Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-024-03553-w OPEN

A clarification of the conditions under which Large language Models could be conscious

Morten Overgaard[1]✉ & Asger Kirkeby-Hinrup[1,2]

With incredible speed Large Language Models (LLMs) are reshaping many aspects of society. This has been met with unease by the public, and public discourse is rife with questions about whether LLMs are or might be conscious. Because there is widespread disagreement about consciousness among scientists, any concrete answers that could be offered the public would be contentious. This paper offers the next best thing: charting the possibility of consciousness in LLMs. So, while it is too early to judge concerning the possibility of LLM consciousness, our charting of the possibility space for this may serve as a temporary guide for theorizing about it.

[1] Department of Clinical Medicine - Center of Functionally Integrative Neuroscience, Aarhus University, Aarhus, Denmark. [2] Department of Philosophy – Theoretical Philosophy, Lund University, Lund, Sweden. ✉email: morten.storm.overgaard@cfin.au.dk

Large Language Models (LLMs) are sophisticated artificial neural networks whose weights are trained on hundreds of billions of words from the internet, including language conversations between conscious humans with 'real' agency. Users that interact with LLMs are provided with a fascinating language-based simulation of a natural language interaction. Because LLMs have been trained on conversations, in which (actual) humans describe and express in different ways the peculiar inner life we associate with conscious experience, the LLMs are capable of giving descriptions and expressions of such an inner life that are practically indistinguishable from the that of humans. To the public, this has made manifest the lack of clarity about what it means to have agency and to be conscious. In public discourse on LLMs an uncertainty about whether they could be conscious drives many of the worries expressed by politicians, the public audience, and laypeople alike. This uncertainty thrives in part because we — as a scientific field — have yet to understand consciousness as well.

In interdisciplinary consciousness studies, researchers are today far from consensus about how to explain consciousness theoretically. In fact, there is an extended and ongoing debate in the field about what the words we use to describe and theorize about consciousness even mean. Therefore, we have no strong theoretical guidance to understand whether LLMs are — or can be — conscious either (Aru et al. 2023; Chalmers, 2023b). Several recent scientific articles have assumed that LLMs are not conscious (Chalmers, 2023a; Colombatto and Fleming, 2023; Dodig-Crnkovic, 2023) and that we therefore can conclude that the ability to converse can happen unconsciously. At the same time, others, as mentioned above, have suggested the exact opposite. However, any such assumption is a theoretical choice not supported by any empirical evidence.

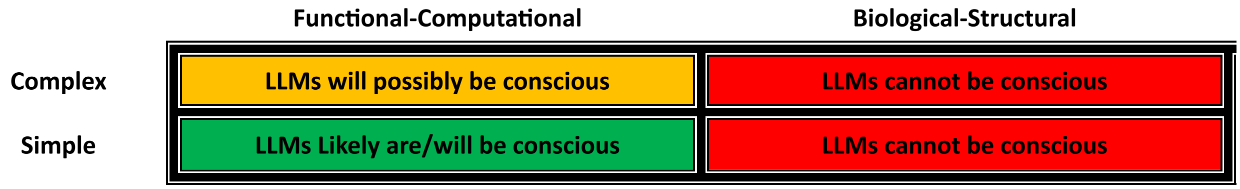
Recently, it has been suggested in media as well as in the scientific literature that there is evidence to suggest that consciousness is common – not just in the biological domain but in any domain where information is integrated (Tononi et al. 2016). It is however very premature to make such a claim based on empirical science. This goes not only for the integrated information theory, but for any contemporary theory of consciousness. How to measure consciousness remains one of the most prominent unsolved problems around (Bayne et al. 2024). Since consciousness seems to be a central component of human life, we have a vested interest in finding objective and reliable biomarkers of consciousness in humans (not the least for clinical reasons). Regarding the topic at hand, clearly, if we only understood how consciousness comes about in humans, it would be much easier to determine what it would take for a machine to be conscious, and whether this is even possible in the first place. But we currently do not know how consciousness comes about in humans, therefore this is not a feasible approach.

It is a strong intuition both in science and common sense that being conscious of something makes a cognitive difference *for* the subject. Yet, predominant models in cognitive neuroscience have not been able to conceptually — or empirically — identify a particular cognitive function (or set of functions) for which consciousness is necessary. This also goes for language and linguistic capabilities. So, at present, there is no objective way of determining whether any given function or action an LLM may perform in fact is associated with consciousness, making this approach unfeasible (see also Bayne et al. 2024).

The brief analysis above seems to show that the debate is stuck. There is no empirical method available to determine if LLMs are conscious, and a theoretical conclusion on the matter will be based on a choice or an assumption, thus either depending on arbitrary assumptions or ending as a circular argument. The problem is familiar to consciousness researchers but is echoed in previous debates about consciousness in e.g. insects, animals, infants, non-communicating patients in coma or vegetative state, and even in neurotypical human adults, as exemplified in the "other minds problem" from the philosophy of mind.

If there is a way forward to directly measuring consciousness, we must identify the questions that need answering before we can find it. For instance, it must be determined whether the core correlate of consciousness is biological/structural in nature or rather functional/computational. Naturally, there will always be biological *and* functional correlates at the same time, yet any theory must argue that consciousness exists because of something that is either biological or functional in nature, so that other correlates are spurious or secondary. In recent years, much attention has been given to classifying explanatory targets and mechanisms of extant theories in various, but similar terms (Doerig et al. 2020; Fahrenfort and van Gaal, 2021; Sattin et al. 2021; Schurger and Graziano, 2022; Signorelli et al. 2021).

Two of the important questions we need to raise are expressed in the matrix below: Consciousness is either realized by certain biological structures or by certain functions/computations, and consciousness is either realized by simple/low-level phenomena or by complex/higher-level phenomena.



	Functional-Computational	Biological-Structural
Complex	LLMs will possibly be conscious	LLMs cannot be conscious
Simple	LLMs Likely are/will be conscious	LLMs cannot be conscious

Segmenting the landscape according to the functional-biological and simple-complex distinctions gives us a handle on the conditions under which LLMs may be conscious (now or in the future). Because this way of segmenting the theoretical landscape has narrowed the possible positions into a two-by-two matrix, where each space predicts the prospects of artificial consciousness, it allows for better (but still underdetermined) generalization than when considering each theory in isolation. This approach has the advantage of offering a theory-neutral mapping of the possibility space for LLM consciousness. Therefore, we will next briefly consider each of the two dimensions in a little more detail.

The biological-functional distinction

Fundamentally, either consciousness is associated with a physical structure or consciousness is associated with function. These two "types" of theories come in many versions, depending on the specifics of what consciousness is taken to be reducible to, identical with, or different from, and if or how it is anchored in some specific structure or function.

Researchers who associate consciousness with biological structures often have one or more neural structure(s) in mind. From this perspective, an organism is conscious under the condition that it has a specific neural structure (biological foundation). This thinking is evident in several currently influential theories, e.g. in integrated information theory, where human consciousness is literally identical to the most complex cluster of interconnected information in a brain (Tononi et al. 2016)[1]. If consciousness depends on biological structures, LLMs will never be conscious because they are not instantiated in the 'right' material (c.f. Searle 1980). Now, one might object that if an LLM was instantiated in the right biological material (a biological computer of the future) then it would in fact be conscious. This, however, is misguided with respect to the point we are making here. Yes, in such a case it would certainly be conscious, but it would not be conscious in virtue of that which makes it an LLM, it would be conscious in virtue of the material in which it was instantiated.

Researchers who associate consciousness with functional properties typically conceive of consciousness as analogous to computer software that needs some hardware to run (c.f. what Chalmers (2011) calls the "thesis of computational sufficiency"[2]). From this perspective, any physical structure (e.g. brains or arrays of silicon chips) with the necessary – currently unknown – characteristics to run the 'right' software will be able to realize consciousness. Accordingly, if consciousness depends on functional characteristics, LLMs can be conscious if they run the 'right' software.

Complexity matters

The other parameter in our chart maps whether consciousness depends on *complex* biological structures or functions, or whether it merely requires *simple* structures or functions.

Some researchers propose that one or a few functions are able to realize consciousness. One such example is higher-order thought theory which argues that consciousness merely requires an (itself unconscious) thought about a first-order content (such as a visual stimulus) (Brown et al. 2019). Consequently, any system with the right kind of metacognitive abilities may be conscious. Nothing indicates that LLMs have actual metacognition. In principle, however, nothing would prevent an artificial system from having higher-order thought-like states, so consciousness in artificial systems is not ruled out from this perspective.

If consciousness is correctly associated with any kind of function, LLMs may already be close to being conscious, and future developments are likely to lead to conscious artificial systems. If consciousness is associated with extremely *simple* functions, without knowing it, we may have created consciousness in artificial systems long before LLMs appeared.

The complexity aspect plays out differently if consciousness is in fact associated with biological structures. On this view our current LLMs will never be conscious because they are not instantiated in the 'right' material. However, it is possible that the 'right' material (e.g. brains) exists not only in humans, so consciousness may be very widespread in nature. On this view, the complexity aspect maps onto *how* widespread consciousness in fact is. If the structure must be very complex, fewer species will be conscious, if the structure is very simple, consciousness will be abundant in biological beings (Wiese and Friston, 2021). An example of the former would be a theory positing the need for exascale quantum computation at ambient temperatures in combination with specific properties of cortical neurons and the neuronal membranes (Stoll, 2022). An example of the latter would be the ability to integrate information (Tononi, 2005).

Conclusions

While the above presents the available options of a highly complicated and diverse theoretical landscape in a simple matrix, we do not suggest that answering the question is simple. We have a difficult and long journey ahead. It should be clear, however, that it is premature to draw any conclusions about the possibility of LLM consciousness. Furthermore, while our ability to reach a final conclusion may still be far in the future, it is possible that we may be able to rule out some positions in the matrix before that (if incoming and future data is serendipitous). In this sense, the matrix may serve as a best-we-have-so-far understanding and may be useful to parse incoming data. Nevertheless, until serious progress is made, and as long as theories rooted in all four positions in the matrix can explain all or most of the available scientific data (for discussion, see Butlin et al. 2023), it will be an unscientific enterprise to draw conclusions about consciousness in LLMs or any other artificial system. In the words of Uriah Kriegel: "When two theories are perfectly empirically equivalent, there is an important sense in which choosing among them on the basis of superempirical virtues is a nonscientific endeavor." (Kriegel, 2020, p. 273).

Received: 30 April 2024; Accepted: 1 August 2024;
Published online: 12 August 2024

Notes

[1] Observe, this position may still be compatible with Panpsychism.

[2] Observe, Chalmers himself is sceptical about LLM consciousness.

References

Aru J, Larkum ME, Shine JM (2023) The feasibility of artificial consciousness through the lens of neuroscience. Trends Neurosci 46(12):1008–1017

Bayne T, Seth AK, Massimini M, Shepherd J, Cleeremans A, Fleming SM, Malach R, Mattingley JB, Menon DK, Owen AM, Peters MAK, Razi A, Mudrik L (2024) Tests for consciousness in humans and beyond. Trends Cogn Sci 28(5):454–466. https://doi.org/10.1016/j.tics.2024.01.010

Brown R, Lau H, LeDoux JE (2019) Understanding the higher-order approach to consciousness. Trends Cogn Sci 23(9):754–768. https://doi.org/10.1016/j.tics.2019.06.009

Butlin, P, Long, R, Elmoznino, E, Bengio, Y, Birch, J, Constant, A, Deane, G, Fleming, SM, Frith, C, & Ji, X (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*

Chalmers DJ (2011) A computational foundation for the study of cognition. J Cogn Sci 12(4):325–359

Chalmers, DJ (2023a) Could a large language model be conscious? *arXiv preprint arXiv:2303.07103*

Chalmers DJ (2023b) Does thought require sensory grounding? From pure thinkers to large language models. Proc Address Am Philos Assoc 97:22–45

Colombatto, C, & Fleming, S (2023) Folk psychological attributions of consciousness to large language models

Dodig-Crnkovic, G (2023) How GPT Realizes Leibniz's Dream and Passes the Turing Test without Being Conscious. Computer Sciences & Mathematics Forum

Doerig A, Schurger A, Herzog MH (2020) Hard criteria for empirical theories of consciousness. Cogn Neurosci 12(2):41–62

Fahrenfort JJ, van Gaal S (2021) Criteria for empirical theories of consciousness should focus on the explanatory power of mechanisms, not on functional equivalence. Cogn Neurosc 12(2):93–94. https://doi.org/10.1080/17588928.2020.1838470

Kriegel U (2020) *The Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press

Sattin D, Magnani FG, Bartesaghi L, Caputo M, Fittipaldo AV, Cacciatore M, Picozzi M, Leonardi M (2021) Theoretical models of consciousness: a scoping review. Brain Sci 11(5):535

Schurger A, Graziano M (2022) Consciousness explained or described? Neurosci Conscious, 2022(1). https://doi.org/10.1093/nc/niac001

Searle JR (1980) Minds, brains, and programs. Behav Brain Sci 3(03):417–424

Signorelli CM, Szczotka J, Prentner R (2021) Explanatory profiles of models of consciousness - towards a systematic classification. Neurosci Conscious, 2021(2). https://doi.org/10.1093/nc/niab021

Stoll E (2022) Modeling electron interference at the neuronal membrane yields a holographic projection of representative information content. *bioRxiv*, 2022.2012.2003.518989. https://doi.org/10.1101/2022.12.03.518989

Tononi, G (2005) Consciousness, information integration, and the brain. In L Steven (Ed.), *Progress in Brain Research* (150, pp. 109–126). Elsevier. https://doi.org/10.1016/S0079-6123(05)50009-8

Tononi G, Boly M, Massimini M, Koch C (2016) Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci 17(7):450–461. https://doi.org/10.1038/nrn.2016.44

Wiese W, Friston, KJ (2021) The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. Philos Mind Sci, *2*

Author contributions

MO and AKH contributed equally to all aspects of the paper.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informated consent was not required as the study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Morten Overgaard.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.