

MIT Press • Open Encyclopedia of Cognitive Science

Consciousness and AI

Patrick Butlin¹

¹Eleos AI

MIT Press

Published on: Feb 05, 2026

DOI: <https://doi.org/10.21428/e2759450.d3359a9d>

License: [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License \(CC-BY-NC-ND 4.0\)](#).

Consciousness is what distinguishes the familiar thoughts, emotions, and sensory experiences of waking life from the processes that go on in our brains without our awareness. Conscious experiences feel some way for the subject and seem to afford the subject a distinctive form of first-person access. Consciousness is often seen as the mark of true intelligence and is also thought to be of great ethical significance, so it is natural to ask whether artificial intelligence (AI) systems could be conscious. Addressing this question is crucial to understanding the kind of change that might be brought about by progress in AI—on one perspective, consciousness would be the difference between systems that are merely sophisticated tools and ones that are our peers. The question also gives new focus to long-standing philosophical debates about the nature and grounds of consciousness.

History

Early pioneers of AI were optimistic about building thinking machines, but the field's founding documents largely set consciousness aside. Alan [Turing \(1950\)](#) urged the field to study behavioral capacities rather than explicitly investigating consciousness, and early manifestos defined the objectives of AI research without mentioning it ([McCarthy et al., 2006](#); [Minsky, 1961](#)) [see [The Turing Test](#)]. However, subsequent projects have aimed to build conscious systems or investigated consciousness as a route to improved performance ([Bengio, 2017](#); [Dennett, 1994](#); [Franklin et al., 2007](#)).

Philosophy of mind in the second half of the 20th century was dominated by functionalism, the view that minds are defined by their functional organization ([Putnam, 1975](#)). This position is friendly to AI minds and consciousness. Putnam's functionalism was motivated partly by a belief in multiple realizability, the claim that mental phenomena could be realized in very different kinds of physical systems. Perhaps the most influential challenge to AI minds was [Searle's \(1980\)](#) “Chinese room” thought experiment. Searle described a person who manipulates symbols according to a book of instructions, and thereby generates apt Chinese utterances, yet does not thereby understand Chinese; he concluded that symbol manipulation is insufficient for intelligence, including consciousness. Functionalists rejected this argument, but it spurred interest in alternative views and in the importance of the mind's connections with the body and environment.

Core concepts

Whether AI consciousness is possible depends on what properties of physical systems are necessary and sufficient for consciousness. The two leading families of views on this question are *computational functionalism* and *biological naturalism*. Computational functionalism claims that implementing computations of the right kind is sufficient for consciousness. This does not straightforwardly imply that AI consciousness is possible because some forms of computation are not possible on the conventional hardware currently used for AI (e.g., analog computation). However, computational functionalism is often motivated by proposed explanations of consciousness in terms of broad functional features of the brain such as its use of gating

mechanisms that select information for further distribution and processing. It appears to be possible to implement such features in conventional systems ([Baars, 1993](#); [Dehaene et al., 2017](#); [Lau, 2022](#)).

In contrast, biological naturalism claims that an organic, living substrate is necessary for consciousness ([Seth, 2025](#)). This implies that conscious artificial systems would have to be very different from today’s AI. There are many possible versions of both computational functionalism and biological naturalism and other views that are different from either; for example, *integrated information theory* claims that consciousness depends on properties that are neither computational nor distinctively biological ([Albantakis et al., 2023](#)).

If AI consciousness is possible, a further question is how to determine whether particular AI systems are conscious. Methods to test for consciousness in other challenging cases, such as nonhuman animals and patients with brain damage, are usually based on either behavior or brain recordings (e.g., functional magnetic resonance imaging or electroencephalogram; [Bayne et al., 2024](#)). However, brain-based methods are likely to be inapplicable in the AI case, and behavioral methods face the *gaming problem* ([Birch, 2024](#)). This is the problem that AI behavior may be generated in very different ways from similar behavior in humans or animals and thus may not provide good evidence of comparable inner processes. This problem is particularly salient when assessing AI systems, like large language models, that have been trained to mimic aspects of human behavior [see [Large Language Models](#)].

Questions, controversies, and new developments

To address the gaming problem, one option is to draw on neuroscientific theories of consciousness, which often describe functional features associated with consciousness in humans. The presence or absence of similar functional features in AI systems may provide evidence about whether they are conscious ([Butlin et al., 2023](#)). For example, global workspace theory claims that consciousness depends on the presence of a shared, limited capacity “workspace” that links multiple specialized subsystems ([Baars, 1993](#); [Mashour et al., 2020](#)). However, it is questionable whether the evidence for these theories, which is largely drawn from studies on humans and primates, supports their extension to the case of AI. In particular, these studies seemingly cannot tell us how similar, in what respects, the features of interest need to be to suffice for consciousness ([Shevlin, 2021](#)).

Relatedly, philosophers have recently argued that, for some AI systems, it could be neither true nor false that they are conscious ([Carruthers, 2019](#); [Papineau, in press](#)). According to reductive materialism, to be conscious just is to instantiate the physical properties that constitute consciousness in humans [see [The Mind–Body Problem](#) and [Physicalism](#)]. AI systems may have some of these properties and not others—for instance, they may have similar functional features differently realized. In this situation, the philosophers claim, it may be indeterminate whether the systems are conscious.

Meanwhile, AI systems are increasingly giving the impression of consciousness to some users ([Colombatto & Fleming, 2024](#); [Shevlin, 2024](#)). This raises empirical questions about the effects of interaction with conscious-seeming systems and ethical questions about their deployment. The social consequences of perceptions of

consciousness in AI are difficult to predict, but they give new urgency to the philosophical and scientific problems mentioned here.

Broader connections

Two further branches of cognitive science can also contribute to debates about AI consciousness. First, ideas about embodied and situated cognition, predictive processing, and the free energy principle may shed light on the possibility and conditions for consciousness in AI [see [The Free Energy Principle](#)]. Second, the issues may be instructively compared with those about consciousness in distant animals such as insects [see [Animal Cognition](#)]. Finally, if AI systems are conscious, it is also important to investigate the nature of their conscious experiences—such as whether they experience human-like emotions.

Further reading

- Andrews, K., & Birch, J. (2023, February 23). What has feelings? *Aeon*. <https://aeon.co/essays/to-understand-ai-sentience-first-understand-it-in-animals>
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. <https://doi.org/10.48550/arXiv.2308.08708>
- Chalmers, D. J. (2023, August 9). Could a large language model be conscious? *Boston Review*. <https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious/>
- Seth, A. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Published online April 21, 2025. <https://doi.org/10.1017/S0140525X25000032>

References

- Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLoS Computational Biology*, 19(10), e1011465. <https://doi.org/10.1371/journal.pcbi.1011465>

↩

- Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

↩

- Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Tests for

↑

- Bengio, Y. (2017). The consciousness prior. *arXiv*. <https://doi.org/10.48550/arXiv.1709.08568>

↑

- Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

↑

- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*. <https://doi.org/10.48550/arXiv.2308.08708>

↑

- Carruthers, P. (2019). *Human and animal minds: The consciousness questions laid to rest*. Oxford University Press.

↑

- Colombatto, C., & Fleming, S. M. (2024). Folk psychological attributions of consciousness to large language models. *Neuroscience of Consciousness*, 2024(1), niae013. <https://doi.org/10.1093/nc/niae013>

↑

- Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486-492. <https://doi.org/10.1126/science.aan8871>

↑

- Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 349(1689), 133-146. <https://doi.org/10.1098/rsta.1994.0118>

↑

- Franklin, S., Ramamurthy, U., D'Mello, S. K., McCauley, L., Negatu, A., Silva L., R., & Datla, V. (2007). LIDA: A computational model of global workspace theory and developmental learning. *Proceedings of the AAI Fall Symposium on Artificial Intelligence and Consciousness*. <https://ccrg.cs.memphis.edu/assets/papers/LIDA%20paper%20Fall%20AI%20Symposium%20Final.pdf>

↑

- Lau, H. (2022). *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press.

↑

- Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious processing and the global neuronal workspace hypothesis. *Neuron*, 105(5), 776-798. <https://doi.org/10.1016/j.neuron.2020.01.026>

↵

- McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence: August 31, 1955. *AI Magazine*, 27(4), 12-14. <https://doi.org/10.1609/aimag.v27i4.190>

↵

- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1), 8-30. <https://doi.org/10.1109/JRPROC.1961.287775>

↵

- Papineau, D. (in press). Consciousness is not the key to moral standing. In G. Lee & A. Pautz (Eds.), *The importance of being conscious*. Oxford University Press.

↵

- Putnam, H. (1975). The nature of mental states. In H. Putnam (Ed.), *Mind, language and reality: Philosophical papers, volume 2* (pp. 429-440). Cambridge University Press.

↵

- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-424. <https://doi.org/10.1017/S0140525X00005756>

↵

- Seth, A. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Published online April 21, 2025. <https://doi.org/10.1017/S0140525X25000032>

↵

- Shevlin, H. (2021). Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, 36(2), 297-314. <https://doi.org/10.1111/mila.12338>

↵

- Shevlin, H. (2024). All too human? Identifying and mitigating ethical risks of social AI. *Law, Ethics and Technology*, 1(2), 0003. <https://doi.org/10.55092/let20240003>

↵

- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460. <https://doi.org/10.1093/mind/LIX.236.433>

↵