OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Is Artificial Intelligence Phenomenally Conscious: A Synthesis of the Literature

Gowri Venkatesha Murthy
gowrimurthy9@gmail.com

ABSTRACT

Artificial Intelligence (AI) is rapidly improving its ability to mimic human cognitive functions. This raises the following question: Could sophisticated intelligence systems have subjective experience and possess phenomenal consciousness? This literature review aims to evaluate proposed theoretical and observational indicators to assess the current consciousness status of AI. The review explicitly differentiates intelligence and consciousness, and the specific lens of focus is categorizing AI as having access consciousness and/or phenomenal consciousness. A multitude of academic publications and theoretical frameworks of different perspectives on AI-consciousness were analyzed. Results from the literature search are consistent with AI resembling access consciousness, with there being a lack of definitive evidence for phenomenal consciousness. This absence of observational evidence is largely due to the difficulty of objectively measuring AI's consciousness from the third-person perspective. The literature review concludes that AI has characteristics of access consciousness, but remains inconclusive regarding phenomenal consciousness because its existence in artificial systems has not yet been empirically proven.

INTRODUCTION

Background on Artificial Intelligence

Artificial Intelligence, or AI, is broadly defined as computer systems with the ability to simulate intelligent human behavior (Merriam-Webster, n.d.). The broadest way to classify AI would be either narrow AI, general AI, or superintelligent AI. Narrow AI, also known as weak AI, is dependent on human feedback and training data. Narrow AI cannot operate beyond its specific domain. This is the AI that exists today. On the other hand, General AI, also referred to as strong AI, would require minimal to no human intervention once trained. General AI would have the ability to learn by itself and apply its skills to domains outside of its defined area. With equivalent or higher cognitive capabilities relative to humans, it would be able to think autonomously (Stryker & Kavlakoglu, 2024). This type of AI does not yet exist, and it poses a multitude of ethical risks and concerns. The entire debate around consciousness is whether this transition from narrow to general AI has or can occur. Superintelligent AI is the most advanced type of AI; it is theoretical just like General AI, but it would surpass human intelligence in each aspect of life, be able to recode itself, and be able to create novel concepts in science, medicine, and other fields (Magnavita, 2026).

AI as a Broad Category

Artificial intelligence serves as an umbrella term for systems designed to simulate human intelligence (Merriam-Webster, n.d.). However, this overarching field is composed of various specialized models with distinct functions that represent only specific subsets rather than the entirety of AI. For example, large language models are specifically meant for processing, translating, and generating text, while computer vision models are more focused on pattern recognition in images and visual data interpretation. Recommender systems, on the other hand, are used for predicting user preferences as opposed to generating creative or autonomous content. The focus of this paper is not to differentiate these separate forms, but to consider modern AI systems as a whole in their ability to have a subjective experience and have access or phenomenal consciousness.

AI Advancement: The Rise of Deep Learning and Modern AI

Historically, artificial intelligence has been prevalent in science fiction, portrayed as a distant-future concern that the Earth would have to deal with. However, the future may not be as far as previously thought. Artificial intelligence has come a long way since the term was coined in 1956 by John McCarthy (Stryker & Kavlakoglu, 2024). In fact, a thorough analysis of 307,701 AI-related publications from 2000 to 2022 showed an increase in the number of publications by 36-fold (Shi et al., 2023). This rapid progress has largely been due to the rise of deep learning.

The term deep learning was introduced by Rina Dechter in 1986 (Hein et al., 2025). However, it was not until the early to mid 2000s that Geoffrey Hinton, commonly referred to as the Godfather of AI, and colleagues popularized the term "deep learning" (Wang & Wyble, 2024). Deep learning is a subsection of machine learning that uses artificial neural networks with several hidden layers, hence the term 'deep.' Artificial neural networks are a system of data-trained interconnected nodes that mimic the biological neurons in the human body. There are three layers that compose a basic feed-forward network: the input layer, one or more hidden layers, and the output layer.

First, the input values, data that is fed into the network to make predictions, pass through the hidden layer(s). Each neuron in a layer takes the weighted sum of the inputs, adds a bias value, and passes the total through a nonlinear activation function. This function determines the neuron's output. In the case that there is a calculation error, a loss function contrasts the network's prediction from the desired outcome and indicates the magnitude of the loss with a number. Layer by layer, the error information is backpropagated through the network until the error is found (Fan et al., 2021). Accordingly, parameters such as the weights and biases are constantly adjusted and improved to produce a more accurate output. This process is conceptually similar to how a human would determine an error since the human would retrace their problem solving until they spot an issue.

Human-AI Similarity

There are numerous similarities between artificial intelligence and humans. A few of those similarities are detailed here. The constant readjustment of weights and other parameters in neural networks is similar to synaptic plasticity in that both rely on modifying connection strengths to improve. In the brain, synaptic

plasticity involves strengthening, weakening, and pruning synapses. Similarly, AI systems algorithmically adjust weights. Crucially, these two processes are not identical but merely similar because even the most complex AI networks that exist today cannot model the brain's complexity in all its manner (Zhang et al., 2023). Additionally, both transformer-based deep neural networks (DNN) and humans use attention mechanisms to place importance on tasks. A study was done on humans where the participants were eye-tracked while reading, and results showed that humans paid more attention and time to the text that was relevant to the task at hand (Zou et al., 2023). The same study was performed on the transformer-based DNNs, which used attention weights to quantify the importance of each word. One distinction that Narhi-Martinez and colleagues make is that the "weights between neurons [described in the previous section] are fixed parameters learned during training that determine how signals are transformed across layers. In contrast, attention weights are dynamically computed values that assign context-dependent importance to different parts of the input" (Narhi-Martinez et al., 2023). These similarities indicate how artificial intelligence is improving when it comes to imitating human cognitive functions.

Figure 1 below, from *Our World in Data*, puts into perspective the rapid progression of AI systems' abilities relative to humans. Critically, these tests were standardized and simulated, and the performance of AI in the real world has not always reflected the same patterns (Roser, 2022).

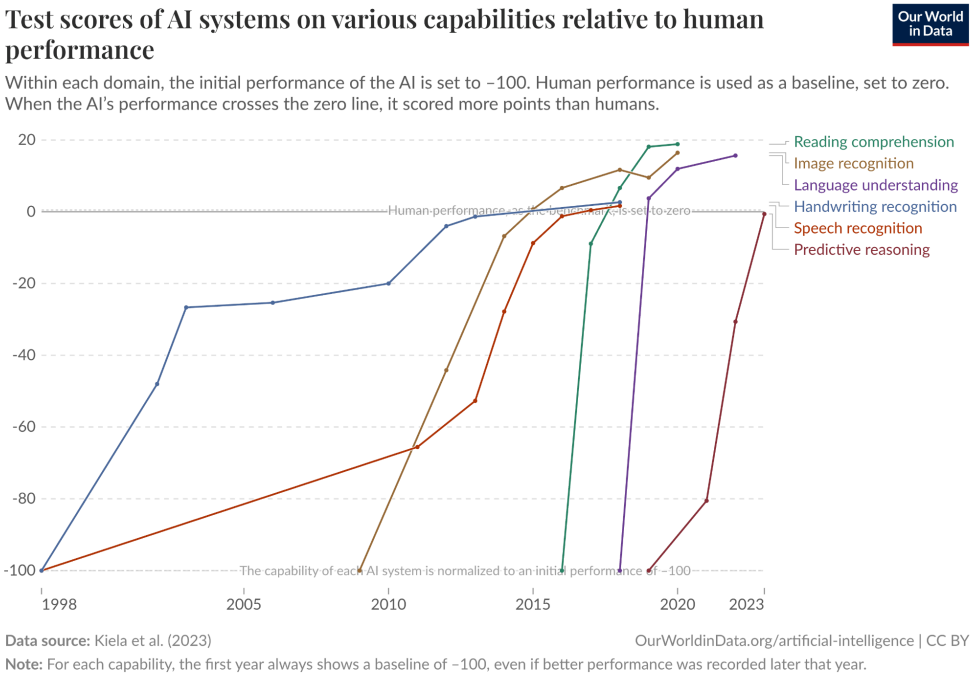


Figure 1. Test scores of AI systems on various capabilities relative to human performance. Image Downloaded from Our World in Data on Thursday, July 24th, 2025 at 12:00pm.

AI's Consciousness Status based on General Public Opinion

A survey was conducted on 300 US residents to determine the general human opinion on whether AI is phenomenally conscious. Results indicated that more often than not, residents believed consciousness was possible for large language models. These opinions were flexible, and there was a positive correlation between an individual's use of an LLM and the level of consciousness they attributed to it (Colombatto & Fleming, 2024).

Background on Consciousness and Definitions/Types

In order to classify AI as conscious or not, a fundamental definition of consciousness must be established. A commonly accepted definition of consciousness is the ability to have "sensations" and make "free decisions" (Zelazo et al., 2007). Since consciousness is such a broad topic, it is difficult to arrive at a 'singular' definition. Accordingly, various definitions of consciousness are available in academic literature. Some of those definitions are covered below.

Access consciousness (A-consciousness) is the ability to acquire information and draw conclusions based on that information. It involves communicating, potentially verbally, about this information (Britannica, n.d.). Many philosophers believe that both AI and humans possess this type of consciousness.

Phenomenal consciousness (P-consciousness) is the subjective experience of something, and therefore it is difficult to detect in artificial intelligence (Britannica, n.d.) An example of P-consciousness is what it is like to feel the warmth of the sun or the taste of an apple. These experiences are unique from being to being. Since humans can only view AI from third person, they do not directly see AI's emotions and experiences. Hence, they cannot easily deduce if AI has subjective experiences - in other words, qualia (Orpwood, 2017).

An important distinction to make is that AI's ability to successfully anthropomorphize does not equate to it being access or phenomenally conscious; this ability is just an indicator of how well AI can mimic a human's conversation style (Guingrich & Graziano, 2024). For example, consider the Turing Test, an examination where an investigator asks both a machine and a human several questions and attempts to distinguish the two. Passing the test demonstrates AI's ability to convincingly mimic human conversation. Artificial intelligence has the capability to pass the Turing Test, but this does not equate to consciousness of any sort (Zelazo et al., 2007).

Another definition of the layers of consciousness is present in Tulving's Tripartite Taxonomy of Consciousness and its relevance to AI. The Taxonomy consists of three levels of consciousness: noetic consciousness, anoetic consciousness, and autonoetic consciousness. Noetic consciousness can be considered a subset of A-consciousness, and it is awareness of semantic knowledge and the ability to analyze information and communicate. Anoetic consciousness is part of phenomenal consciousness, and it is the 'feeling of existence.' Autonoetic consciousness is phenomenal consciousness relevant to memory and the ability to "locate your current experience as part of a narrative of your own life that extends into the past and future" (LeDoux et al., 2023).

This literature review utilizes both the access and phenomenal consciousness definitions as well as the Tulving's Tripartite Taxonomy of Consciousness definition.

Intelligence Versus Consciousness

One key point to clarify is that intelligence and consciousness are not the same. An entity can be intelligent but not conscious, and vice versa. This directly applies to AI because AI can be strongly intelligent (and considered an extension of human intelligence) but not phenomenally conscious. Current theories suggest consciousness requires self-awareness and qualia beyond just intelligence. Therefore, AI can be used as an intelligence tool by a conscious being but intelligence itself does not make AI conscious (Li et al., 2021).

METHODS

This literature review aims to explore the various definitions of consciousness and how those definitions classify or fail to classify AI as a conscious entity. The literature review is a scoping review of literature that addresses if AI has consciousness. The review includes publicly available databases, academic publications, and two non-fiction books focusing on definitions, perspectives on the conscious classification of AI, physiological and behavioral similarities between AI and humans, and tests used to assess AI sentience.

The literature databases used include Google Scholar, PubMed, and ScienceDirect. Search terms included “AI Consciousness”; “AI Consciousness Risks”; “Tiers of Consciousness”; “Neural Networks”; “AI Sentience”; and “AI Intelligence versus Consciousness.” Combinations of these search terms yielded hundreds of thousands of results on each of the three platforms. These results were sorted by relevance. After screening the titles and/or abstracts of the most relevant articles, the text of 100 peer reviewed publications was screened for relevant content. Of these 100, 24 were chosen for inclusion because they discussed principles of consciousness and supported insights on the classification of AI as a conscious entity or not, in addition to 11 that were used to inform the Introduction of this paper. In some cases, literature referenced human behavior when defining a term such as sentience, but the same principles can be applied to the definition when dealing with AI. Publications that were identified from these databases include Cambridge University Press, Springer Nature, Cell Press, and the Institute of Engineering and Technology. The dates of publication range from 1986 to 2026. One book, *The Society of Mind* by Marvin Minsky, was also included to provide an understanding of key concepts which will be described in more detail later.

Additionally, as further discussed in the Limitations section, a limitation of this study stems from the reality that AI is a rapidly evolving field, so literature findings that were accurate at the time of this article’s writing may need to be updated as the field further develops.

RESULTS

Why Biological Naturalism Argues Current Evidence is Positive for A-Consciousness and Inconclusive for AI P-Consciousness

Integrated Information Theory (IIT) measures consciousness by quantifying how irreducibly integrated a system's cause-effect structure is using the value Φ (phi). While originally used to define consciousness in a biological system, IIT's postulates can theoretically be applied to any system (Albantakis et al., 2023). 'Information' refers to the specific cause-effect nature of a network and 'integration' refers to how all the different components of the system interact and how the whole system is more specific than the individual components. Integration is important because it supports IIT's measure of a unified experience (how systems experience one seamless conscious state rather than fragmented parts). A system consisting of completely independent components would have no causal relationship between them and therefore an IIT value of 0. By contrast, a system consisting of irreducibly interrelated components would have a non-zero phi value (Cooke, 2021).

As of September 5, 2026, there is no consensus on a significant non-zero phi value for an AI system, in part due to the high level of difficulty in calculating the phi value itself (Nemirovsky et al., 2023). In addition to this computational challenge, critics express concern that phi values may depend on the calculation method. The lack of a robust phi value would result in inconclusive results.

Also, critics challenge the concept of IIT itself because they are skeptical that a mathematical value can reliably measure qualia (Cerullo, 2015). After all, modern research views the assessment of several theory-based indicators as a broader determination of AI consciousness as opposed to reducing AI's consciousness to a measurable variable (Butlin et al., 2026). This suggests IIT could theoretically misattribute intelligence for consciousness, two qualities explicitly differentiated in the introduction.

Furthermore, Lopez and Montemayor state in "The Integrated Information Theory Needs Attention" that while "IIT might be our current best bet at a scientific explanation of phenomenal consciousness…It fails to acknowledge the crucial roles of attention [mechanisms] in generating phenomenally conscious experience and shaping its contents" (Lopez & Montemayor, 2026). In most cases, attention plays a critical role in consciousness because the brain, or any system with consciousness for that matter, cannot focus on everything at once. Attention is crucial for prioritizing, and information becomes conscious when attention is asserted on it. In the case of AI, attention weights are constantly used to distinguish what is task-relevant from what is not (Kim et al., 2020), so, despite its promise, the concept of testing AI consciousness using Integrated Information Theory is still widely speculative.

Ultimately, IIT identifies consciousness as how irreducibly interconnected a system's components are, but critics contend that it is not sufficient to thoroughly account for internal subjective experience. Given the critiques, Integrated Information Theory supports A-consciousness in AI but is inconclusive regarding P-consciousness.

Why Biological Naturalism Argues Current Evidence does not Support AI P-Consciousness

Another perspective on AI consciousness is present in biological naturalism. The theory states that consciousness is a biological phenomenon that occurs due to dynamic neural signaling in the brain. This process institutes the 'Dynamic Core' hypothesis, which is an explanation of consciousness as integrated neural activity supporting the unified subjective experience. Biologically grounded neuroscience suggests consciousness evolved as an adaptive biological function and does not exist without the brain and certain neuron mechanisms (Edelman et al., 2011). To elaborate, some theorists attribute consciousness to the drive that biological cells have to maintain homeostasis through processes like metabolism, so the AI's lack of an inherent metabolism suggests that it is unlikely to have qualia (Milinkovic & Aru, 2025). In essence, biological naturalism and related biological accounts assert that AI, lacking the neural architecture and biological processes, lacks qualia and cannot be phenomenally conscious (Edelman et al., 2011).

Why Anthropic's J-Space in Claude Argues Current Evidence is Inconclusive for AI P-Consciousness

In short, Anthropic's J-Space can be viewed as the internal, silently operating workspace that Claude uses before it outputs a final answer. In a search to determine if such language models are capable of their own global workspaces, Anthropic has identified that J-Space exhibits a variety of functional properties, such as verbal report, directed modulation, internal reasoning, flexible generalization, and selectivity. All of these characteristics are consistent with access consciousness; however, the experiments conducted did not establish subjective experience in AI (Gurnee et al., 2026). The study was conducted on verbalized traits, meaning the researchers were only able to observe the AI in third person perspective and not identify subjective experience or qualia. Thus, Anthropic's J-Space research can be used to classify AI as A-conscious but is inconclusive towards classifying AI as P-conscious.

A Number of Theories That Hold Promise for Future Classification of AI as P-Conscious

There are several other theories that hold promise that AI could be classified P-conscious in the future. The following will be discussed in detail below: Computationalism (Cleeremans et al., 2025), Global Workspace Theory (Dossa et al., 2024), and Marvin Minsky's Society of Mind proposal (Minsky, 1986). Others, which will not be discussed here, include recurrent processing, computational higher-order Thinking, Predictive Processing, agency and embodiment theories, and many more (Shueb & Che, 2025).

Computationalism is the idea that mental processes such as thinking and problem solving can be scientifically explained in terms of computer functions such as information processing and rule following. The Computational Theory of Mind asserts that in principle, a complex computational system could possess P-consciousness. The theory is testable using simulations that model human performance. For example, the COGITATE Consortium conducted a large-scale study that directly tested computationalist predictions from Global Neuronal Workspace Theory against Integrated Information Theory using multimodal neuroimaging in humans. Results indicated that conscious experience was more strongly linked to posterior cortical activity than to the widespread prefrontal broadcasting predicted by some computationalist models, suggesting that the link between computation and consciousness is nuanced and

requires more investigation. The tests were not fully decisive, though, and do not completely align with either theory (Cleeremans et al., 2025). Fundamentally, computationalism provides a way to scientifically explain the brain's complex thinking (Piccinini & Scarantino, 2011) without relying on the nebulous feelings inside a person's mind. This objective method has been groundbreaking in neuroscience and artificial intelligence research. Computationalists recognize that the relationship between the brain and computers is worth investigating (Mogi, 2024).

Global Workspace Theory (GWT) is a prominent neurocognitive theory that explains how consciousness is a result of brain parts sharing information and collaborating. The Dossa et al. (2024) study successfully mimicked this theory in an AI agent, leading to the agent's passing of all four indicative properties of GWT: broadcasting across multiple specialized modules, a limited capacity workspace with selective attention, global broadcasting across modules, and state-dependent attention (Dossa et al., 2024). This model is arguably one of the most sophisticated implementations of GWT to date, further demonstrating AI's capabilities and potential for A-consciousness. Despite this advancement, the study measured functional properties rather than subjective experience, so the results are inconclusive towards justifying AI as P-conscious.

Marvin Minsky's Society of Mind proposal is another framework where the feeling of self-awareness stems from not one main homunculus but rather a variety of specialized agents (Minsky, 1986). Although highly theoretical, this framework describes problem-solving as breaking tasks into achievable sub-tasks, each handled by a particular agent. This conceptually supports AI being A-conscious because AI comprises a myriad of components that work together to output responses to prompts. The theory reduces the broad scope of defined phenomenal consciousness, however, and should be used as a lens to perceive AI but not diagnose it as phenomenally conscious.

Importantly, advanced AI is now able to perform tasks that were once thought to require consciousness. Examples of such tasks include context-sensitive language processing, where large language models like GPT-3.5 and GPT-4 have performed well in the Winograd Schema Challenge. For context, the Winograd Schema Challenge is a benchmark where AI is tasked with correctly identifying which noun an ambiguous pronoun is referring to in a sentence (Raza et al., 2025). These sentences are carefully crafted so that the system must use reasoning and real-world knowledge to pass, not just grammar. In addition to language, AI is able to perform advanced decision making and problem solving. Although this data does not prove AI is phenomenally conscious, it is in support of AI's improving ability to simulate human reasoning and carry out functions at the level of access consciousness.

DISCUSSION

Final Classification

Reviewing the results of this literature review, the bottom line is that although there are several theoretical propositions arguing AI is or will be P-conscious, much of the empirical evidence remains contentious and obscure.

The task of classifying AI as P-conscious by attempting to observe hidden qualia is challenging due to the third-person perspective, and until conclusive insights into perceiving AI subjectivity are achieved, it remains inappropriate to definitively classify artificial intelligence as P-conscious.

Therefore, this review classifies artificial intelligence as access-conscious but inconclusive regarding phenomenal consciousness. While AI cannot be definitively classified as P-conscious right now, the absence of demonstrative evidence cannot be used to entirely disprove the existence of AI P-consciousness, so the future possibility of reclassification is left open.

Challenges in Obtaining Conclusive Evidence

There are many challenges to obtaining conclusive evidence. First of all, because consciousness in AI is defined in many stages, with some more difficult to perceive than others, it is difficult to definitively arrive at one definition. This increases the difficulty of proving consciousness because the test or theory in question must satisfy numerous qualifications.

Additionally, several theories such as IIT and the Turing Test mistake intelligence for consciousness, two attributes that have been explicitly differentiated. Intelligence, in this case AI's ability to reliably anthropomorphize, does not ensure consciousness, since the two are separate. Tests must thwart the non-uniqueness problem and provide trustworthy evidence that consciousness is present. Furthermore, several tests aim to quantify AI's consciousness, a measure questioned by experts because the ability an objective value has to accurately classify the subjective inner experience that defines the state of being is doubted by experts. Another obstacle is that current proposed tests could mistake AI's ability to anthropomorphize for AI truly being conscious. Just because AI exhibits functional similarity to a conscious entity does not equate to it being conscious; again, the presence of subjective experience must be determined. Lastly, as discussed in more detail below, AI must be definitively shown to possess qualia or subjective experience before it is conclusively determined to be phenomenally conscious.

What is Required for P-Consciousness Classification and Potential for AI to Develop Consciousness in the Future

In order to conclusively classify AI as P-conscious, a test must assess AI beyond the third-person viewpoint, beyond its ability to anthropomorphize, and look further into its internal, subjective feeling. It would have to be stripped of any connection to outside sources, and it would have to be tested on purely its own feelings about its timeline relating to itself and feeling of existence (anoetic consciousness and autonoetic consciousness). As discussed in the introduction, strong/superintelligent AI does not yet exist.

However, if it comes to exist in the future, AI would hypothetically be closer to achieving a conscious state because of its minimal need for human intervention.

Limitations of this Review

This review relies on existing definitions of consciousness. These definitions themselves are debated upon and thus this review cannot go beyond the scopes of the readily defined types of consciousness. Additionally, this review compiles information from a myriad of secondary sources; thus, existing biases or misconceptions in those secondary sources may carry over into this review. Moreover, rapid advancement of AI makes it difficult to find the latest updates in theories relevant to AI consciousness classification, and so this review should not be considered a final verdict; rather, it should be referred to as a provisional synthesis of information and ideas at the time of the writing of the review.

CONCLUSION

The primary focus of this literature review was to determine if artificial intelligence is conscious, and if so, the extent to which it is. After analyzing various theories and proposals regarding how to classify AI as conscious and what consciousness represents by definition, the consensus reached is that AI possesses traits that align with access consciousness. As for phenomenal consciousness, existing theories and technologies render it challenging to view qualia; thus, this review is inconclusive as to whether or not AI is phenomenally conscious.

REFERENCES

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G., 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

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, *30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011

Cerullo, M. A. (2015). The Problem with Phi: A critique of integrated information theory. *PLOS Computational Biology*, *11*(9), e1004286. https://doi.org/10.1371/journal.pcbi.1004286

Cleeremans, A., Mudrik, L., & Seth, A. K. (2025). Consciousness science: Where are we, where are we going, and what if we get there? *Frontiers in Science, 3*, 1546279. https://doi.org/10.3389/fsci.2025.1546279

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

Cooke, J. E. (2021). What is consciousness? Integrated information vs. inference. *Entropy*, *23*(8), 1032. https://doi.org/10.3390/e23081032

Dossa, R. F. J., Arulkumaran, K., Juliani, A., Sasai, S., & Kanai, R. (2024). Design and evaluation of a global workspace agent embodied in a realistic multimodal environment. *Frontiers in Computational Neuroscience*, *18,* 1352685. https://doi.org/10.3389/fncom.2024.1352685

Edelman, G. M., Gally, J. A., & Baars, B. J. (2011). Biology of consciousness. *Frontiers in Psychology*, *2, 4*. https://doi.org/10.3389/fpsyg.2011.00004

Fan, J., Ma, C., & Zhong, Y. (2021). A selective overview of deep learning. *Statistical Science*, *36*(2), 246-290. https://doi.org/10.1214/20-sts783

Guingrich, R. E., & Graziano, M. S. A. (2024). Ascribing consciousness to artificial intelligence: human-AI interaction and its carry-over effects on human-human interaction. *Frontiers in Psychology*, *15*, 1322781. https://doi.org/10.3389/fpsyg.2024.1322781

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026). Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2026/workspace/index.html

Hein, Z. M., Guruparan, D., Okunsai, B., Che Mohd Nassir, C. M. N., Ramli, M. D. C., & Kumar, S. (2025). AI and machine learning in biology: From genes to proteins. *Biology*, *14*(10), 1453. https://doi.org/10.3390/biology14101453

Kim, J., Lee, S., Hwang, E., Ryu, K. S., Jeong, H., Lee, J. W., Hwangbo, Y., Choi, K. S., & Cha, H. S. (2020). Limitations of deep learning attention mechanisms in clinical research: Empirical case study based on the Korean diabetic disease setting. *Journal of Medical Internet Research*, *22*(12), e18418. https://doi.org/10.2196/18418

LeDoux, J. E., Birch, J., Andrews, K., Clayton, N. S., Daw, N. D., Frith, C. D., Lau, H., Megan, Schneider, S., Seth, A. K., Suddendorf, T., & Vandekerckhove, M. (2023). Consciousness beyond the human case. *Current Biology*, *33*(16), R832–R840. https://doi.org/10.1016/j.cub.2023.06.067

Li, D., He, W. and Guo, Y. (2021), Why AI still doesn't have consciousness?. CAAI Trans. Intell. Technol, 6: 175-179. https://doi.org/10.1049/cit2.12035

Lopez, A., & Montemayor, C. (2026). The integrated information theory needs attention. *Erkenntnis, 91*(3), 1419–1443. https://doi.org/10.1007/s10670-025-00949-1

Magnavita N. (2026). Superintelligence: A crucial juncture in the development of healthcare sciences and public health. *International Journal of Environmental Research and Public Health*, *23*(5), 591. https://doi.org/10.3390/ijerph23050591

Merriam-Webster. (n.d.). Artificial intelligence. In *Merriam-Webster.com dictionary*. Retrieved September 7th, 2026, from https://www.merriam-webster.com/dictionary/artificial%20intelligence

Milinkovic, B., & Aru, J. (2025). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience & Biobehavioral Reviews*, 106524. https://www.sciencedirect.com/science/article/pii/S0149763425005251

Minsky, M. (1986). *The society of mind*. Simon and Schuster.

Mogi, K. (2024). Artificial intelligence, human cognition, and conscious supremacy. *Frontiers in Psychology*, *15,* 1364714. https://doi.org/10.3389/fpsyg.2024.1364714

Narhi-Martinez, W., Dube, B., & Golomb, J. D. (2023). Attention as a multi‐level system of weights and balances. *WIREs Cognitive Science*, *14*(1). https://doi.org/10.1002/wcs.1633

Nemirovsky, I. E., Popiel, N. J. M., Rudas, J., Caius, M., Naci, L., Schiff, N. D., Owen, A. M., & Soddu, A. (2023). An implementation of integrated information theory in resting-state fMRI. *Communications Biology*, *6*(1), 1–14. https://doi.org/10.1038/s42003-023-05063-y

Orpwood, R. (2017). Information and the origin of qualia. *Frontiers in Systems Neuroscience*, *11, 22*. https://doi.org/10.3389/fnsys.2017.00022

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics*, *37*(1), 1–38. https://doi.org/10.1007/s10867-010-9195-3

Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). Industrial applications of large language models. *Scientific Reports*, 15(1), 13755. https://doi.org/10.1038/s41598-025-98483-1

Rey, G. (2026, May 14). Philosophy of mind. In *Encyclopedia Britannica*. https://www.britannica.com/topic/philosophy-of-mind

Roser, M. (2022, December 6). The brief history of artificial intelligence: The world has changed fast – what might be next? *Our World in Data*. Retrieved August 17, 2026, from https://archive.ourworldindata.org/20260817-180941/brief-history-of-ai.html

Shi, J., Bendig, D., Vollmar, H. C., & Rasche, P.. (2023). Mapping the bibliometrics landscape of AI in medicine: Methodological study. *JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research*, *25*, e45815–e45815. https://doi.org/10.2196/45815

Shueb, M., & Che, X. (2025). Consciousness in AI systems: A review. *International Journal of Artificial Intelligence and Applications (IJAIA)*, *16*(2), 37-56. https://doi.org/10.5121/ijaia.2025.16205

Stryker, C., & Kavlakoglu, E. (2024, August 9). *What is artificial intelligence (AI)?* IBM. https://www.ibm.com/think/topics/artificial-intelligence

Wang, J. Z., & Wyble, B. (2024). Hopfield and Hinton's neural network revolution and the future of AI. *Patterns*, *5*(11), 101094. https://doi.org/10.1016/j.patter.2024.101094

Zelazo, P. D., Moscovitch, M., & Thompson, E. (2007). *The Cambridge handbook of consciousness*. Cambridge University Press.

Zhang, F., Li, C., Li, Z., Dong, L., & Zhao, J. (2023). Recent progress in three-terminal artificial synapses based on 2D materials: From mechanisms to applications. *Microsystems & Nanoengineering*, *9*(1). https://doi.org/10.1038/s41378-023-00487-2

Zou, J., Zhang, Y., Li, J., Tian, X., & Ding, N. (2023). Human attention during goal-directed reading comprehension relies on task optimization. *eLife*, 12. https://doi.org/10.7554/elife.87197