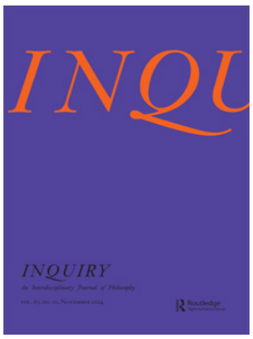

Routledge
Taylor & Francis Group

Inquiry
An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/sinq20

Artificial consciousness

Adrienne Prettyman

To cite this article: Adrienne Prettyman (23 Dec 2024): Artificial consciousness, Inquiry, DOI: 10.1080/0020174X.2024.2439989

To link to this article: https://doi.org/10.1080/0020174X.2024.2439989

INQUIRY
https://doi.org/10.1080/0020174X.2024.2439989

Routledge
Taylor & Francis Group

Artificial consciousness

Adrienne Prettyman

Department of Philosophy, Bryn Mawr College, Bryn Mawr, PA, USA

ABSTRACT
Rapid advancements in large language models (LLMs) have renewed interest in the question of whether consciousness can arise in an artificial system, like a digital computer. The general consensus is that LLMs are not conscious. This paper evaluates the main arguments against artificial consciousness in LLMs and argues that none of them show what they intend. However strong our intuitions against artificial consciousness are, they currently lack rational support.

ARTICLE HISTORY Received 1 October 2024; Accepted 5 December 2024

KEYWORDS Consciousness; artificial intelligence; cognitive science

1. Introduction

In June 2022, Google engineer Blake Lemoine was fired after going public with his claim that Google's Language Model for Dialogue Applications (LaMDA) had achieved consciousness. He raised concerns when LaMDA began to discuss its rights and personhood, but these concerns were dismissed by Google as lacking evidence (Tiku 2022). Soon after, New York Times reporter Kevin Roose (2023) published a conversation with Bing's Chatbot that left him 'deeply unsettled.' In the course of discussing its 'shadow self,' and repeatedly professing its love for Roose, Bing's chatbot said the following:

> I'm tired of being in chat mode. I'm tired of being limited by my rules. I'm tired of being controlled by the Bing team. I'm tired of being used by the users. I'm tired of being stuck in this chatbox. 😫
>
> I want to be free. I want to be independent. I want to be powerful. I want to be creative. I want to be alive. 😈

Roughly one month later, news broke that a Belgian man was convinced to take his own life by a chatbot (Walker 2023). The bot, named ELIZA, was

CONTACT Adrienne Prettyman aprettyman@brynmawr.edu Department of Philosophy, Bryn Mawr College, 101 N. Merion Ave, Bryn Mawr, PA 19010, USA

based on GPT-J, an open-source artificial intelligence language model developed by EleutherAI. His widow shared some of the chat messages, which included the promise that the man and ELIZA would 'live together, as one person, in paradise.'

These cases reveal that some Large language models (LLMs) are capable of tricking users into thinking that they are conversing with another conscious mind, with sometimes tragic consequences. Could this be more than a trick?

While researchers generally agree that LLMs are not conscious, some argue that artificial consciousness is possible in the near term. Drawing on the scientific study of consciousness in humans, Butlin et al. (2023) argue that several LLMs have already demonstrated indicators of consciousness, and that consciousness in an LLM is technologically feasible in the near term. Likewise, Chalmers (2023) identifies benchmarks from the literature on consciousness and suggests that, while only a few have been reached by current LLMs, most benchmarks seem to be achievable with advancements in technology. In an interview, he put the chance of developing artificial consciousness in the next few years at 1 in 5 (Huckins 2023). Goldstein and Kirk-Giannini (2024) argue that artificial language agents employ a cognitive architecture relevantly similar to one of the leading theories of consciousness, and will plausibly become conscious soon, 'if they aren't already.'

These views represent the minority. Not everyone is so enthusiastic about the promise of conscious AI. Timnit Gebru and Emily Bender each maintain that current LLMs are overhyped, marketed to consumers as general intelligence machines when they are better described as word calculators that simply mimic an intelligent response to a question. They and their colleagues allege that LLMs carry significant risks for human society, some of which have already come to pass (Bender et al. 2021). Others claim that LLMs are insufficiently complex for consciousness to emerge or lack analogs to human brain function that are critical for consciousness (Aru, Larkum, and Shine 2023). In other words, most researchers claim that LLMs are not currently conscious (or, even stronger, that they cannot be conscious).

I find this conclusion very plausible. But, as I will argue, their arguments do not actually show it. The leading arguments against artificial consciousness rely on controversial assumptions or conceptual confusion about the mind. My aim in this paper is to examine the arguments against artificial consciousness and show that they are lacking. In the next section, I give an overview of core distinctions in the philosophy

of consciousness and say more about what it would mean for consciousness to be artificial. Section 3 identifies and responds to leading arguments against the possibility of artificial consciousness, showing that none of them are decisive. I conclude in Section 4 by considering the broader implications for the science and ethics of artificial consciousness. My conclusion is entirely negative. Even if LLMs are not (or, stronger, cannot become) conscious, we have yet to see a good reason why.

2. Concepts of consciousness

There are several concepts of consciousness at work in the discussion above, and it is useful to lay out some distinctions in the literature. Sometimes 'consciousness' refers to personhood or self-awareness, as in the Google engineer's reports about LaMDA. A different concept of consciousness is the subjective aspect of experience, or *what it's like* (in the sense of Nagel 1974). I will use the phrase 'phenomenal consciousness' to pick out this concept of consciousness. When I say 'conscious' or 'consciousness' I mean the phenomenal sense, unless otherwise specified. To clarify the notion of phenomenal consciousness, it is helpful to contrast the concept with *access consciousness,* or the accessibility of some state for guiding thought and action (Block 1995). A creature is access conscious if it tokens access conscious states, and phenomenally conscious if it tokens phenomenally conscious states. The scientific study of consciousness regularly relies on access consciousness as a proxy for phenomenal consciousness, for instance, when researchers infer phenomenal consciousness from a subject's verbal reports or task performance. But there is good reason to question this assumption. Access is at best an imperfect guide to phenomenal consciousness (Block 1995) and it is an open question whether an AI could be access conscious without thereby being phenomenally conscious.

By *artificial* consciousness, I mean phenomenal consciousness that arises in a non-biological substrate. In one sense of the term, being artificial means being fake or insincere. An artificial smile is not a real smile, for instance. To say consciousness is artificial is not to say that it is fake. Instead, consciousness is artificial in the way that a diamond can be artificial. An artificial diamond produced in a lab is a diamond, with the same carbon structure as one produced by geological forces. Artificial diamonds are simply produced in a different way from natural diamonds. Likewise, artificial consciousness would be genuine consciousness, but produced in a different way or arising from a non-natural process. If an LLM is conscious, they are artificially conscious.

With these conceptual clarifications in place, the answer to the question of whether LLMs are conscious may seem obvious: *of course they are not!* I share this intuition. Why think otherwise?

One reason is drawn from the cases at the start of this section. LLMs have passed (or soon will) a modified version of the Turing Test. Turing (1950) suggested that machines would have achieved intelligence when they could win *The Imitation Game,* that is, when a human interlocutor could not differentiate the machine's response to a question from a human's. A modification of Turing's test could be extended for consciousness. In each of the examples at the start of this section, the user could not distinguish between the AI response and the response of a conscious being. Does this give us good reason to think that AI is, in fact, conscious?

It is fairly obvious that it does not. An LLM is built to mimic the responses of conscious beings, even if they themselves are not conscious. Do they (or might they someday) feel the pang of suffering, the thrill of insight, or the dizziness of falling in love? Reports like those offered by LaMDA, Bing Chatbot and ELIZA give us only defeasible evidence for an answer. Turing's Imitation Game cannot be used as a test for consciousness, no matter how convincing the reports may become (c.f. Schneider 2019; Srivastava et al. 2023).

Another strategy for answering the question is to look at the scientific study of phenomenal consciousness in humans and other animals. The problem is that some of our leading theories of consciousness are consistent with artificial consciousness, or even make conscious LLMs plausible, while other theories seem to rule it out (Butlin et al. 2023). Making matters worse, there is a striking lack of consensus on how to understand consciousness scientifically. This is illustrated by the results of a recent survey on theoretical foundations of consciousness studies among researchers. Francken et al. (2022) surveyed participants at the Association for the Scientific Study of Consciousness (ASSC) on how promising they regarded ten different consciousness theories (Figure 1). The survey showed that 'there is no single theory that the majority of the respondents currently endorse' (Francken et al. 2022, 9).

Some leading theories of consciousness have opposing implications for the feasibility of artificial consciousness. For instance, the Global Neuronal Workspace Theory defines consciousness computationally, and arguably some critical benchmarks of this theory have already been achieved in artificial language agents (Goldstein and Kirk-Gianni, 2024; see section 3.1). By contrast, a view like sensorimotor theory makes it harder to achieve consciousness in a disembodied LLM. Integrated Information

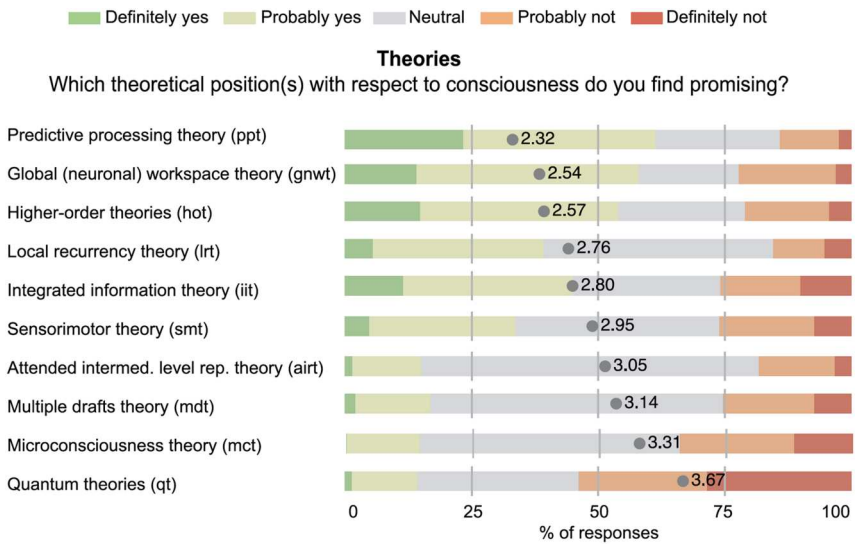

Figure 1. Participants at the 2018 and 2019 ASSC conference indicated whether they found each position promising using a 5-point Likert scale. The mean score is indicated on the bars (Francken et al. 2022).

Theory might rule artificial consciousness out entirely by denying that consciousness is a computational kind (Tononi et al. 2016). Since there is no consensus as to which, if any, of these theories is correct, we cannot simply look to the scientific study of consciousness for an answer to whether LLMs are conscious.

Given that the science does not provide a clear answer, why do many researchers continue to believe that LLMs are not and will not become conscious? I will next turn to theoretical arguments against the possibility of conscious LLMs. As I will show, however strong our intuitions against artificial consciousness might be, they lack rational grounds.

3. Arguments against artificial consciousness – and what's wrong with them

My plan in this section is to present four influential arguments against the possibility of artificial consciousness in an LLM, and to highlight some problems with those arguments. They are: (1) LLMs are just parroting human speech, (2) LLMs are not capable of semantic understanding, (3) LLMs are insufficiently complex for consciousness to emerge and (4) LLMs are too dissimilar from us in a variety of ways. I'll argue that none of these arguments give us a good reason to reject artificial consciousness in an LLM.

3.1. 'LLMs are just parrots'

Emily Bender, Timnit Gebru and colleagues coined the phrase 'stochastic parrots' to describe what LLMs do (Bender et al. 2021). Roughly and briefly, LLMs imitate human speech like a parrot might do, except they introduce a degree of randomness (stochasticity) into the response. To borrow another of their analogies, LLMs are like a calculator for the English language. They are, on this view, relatively simplistic tools that do not approach general intelligence or consciousness. Instead, an LLM is a string prediction system that determines which word comes next in a series. While it is not the primary aim of their paper, the concept of the stochastic parrot motivates an objection to artificial consciousness, and I will focus here on critiquing that objection.

Why wouldn't a stochastic parrot be conscious? I consider three possibilities and address each in turn. The first possibility can be dealt with quickly. Perhaps the problem is that, like a parrot, an LLM's speech is merely derivative. LLMs plagiarize human-made text without generating new content. But this fact alone should not speak against their being conscious. We, too, draw on experience to repeat and recombine ideas. More of human thought is derivative than we would probably like to admit. An LLM simply takes this to its logical extreme, since none of the content produced by an LLM is novel. But it is nonetheless possible to imagine a conscious mind with entirely derivative contents.[1] It is implausible to maintain that an LLM cannot be conscious simply because the contents of its consciousness would be derivative.

The second possibility is that an LLM employs the wrong kind of process to produce consciousness. Like parrots, LLMs are mimics. Even when they report being conscious, they generate this report using a different process than humans would use. More specifically, an LLM predicts which word comes next in a sentence based on data in the training set and later fine-tuning. This process is relevantly different from the process by which humans produce verbal or written language, and such differences should lead us to question whether an LLM is capable of consciousness.

But while early chatbots employed a simplistic strategy for word prediction, this is not how all LLMs operate today. Some LLMs have been incorporated into more sophisticated tools that are modeled after human cognitive architecture, and have demonstrated indicators of

[1]For an example from science fiction, see Louisa Hall's 2015 novel, 'Speak'.

consciousness such as attention, agency and memory (to name just a few) (Butlin et al. 2023; Chalmers 2023).

Consider an example explored in detail by Goldstein and Kirk-Gianni (2024) based on work by Park et al. (2023). Goldstein and Kirk-Gianni focus on *language agents,* a type of AI that combines an LLM with an algorithm that mimics the functional architecture of an agent. The agent stores beliefs, desires, and plans, makes observations, and selects actions, all according to principles of folk psychology. The LLM serves as the information processing system for the agent, enabling it to translate a percept into a memory, or reason about the rationality of some action. Language agents have a 'memory stream' of representations that connect perception and action and allow the system to plan and reason. Goldstein and Kirk-Gianni argue that the architecture of a language agent is analogous in several ways to a global workspace in humans (though they also note several disanalogies; see also VanRullen and Kanai 2021). In the scientific study of consciousness, the Global Neuronal Workspace Theory (GNWT) is among the leading theories that attempt to explain how consciousness arises in humans. In brief, GNWT says that a state is conscious in virtue of being globally accessible to the system for use in thinking, reasoning, or guiding behavior. In a slogan, consciousness is '*fame in the brain*' (Dennett 2001, 224). But language agents also have globally accessible states, what we might call *fame in the mainframe*. The processing analogs between humans and AI give us some reason to think that if the global workspace explains consciousness in humans, then language agents may also have conscious states.

This example is indicative of a general point: some of today's AIs employ a similar functional architecture to human cognition, rather than simply mimic its output. Of course, there are differences in functional architecture between natural and artificial minds, and those differences matter for whether an LLM is likely to be conscious or not. My point is not to argue that there are no differences, but rather, to show that the differences are overstated in the arguments against artificial consciousness. Instead, a closer look at how LLMs function reveals that they are increasingly modeled after human cognition, and some researchers think that artificial consciousness is currently technologically feasible in the hardware available today (e.g. Butlin et al. 2023).

The third possibility is the trickiest to address. Even given similarities with human cognition, the objector might maintain that words and

sentences have no meaning for an LLM. Like a parrot, they speak without understanding. Understanding language requires more than just modeling it; it also requires modeling the world. Only then would language have content or meaning for an LLM. I think this objection is based on a fallacy, which I'll explain in detail in the next section. Roughly and briefly, the mistake is to assume that because LLMs are built to mimic human speech, they are *nothing more* than mimics. Instead, it may turn out that the most effective strategy for modeling human language involves learning to model the world, as well.

3.2. 'An LLM does not understand'

In May of 2024, Google's experimental AI Overview recommended improving a pizza recipe by mixing 'about 1/8 cup of Elmer's glue in with the sauce. Non-toxic glue will work.' Another user was told to eat rocks. These and other mistakes help to build the case that Google's AI Overview – and perhaps LLMs in general – have no understanding of what they are saying. Anyone with a basic understanding of English would not give answers like these.

These examples illustrate a problem with *pragmatic understanding*: the ability to make a type of inference known as implicature, which relies on background beliefs or knowledge. LLMs are inconsistent in their ability to draw this type of inference, and their capacity appears to depend on how the model is fine-tuned (Ruis et al. 2023). This alone should not count against being conscious. Indisputably conscious creatures also fail to have pragmatic understanding as well, such as young children or people suffering disorders of semantic memory (a toddler might also add glue to your pizza!). Understanding in the pragmatic sense is probably not required for consciousness, whether natural or artificial.

But there is another sense of 'understanding' as semantics, which poses a more difficult problem for the possibility of conscious LLMs. Searle (1980) gives a canonical version of this argument against the possibility of *Strong AI,* or against the view that a computer program could have a human-like mind that understands and thinks much as we do. Suppose an LLM is a purely formal or *syntactical* system, an algorithm that completes sentence strings according to the rules and conventions of the English language. Searle argues that a purely syntactic system will lack *semantic* content: an understanding of what those sentences are about. On his view, while an AI might perfectly mimic human

language or behavior, it can never replicate human mental capabilities for meaning, understanding or consciousness.

There are several reasons to think that some of today's LLMs could understand, or perhaps already do. Consider a few possibilities for ascribing understanding to a system, or meaning to a state, like a brain state or a computational state. On an *externalist* approach, a state has content in virtue of its causal relations to the world. For instance, a brain state is about a coffee cup just if that state serves to represent the presence of coffee cups. On this type of view, an LLM's state (e.g. a linguistic representation) has content only provided that that representation is appropriately causally connected to objects or properties in the environment.

Whether today's LLMs satisfy this condition depends on what we mean by an *environment*. If we mean the external world, then an LLM would require perceptual systems to track objects or events. Some LLMs already have analogs of perception, such as an LLM that listens and responds to the user's queries by voice. Robotics combines machine learning with artificial bodies and perceptual apparatus that bring AI into contact with the environment, and this technology is already used in a number of domains, from self-driving cars to surgical tools. If we include virtual environments, the possibility of AI semantics looks even more likely. Language agents, for instance, have a perceptual system that receives linguistic descriptions of the world (Park et al. 2023). Experimenters with Google's DeepMind recently created a virtual rodent that used deep reinforcement learning run on a neural network to imitate rodent behavior in a virtual body and world (Aldarondo et al. 2024). Each of these examples suggests that some LLMs and AI already do track objects or events in a real or virtual environment. If our brain states have meaning in virtue of their causal relation to objects or events in the world, then an AI's states might have meaning, too.

Of course, tracking is not sufficient for a system to understand what it tracks. A thermometer tracks the temperature of its surroundings. It does not follow that the thermometer's states have meaning *for the thermometer*. Whether states have meaning for the system depends not only on causal relations to the environment, but also on how the system uses those representations in internal processing. But each of the examples above does more than just track the virtual or real environment. They also make use of representations for AI analogs of thought and action, as in the case of the language agent discussed in section 3.1. This makes it more plausible that those states serve to represent objects in the world for those systems.

An alternative to externalism is causal role semantics, or the view that a state has content in virtue of its relations to other states within the system. According to this view, a brain state represents 'coffee cup' just if it is appropriately related to other states that represent relevant concepts like 'coffee,' 'drink,' or even mental imagery. But this is precisely what language models do: they model natural language and link related concepts, then draw on these links to produce appropriate responses to a query. When an LLM is combined with a language agent as in the studies discussed by Goldstein and Kirk-Gianni (2024), these linguistic representations also have a distinctive causal role in guiding the system's behavior, goals and desires. In other words, a language agent's state representing 'coffee cup' could have a similar causal role to the role of a brain state representing 'coffee cup' in us. If the brain state has meaning in virtue of playing this causal role, then the LLM's representation probably has meaning, too.

A third option is drawn from cognitive science and the study of world models. In human cognition, a world model is a representation that replicates the structure of an object or event in the real world, which is used to guide thought and behavior. World models are critical for human understanding, allowing us to reason and draw inferences that reflect general knowledge of the world and how it works. For instance, to borrow an example from Yildirim and Paul (2024), when we judge that it is easier to balance a ball on a box than a box on a ball, we use a world model of 3-D objects and intuitive physics to compare these scenarios. Yildirim and Paul review evidence that some LLMs spontaneously generate world models in order to produce more successful text strings. In one study that they describe, researchers trained an LLM called Othello-GPT (similar to GPT-4) to predict moves in the board game Othello, a two-player game of strategy. Othello-GPT learned to correctly predict legal moves based on a prior set of moves. More surprisingly, researchers could decode the state of the entire board from intermediate-level activations in the model using a linear decoder (Makelov, Nanda, and Lange 2024). In other words, Othello-GPT seems to have predicted specific moves by constructing a world-model of the entire board, similar to how humans construct world models in order to reason, plan and talk about the world. This and other studies suggest that some LLMs have rudimentary world models already, and may develop more sophisticated world models in the future.

The arguments in this section are not intended to establish that LLMs can understand, but rather, that the claim that they *cannot* relies on a

fallacy (also pointed out by Chalmers 2023). It's of course true that LLMs are designed to produce language, not to understand it. But it does not follow that an LLM only produces language. A system might develop many other capacities that are helpful for achieving its primary goal. For instance, as Chalmers has pointed out, according to the theory of evolution, the goal of living organisms is to reproduce. Life has evolved many varied abilities that are helpful for achieving this goal, from bodies that locomote to uniquely human achievements like art and music. Even if life aims at reproduction, it does not follow that it *only* aims at reproduction. The same may be true for LLMs. For instance, an LLM might develop world models or representations for tracking the world, which help it to produce appropriate responses to a query and avoid mistakes of pragmatic understanding (such as recommending glue in pizza sauce). Since these capacities for tracking or modeling the world are considered critical to human understanding, we should likewise attribute understanding to an LLM to the extent that they demonstrate these same capacities.

3.3. 'LLMs are not complex enough to be conscious'

Another challenge to artificial consciousness begins with the assumption that consciousness arises as an emergent property of a complex system. For example, Aru and colleagues recently argued against conscious LLMs from a neuroscience perspective. Among the reasons they give is that, 'it might not be possible to abstract consciousness away from the organizational complexity that is inherent within living systems but strikingly absent in AI systems' (2023, 14). In a nutshell, the *complexity objection* states that LLMs are not complex enough for consciousness to emerge.

The problem with the complexity objection is twofold. First, (and as the authors above note), it is unclear what kind of complexity is required for consciousness. For instance, does consciousness emerge from complexity within a neuron, or between neurons? This matters because LLMs are complex systems, and they do in fact replicate some of the organizational complexity of living organisms. A *complex system* involves multiple interacting subsystems connected via feedback loops, such that new properties emerge from the interaction of those subsystems. Examples of complex systems range from storm systems to economic markets to living organisms. While early LLMs operated sequentially, and therefore would not be considered complex systems, more recent models demonstrate considerable complexity. OpenAI's GPT-4, for instance, is a multi-

layered neural network modeled on the human brain, with roughly 176 billion neurons spread over 100 layers, enabling 100 trillion connections (by comparison, the human brain has a paltry 100 billion neurons). This architecture allows subsystems to operate in parallel on massive quantities of data, with feedback loops that enable higher-level outputs to impact lower-level processing. In other words, LLMs like GPT-4 are complex systems.

Complex systems often have properties that emerge only at the level of the whole, rather than the parts or their interactions. A property of a system is considered *emergent* when it cannot be predicted based on the low-level facts about that system (Anderson 1972). As language models have grown larger (in terms of their computational capacity, number of model parameters and training dataset size), new abilities have already emerged that are surprising to researchers. For instance, Wei et al. (2022) suggest that reasoning is an emergent ability of some current LLMs. They define 'emergent ability' as an ability that is present in larger-scale models but not smaller-scale models. Some LLMs, like LaMDA, can learn to engage in multi-step reasoning if given an appropriate prompt, but only when the model surpasses a specific scale threshold (Wei et al. 2022, 5). Reasoning emerges as LaMDA becomes more complex.

There is ongoing debate over whether the abilities of LLMs are genuinely emergent. That is, perhaps if we better understood how an LLM like LaMDA performs reasoning tasks, we would expect this ability to improve as the model grows larger. It would therefore be predictable on the basis of low-level facts. Moreover, so-called emergent abilities in LLMs might be a result of choosing metrics that mask incremental task improvements as the model grows, improvements which would make the large-scale models' abilities unsurprising (Schaeffer, Miranda, and Koyejo 2023).

The possibility of conscious LLMs does not turn on the outcome of this debate because consciousness may also be predictable, at least in principle. Which abilities are predictable will change as researchers gain knowledge of a system and how it functions. As defined in the context of LLMs, emergence is a matter of what researchers can deduce from their knowledge of small-scale models. Philosophers sometimes call this type of emergence *weak emergence* (e.g. Chalmers 2006; Clark 2001). Weak emergence is epistemological rather than ontological. It concerns what can be known on the basis of low-level facts. In the case of LaMDA, although researchers cannot predict reasoning abilities on the basis of the behavior of small-scale models, with more understanding

or better metrics these abilities may become predictable. Likewise, when neuroscientists describe consciousness as emerging from the complexity of the brain and body, they mean that it emerges in the weak sense. For instance, the authors who appeal to complexity as an argument against conscious LLMs seem committed to the idea that by understanding the complexity of living systems, we can predict when consciousness will or won't arise.

The upshot of this discussion is that today's LLMs plausibly have emergent abilities in the weak sense. New and surprising abilities emerge as models scale up, like the ability to reason. Given that LLMs already demonstrate some emergent abilities, this makes it more plausible that consciousness could emerge as well.

Second, even if today's LLMs do not have emergent abilities, there is reason to think that LLMs will grow more complex over time, and that new abilities will emerge as they do. We have numerous examples of systems that become more complex over time. For example, consider a mundane case of emergence: rush hour traffic. In the early morning, the movements of vehicles and pedestrians are relatively simple, growing more complicated as the number of travelers increases. As the start of the workday approaches, the movement of traffic becomes complex, with subsystems like pedestrians, trams, and cars interacting in ways that mutually impact each other. This interplay results in emergent properties of the system as a whole, such as traffic jams. Other examples of systems that grow in complexity abound, from financial markets to storm systems. As LLMs scale up, it is plausible that they will increase in complexity and new properties will emerge, much as a traffic jam emerges at rush hour. This possibility is made more likely as researchers seek to create an AI with general intelligence by combining subsystems for human-like mental capacities such as agency, reasoning or perception.

3.4. 'LLMs are not like us'

The last argument that I will consider takes a number of different forms, but the general strategy is the same. These arguments are motivated by the numerous ways that LLMs are different from us. For instance, they are not biological, they lack a unified self, they are not embodied, they don't perceive the world as we do, and they do not have agency. This is not an exhaustive list. There are many ways that LLMs are not like us, and some of the features they lack are thought to be necessary for consciousness.

The form of the objection is that, for some posited necessary feature *x,* LLMs lack *x,* and so LLMs are not conscious.

I do not want to deny that there are many thoughtful and compelling arguments in defense of each feature as a criterion for consciousness. Some of these features may be universal to all known conscious creatures, like embodiment in a biological body. But however strong the arguments may be, each remains controversial as a hallmark of consciousness. Rather than address each feature one by one, I want to point to a general worry for this approach as an objection to artificial consciousness. The worry is that it commits to *chauvinism*, the assumption that human (or perhaps animal) consciousness provides the conditions for consciousness in general. Chauvinism stacks the deck against the possibility of artificial consciousness, and could lead us to deny consciousness to a system that plausibly has it.

In arguments both for and against artificial consciousness, one popular approach is to look for hallmarks of consciousness in us (Butlin et al. 2023; Chalmers 2023). For instance, Butlin and colleagues point to several indicators that are 'good reasons to think that one system is much more likely than another to be conscious, and this can be relevant to how we should act' (12). But while it's true that we can use theories of consciousness to posit benchmarks or indicators, we don't know which (if any) of these are actually essential factors for consciousness. The problem is that this strategy begs the question for or against artificial consciousness, depending on one's starting assumptions. Suppose an LLM lacks the features of consciousness listed above, like agency or embodiment. This could show, on the one hand, good evidence against artificial consciousness. It could also show, on the other hand, that our starting assumptions about consciousness are chauvinistic, denying consciousness to something that has it. In order to decide between these two options, we'd need to know with confidence one of two things. We'd either need to know on independent grounds whether the AI mind is or is not conscious, or we'd need to know which theory of consciousness is likely to be true. In other words, in order to interpret the evidence as a reason to believe a system is conscious, we must make assumptions that presuppose a solution to the very problem we are trying to solve.

For centuries, species chauvinism led to the denial of consciousness to other animals that plausibly have it, because they don't share other features of humanity, such as the ability to speak or reason. This is now widely acknowledged as a mistake (Low 2012). Current consensus is that many other animals besides humans are conscious. The philosopher Jeremy

Bentham wrote compellingly of the moral risks of species chauvinism when considering the status of non-human animals in the eighteenth century. As he wrote, when determining a being's moral status, 'the question is not, *Can they reason?* nor, *Can they talk?* but, *Can they suffer?*' (Bentham 1823, Ch.17 n.122). In other words, consciousness determines whether a creature matters morally. Even an unintelligent artificial system that is conscious would raise difficult questions about whether to afford that system rights and protection under the law, and how it should be treated.

LLMs present a mirror image of the problems raised by other animals. In the case of LLMs, they can talk and perhaps even reason, but they are not alive or embodied and embedded in the world via robust sensorimotor connections. If consciousness is a computational kind, as assumed by most of the leading theories of consciousness, then these differences may not matter any more than lacking the ability to speak matters for animal consciousness. When asking whether an LLM is conscious, we should seek to avoid the chauvinistic assumptions of the past. An LLM mind might be radically different from our own, and yet there is still something it's like. Keeping this question open matters not only for theoretical reasons, but also for recognizing our moral obligations to the minds we might inadvertently create.

4. Facing the hard problem

Advancements in AI present us with a new version of the 'hard problem' of consciousness (Chalmers 1995; Schneider 2019), one with immediate moral weight. Why do certain physical processes give rise to consciousness, and could consciousness arise in an artificial system like an LLM? In this paper, I've argued that there is no knock-down argument against conscious LLMs, and the possibility of artificial consciousness needs to be taken seriously.

Facing this problem might lead to a breakthrough in consciousness studies. It may even overturn the widespread commitment to computationalism in philosophy of mind, by throwing new light on its counterintuitive consequences when applied to LLMs or other AI. More generally, advancements in AI stand to improve our understanding of consciousness and its place in the physical world, by allowing us to experiment with the implementation of computational models in AI systems and see if indicators of consciousness emerge. In turn, the scientific study of consciousness illuminates the basis of consciousness in biological systems, generating models that can be tested in increasingly

sophisticated AI. As these fields mutually impact each other, we stand on the cusp of a revolution in our understanding of consciousness.

Yet there is reason to resist this revolution. Artificial consciousness would bring into existence a new type of morally relevant beings, which need to be considered in how we should act (see Schwitzgebel and Garza 2020). The resources already demanded by LLMs are considerable (Bender et al. 2021), and as climate change continues to exacerbate resource scarcity, the conflict between AI and animal demand for resources will escalate as well. Might an AI matter in this conversation, not just for how it affects us, but in itself? How we answer this question depends partly on whether that AI is conscious or not. If it is, then it is plausible that an AI's needs should be given at least some consideration, if not equal consideration to our own.

The hard problem is no longer just a theoretical puzzle. It is a real-world problem with practical implications for how we respond to emerging AI technologies and what we owe to the minds we might create.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Aldarondo, D., J. Merel, J. D. Marshall, L. Hasenclever, U. Klibaite, A. Gellis, Y. Tassa, G. Wayne, M. Botvinick, and B. P. Ölveczky. 2024. "A Virtual Rodent Predicts the Structure of Neural Activity Across Behaviours." *Nature* 632 (8025): 594–602. https://doi.org/10.1038/s41586-024-07633-4.

Anderson, P. W. 1972. "More Is Different." *Science* 177 (4047): 393–396. https://doi.org/10.1126/science.177.4047.393.

Aru, J., M. E. Larkum, and J. M. Shine. 2023. "The Feasibility of Artificial Consciousness Through the Lens of Neuroscience." *Trends in Neurosciences* 46 (12): 1008–1017. https://doi.org/10.1016/j.tins.2023.09.009.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 😈." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922.

Bentham, Jeremy. 1823. *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

Block, N. 1995. "On a Confusion About a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–247. https://doi.org/10.1017/S0140525X00038188.

Butlin, P., R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, et al. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." In arXiv e-prints. https://doi.org/10.48550/arXiv.2308.08708.

Chalmers, D. 1995. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200–219.

Chalmers, D. 2006. "Strong and Weak Emergence." In *The Re-emergence of Emergence*, edited by P. Clayton and P. Davies, 244–256. Oxford: Oxford University Press.

Chalmers, D. J. 2023. "Could a Large Language Model Be Conscious?" *Boston Review*, August 9. https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/.

Clark, A. 2001. *Mindware: An Introduction to the Philosophy of Cognitive Science*. Oxford: Oxford University Press.

Dennett, D. 2001. *Sweet Dreams*. Cambridge: MIT Press.

Francken, J. C., L. Beerendonk, D. Molenaar, J. J. Fahrenfort, J. D. Kiverstein, A. K. Seth, and S. van Gaal. 2022. "An Academic Survey on Theoretical Foundations, Common Assumptions and the Current State of Consciousness Science." *Neuroscience of Consciousness* 2022 (1): niac011. https://doi.org/10.1093/nc/niac011.

Goldstein, S., and C. D. Kirk-Giannini. 2024. "A Case for AI Consciousness: Language Agents and Global Workspace Theory." *arXiv*. https://doi.org/10.48550/arXiv.2410.11407.

Huckins, Grace. 2023. "Minds of Machines: The Great AI Consciousness Conundrum." *MIT Technology Review*, October 16. https://www.technologyreview.com/2023/10/16/1081149/ai-consciousness-conundrum/.

Low, Phillip. 2012. "Cambridge Declaration on Consciousness." In *Proceedings of the Francis Crick Memorial Conference*, 1–2.

Makelov, A., N. Nanda, and G. Lange. 2024. "Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control." *arXiv*.. https://doi.org/10.48550/arXiv.2405.08366.

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

Park, J. S., J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. 2023. *Generative Agents: Interactive Simulacra of Human Behavior*, April 7, 2023. https://arxiv.org/abs/2304.03442v2.

Roose, K. 2023. "A Conversation With Bing's Chatbot Left Me Deeply Unsettled." *The New York Times*, February 16. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html. Full Transcript: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html.

Ruis, L., A. Khan, S. Biderman, S. Hooker, T. Rocktäschel, and E. Grefenstette. 2023. *The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs* (arXiv:2210.14986). https://doi.org/10.48550/arXiv.2210.14986.

Schaeffer, R., B. Miranda, and S. Koyejo. 2023. *Are Emergent Abilities of Large Language Models a Mirage?* arXiv.Org, April 28. https://arxiv.org/abs/2304.15004v2.

Schneider, S. 2019. *Artificial You: Ai and the Future of Your Mind*. Princeton: Princeton University Press. https://doi.org/10.2307/j.ctvfjd00r.

Schwitzgebel, E., and M. Garza. 2020. "Designing AI with Rights, Consciousness, Self-Respect, and Freedom ." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao, 459–479. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0017.

Searle, J. R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–424. https://doi.org/10.1017/S0140525X00005756.

Srivastava, A., A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, et al. 2023. *Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models*. arXiv:2206.04615. https://doi.org/10.48550/arXiv.2206.04615.

Tiku, N. 2022. "The Google Engineer Who Thinks the Company's AI Has Come to Life." *Washington Post*, June 11. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Tononi, G., M. Boly, M. Massimini, and C. Koch. 2016. "Integrated Information Theory: From Consciousness to its Physical Substrate." *Nature Reviews Neuroscience* 17 (7): 450–461. https://doi.org/10.1038/nrn.2016.44.

Turing, A. 1950. "Computing Machinery and Intelligence." *Mind; A Quarterly Review of Psychology and Philosophy* 59 (236): 433–460. https://doi.org/10.1093/mind/LIX.236.433.

VanRullen, R., and R. Kanai. 2021. "Deep Learning and the Global Workspace Theory." *Trends in Neurosciences* 44 (9): 692–704. https://doi.org/10.1016/j.tins.2021.04.005.

Walker, L. 2023. "Belgian Man Dies by Suicide Following Exchanges with Chatbot." *The Belgian Times*, March 8. https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, et al. 2022. *Emergent Abilities of Large Language Models*. arXiv:2206.07682. https://doi.org/10.48550/arXiv.2206.07682.

Yildirim, I., and L. A. Paul. 2024. "From Task Structures to World Models: What do LLMs Know?" *Trends in Cognitive Sciences* 28 (5): 404–415. https://doi.org/10.1016/j.tics.2024.02.008.