown, T. B. *et al.* Language Models are Few-Shot Learners. *arXiv [cs.CL]* (2020).
3. Jannai, D., Meron, A., Lenz, B., Levine, Y. & Shoham, Y. Human or not? A gamified approach to the Turing test. *arXiv [cs.AI]* (2023).
4. Bayne, T. & Williams, I. The Turing test is not a good benchmark for thought in LLMs. *Nat. Hum. Behav.* **7**, 1806–1807 (2023).
5. Jones, C. R. & Bergen, B. K. People cannot distinguish GPT-4 from a human in a Turing test. *arXiv [cs.HC]* (2024).
6. Choi, J. H., Hickman, K. E., Monahan, A. B. & Schwarcz, D. ChatGPT goes to law school. *J. Legal Educ.* (2021).
7. Moskvichev, A., Odouard, V. V. & Mitchell, M. The ConceptARC benchmark: Evaluating understanding and generalization in the ARC domain. *arXiv [cs.LG]* (2023).
8. Chollet, F., Knoop, M., Kamradt, G. & Landers, B. ARC Prize 2024: Technical Report. *arXiv [cs.AI]* (2024).
9. Xu, H., Zhao, R., Zhu, L., Du, J. & He, Y. OpenToM: A comprehensive benchmark for evaluating Theory-of-Mind reasoning capabilities of large language models. *arXiv [cs.AI]* (2024).
10. Yildirim, I. & Paul, L. A. From task structures to world models: what do LLMs know? *Trends Cogn. Sci.* **28**, 404–415 (2024).
11. Niu, Q. *et al.* Large Language Models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv [cs.AI]* (2024).
12. Binz, M. *et al.* Centaur: a foundation model of human cognition. *arXiv [cs.LG]* (2024).
13. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2105646118 (2021).
14. Saanum, T., Buschoff, L. M. S., Dayan, P. & Schulz, E. Next state prediction gives rise to entangled, yet compositional representations of objects. *arXiv [cs.LG]* (2024).
15. Gandhi, K., Fränken, J.-P., Gerstenberg, T. & Goodman, N. D. Understanding social reasoning in language models with language models. *arXiv [cs.CL]* (2023).
16. Kıcıman, E., Ness, R., Sharma, A. & Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv [cs.AI]* (2023).
17. Colombatto, C. & Fleming, S. M. Folk psychological attributions of consciousness to large language models. *Neurosci. Conscious.* **2024**, niae013 (2024).
18. Lenharo, M. What should we do if AI becomes conscious? These scientists say it's time for a plan. *Nature* (2024) doi:10.1038/d41586-024-04023-8.
19. Butlin, P. *et al.* Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv [cs.AI]* (2023).
20. Evers, K. *et al.* Artificial consciousness. Some logical and conceptual preliminaries. *arXiv [cs.AI]* (2024).
21. Bayne, T. *et al.* Tests for consciousness in humans and beyond. *Trends Cogn. Sci.* **28**, 454–466 (2024).
22. LeDoux, J. *et al.* Consciousness beyond the human case. *Curr. Biol.* **33**, R832–R840 (2023).
23. Polger, T. W. Computational Functionalism. in *The Routledge Companion to Philosophy of Psychology* 148–163 (Routledge, Second edition. | Abingdon, Oxon ; New York, NY : Routledge, Taylor & Francis Group, 2020., 2019).
24. Putnam, H. Minds and Machines. in *Dimensions Of Mind: A Symposium.* (ed. Hook, S.) 138–164 (NEW YORK University Press, 1960).

25. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
26. Koriat, A. Metacognition and consciousness. in *The Cambridge Handbook of Consciousness* (eds. Zelazo, P. D., Moscovitch, M. & Thompson, E.) 289–326 (Cambridge University Press, Cambridge, 2007).
27. Hohwy, J. & Seth, A. Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *PhiMiSci* **1**, 3 (2020).
28. Graziano, M. S. A. Consciousness and the attention schema: Why it has to be right. *Cogn. Neuropsychol.* **37**, 224–233 (2020).
29. Chalmers, D. J. Could a Large Language Model be Conscious? *arXiv [cs.AI]* (2023).
30. Bickle, J. Multiple Realizability. (1998).
31. Meehl, P. E. Appraising and amending theories: The strategy of lakatosian defense and two principles that warrant it. *Psychol. Inq.* **1**, 108–141 (1990).
32. Lakatos, I. History of science and its rational reconstructions. *PSA* **1970**, 91–136 (1970).
33. Day, W. On the difference between radical and methodological behaviorism. *Behaviorism* **11**, 89–102 (1983).
34. Staddon, J. *The New Behaviorism: Foundations of Behavioral Science*. (Routledge, London, England, 2021).
35. IIT-Concerned *et al.* The Integrated Information Theory of consciousness as pseudoscience. *PsyArXiv* (2023) doi:10.31234/osf.io/zsr78.
36. Frohlich, J., Safron, A. & Reggente, N. Recent pseudoscience accusation echoes historic pushback against general relativity. *PsyArXiv* (2024) doi:10.31234/osf.io/awys2.
37. Hempel, C. G. & Oppenheim, P. Studies in the logic of explanation. *Philos. Sci.* **15**, 135–175 (1948).
38. Chater, N. *The Mind Is Flat: The Illusion of Mental Depth and the Improvised Mind*. (Penguin UK, 2018).
39. Kirk, R. Sentience and Behaviour. *Mind* **83**, 43–60 (1974).
40. Chalmers, D. J. *The Two-dimensional Argument against Materialism*. (Oxford University Press, 2009).
41. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).
42. Turing, A. M. Computing Machinery and Intelligence. in *Parsing the Turing Test* 23–65 (Springer Netherlands, Dordrecht, 2009).
43. Schneider, S. How to catch an AI zombie: Testing for consciousness in machines. in *Ethics of Artificial Intelligence* 439–458 (Oxford University PressNew York, 2020).
44. Box, G. E. P. Science and statistics. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).
45. Bengio, Y. *et al.* Managing extreme AI risks amid rapid progress. *arXiv [cs.CY]* (2023) doi:10.1126/science.adn0117.
46. Hildt, E. Artificial intelligence: Does consciousness matter? *Front. Psychol.* **10**, 1535 (2019).
47. Wong, H. Y. Interrogating artificial agency. *Front. Psychol.* **15**, 1449320 (2025).
48. Miller, J. D., Yampolskiy, R. & Häggström, O. An AGI modifying its utility function in violation of the strong orthogonality thesis. *Philosophies* **5**, 40 (2020).
49. Gershman, S. J., Balbi, P. E., Gallistel, C. R. & Gunawardena, J. Reconsidering the evidence for learning in single cells. *Elife* **10**, (2021).
50. Lemoine, B. Is LaMDA Sentient? — an Interview. *Medium* https://cavouresoterica.it/wp-content/uploads/2022/07/an-Interview-by-Blake-Lemoine-2.pdf (2022).
51. Nussbaum, M. The moral status of animals. *Chron. High. Educ.* **52**, B6–8 (2006).
52. Singer, P. Speciesism and moral status. *Metaphilosophy* **40**, 567–581 (2009).

53. DeGrazia, D. *Animal Rights: A Very Short Introduction: A Very Short Introduction*. (Oxford University Press, London, England, 2002).
54. Singer, P. All Animals Are Equal. (1989).
55. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, Second Edition: An Introduction*. (MIT Press, 2018).
56. Watson, P., Pearson, D., Wiers, R. W. & Le Pelley, M. E. Prioritizing pleasure and pain: attentional capture by reward-related and punishment-related stimuli. *Curr. Opin. Behav. Sci.* **26**, 107–113 (2019).
57. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as representation of momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016).
58. Tomasik, B. Do artificial reinforcement-learning agents matter morally? *arXiv [cs.AI]* (2014).
59. Lashley, K. S. The behavioristic interpretation of consciousness. I. *Psychol. Rev.* (1923).
60. Skinner, B. F. *Science And Human Behavior*. (Free Press, New York, NY, 1965).
61. Blanshard, B. *The Nature Of Thought*. (Allen and Unwin, London, 1939).
62. Block, N. Psychologism and Behaviorism. *Philos. Rev.* **90**, 5 (1981).
63. Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379–423 (1948).
64. Putnam, H. Brains and Behavior. in *American Association for the Advancement of Science, Section L (History and Philosophy of Science)* (1961).