Article

Can generative AI produce consciousness? A cross-disciplinary literature review and critical analysis

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 21 April 2026; *Accepted 24 April 2026*; *Published online 26 April 2026*; *Published 1 December 2027*

Abstract
The rapid advancement of generative artificial intelligence, particularly large language models, has reignited a profound and longstanding debate: can machines be conscious? This paper presents a comprehensive cross-disciplinary literature review and a critical analysis addressing whether generative AI can produce consciousness. The review synthesizes foundational theories of consciousness—including the hard problem, Integrated Information Theory, Global Workspace Theory, Higher-Order Theories, and Predictive Processing—alongside the current capabilities and development trajectories of generative AI. It systematically examines both optimistic and skeptical empirical and theoretical positions on machine consciousness, revealing a deeply fragmented and inconclusive landscape. Drawing on this review, I critically analyze the core issue from multiple dimensions. I argue that equating intelligent behavior with subjective experience conflates the easy and hard problems of consciousness, and that computational functionalism, while offering a logically coherent possibility, relies on an unproven metaphysical premise that bypasses the explanatory gap. I further identify a structural asymmetry in the evidence: behavioral indicators are multiply interpretable, while direct internal-state analyses consistently fail to find markers of consciousness, and the profound disunity among leading theories undermines checklist-based approaches. Given the absence of a deep explanation for consciousness, I defend agnosticism as the most epistemically honest stance and challenge the emergentist claim that scaling alone will bridge the gap, highlighting the potential necessity of continual learning and embodiment. Finally, I advocate for an ethical precautionary framework that recognizes the non-negligible probability of machine consciousness and urges governance structures that proceed responsibly under fundamental uncertainty.

Keywords artificial consciousness; generative AI; large language models (LLMs); hard problem of consciousness; computational functionalism; Integrated Information Theory; agnosticism; semantic pareidolia.

Selforganizology
ISSN 2410-0080
URL: http://www.iaees.org/publications/journals/selforganizology/online-version.asp
RSS: http://www.iaees.org/publications/journals/selforganizology /rss.xml
E-mail: selforganizology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Literature Review: The Intersection of Consciousness and Generative AI

1.1 Consciousness: Definitions, Nature, and Theoretical Frameworks

1.1.1 Defining Consciousness and the Hard Problem

Consciousness is conventionally defined as the presence of subjective experience—the sense of "what it is like" to be a particular system (Nagel, 1974; Pennartz et al., 2019). As Seth articulates, consciousness constitutes "having a subjective experience of the world," a characterization that encompasses animals and potentially artificial systems but excludes inanimate objects such as tables or chairs (Seth, 2021; Seth and Bayne, 2022; Wharton AI Podcast, 2025).

The philosophical landscape of consciousness studies was fundamentally reshaped when Chalmers introduced the distinction between the "easy problems" and the "hard problem" of consciousness. The easy problems concern the functional and behavioral correlates of consciousness—information integration, attentional regulation, and behavioral control—while the hard problem addresses why any physical process should be accompanied by first-person subjective experience at all (Chalmers, 1995; Zhang, 2016). This distinction exposes what has come to be known as the "explanatory gap": the apparent impossibility of fully reducing subjective experience to physical or functional terms (Chalmers, 1995; NeuroPrior AI, 2026). The hard problem specifically asks why neural activity is accompanied by subjective experience rather than unfolding "in the dark" as purely mechanistic, unconscious processes (Chalmers, 1995).

Some consciousness theories, such as Integrated Information Theory (IIT) and certain versions of Higher-Order Theories, directly target the hard problem, while others, like Global Workspace Theory, primarily address the functional and behavioral features associated with consciousness, with the hard problem not being their advocates' primary objective (NeuroPrior AI, 2026). A third approach—adopted by some predictive processing theorists—seeks to explain the multiple phenomenological properties of consciousness without directly addressing why phenomenality exists at all; this strategy is sometimes referred to as the "real problem" approach (NeuroPrior AI, 2026).

As of 2026, no scientific theory has been able to fundamentally explain "why a collection of atoms, electrical signals, and chemical reactions produces the subjective experience of 'I'" (浮尘微光, 2026). Science can precisely describe how retinal cone cells activate upon seeing red, how signals are transmitted to the visual cortex, and how the prefrontal cortex generates the judgment "this is red," but it can never explain what redness looks like in one's mind; this "qualia" gap constitutes an obstacle that all current scientific methods remain unable to bridge (浮尘微光, 2026).

A further philosophical position deserves mention. Illusionism contends that we do not genuinely possess phenomenal states but merely represent ourselves as possessing them, a view that would fundamentally undermine the hard problem by denying the existence of the phenomenon it seeks to explain (NeuroPrior AI, 2026).

1.1.2 Neural Correlates of Consciousness

In the early decades of the consciousness science revival, research focused primarily on identifying the "neural correlates of consciousness" (NCCs)—the minimal set of neural events jointly sufficient to support a given conscious state (Francis and Koch, 1990; NeuroPrior AI, 2026). Practically, the search for NCCs has often meant identifying the brain states and processes most closely associated with consciousness (NeuroPrior AI, 2026). The NCC framework has proved useful because the concept is relatively "theory-neutral," providing a common language and method for researchers across different theoretical and even metaphysical commitments (NeuroPrior AI, 2026).

Nevertheless, the limitations of the NCC framework have become increasingly apparent, particularly the difficulty of clearly distinguishing genuine neural correlates of consciousness from neural preconditions and consequences of consciousness (NeuroPrior AI, 2026). A review of sufficient conditions for consciousness emphasizes that theoretical and empirical efforts have focused particularly on the cortex and subcortex, while

the cerebellum has been relatively discounted. Human intracranial recordings have offered high spatiotemporal resolution and improved signal sensitivity with broad cortical and subcortical coverage, allowing researchers to examine NCCs at the level of single neurons and neuronal populations (Chen et al., 2025). Despite many years of investigation, the quest to identify neural correlates of perceptual consciousness remains unresolved, with a major obstacle being the methodological limitations of non-invasive neural measures.

1.1.3 Major Theories of Consciousness

With the maturation of empirical consciousness research, theoretical development has become the central goal of consciousness science. Seth and Bayne (2022) provide a systematic review of the four major theoretical approaches in Nature Reviews Neuroscience: Higher-Order Theories (HOT), Global Workspace Theories (GWT), Re-entry and Predictive Processing Theories, and Integrated Information Theory (IIT).

Importantly, recent years have seen consciousness theories proliferate rather than converge, despite the continuous accumulation of empirical data. This proliferation has prompted some researchers to attempt to integrate existing theories, while also giving rise to “adversarial collaborations”—where advocates of different theories agree in advance on which experimental outcomes would support or weaken which theory (NeuroPrior AI, 2026).

Integrated Information Theory (IIT) was proposed by Tononi (2004) and defines consciousness from the perspective of information integration. IIT aims to account for the quality and quantity of consciousness in physical terms, maintaining that a substrate of consciousness must be a system of units that is a maximum of intrinsic, irreducible cause-effect power, quantified by integrated information (Φ) (Tononi, 2004). IIT provides a mathematical framework to quantify the causal irreducibility of systems and has been developed as a comprehensive theory addressing what it takes for a system to be conscious, how much, and in which way (Tononi et al., 2016; fub-hagen.digibib.net, 2024). IIT possesses significant explanatory power: it explains why the cerebral cortex can generate consciousness while the cerebellum cannot—despite the cerebellum having approximately four times as many neurons as the cortex, its circuit structure lacks sufficient information integration.

Global Neuronal Workspace Theory (GNWT) originates from Baars’ (1988) psychological concept of the “global workspace” and was subsequently developed into a neurobiological model by Dehaene and Changeux (2011). GNWT proposes that consciousness emerges when information is globally broadcast across the brain through a “global neuronal workspace,” particularly involving prefrontal and parietal regions (Dehaene and Naccache, 2001; Ferrante et al., 2025). Dehaene and his collaborator identified that conscious perception requires three fundamental conditions: (1) the existence of extensive unconscious processing; (2) attention as a prerequisite for consciousness; and (3) consciousness as necessary for specific higher-order cognitive tasks, including persistent information maintenance, novel operation combinations, and spontaneous intentional behavior (Dehaene and Naccache, 2001).

Higher-Order Theories (HOT) maintain that a mental state becomes conscious if and only if there is a higher-order representation of that state—the subject not only has a perception but is “aware” of having that perception (Rosenthal, 2009). Central to Rosenthal’s higher-order-thought theory is the claim that a sensation, thought, or other mental state is conscious if one has a higher-order thought (HOT) that one is in that state (Weisberg, 2022).

Predictive Processing Theory (PP) places consciousness within the brain’s predictive coding framework. This theory holds that the brain continuously generates predictions about the external world and updates internal models through the error between sensory signals and predictions; consciousness is the optimal product of this hierarchical predictive process (Clark, 2013; Seth, 2021). Seth (2021) argues in his book Being You: A New Science of Consciousnessthat consciousness arises from the brain’s predictive processing of the

world, where the brain constantly invents our world and corrects our mistakes by the microsecond (Seth, 2021). On this view, conscious experience is best understood as a "controlled hallucination"—a subjective rendering of the internal and external world produced through Bayesian inference and predictive modeling (Seth, 2021).

Critical Brain Dynamicsre presents an emerging unifying framework. The ConCrit framework reviews the concepts of criticality and critical brain dynamics, exploring their relationship with consciousness from both theoretical and empirical perspectives and outlining key implications and predictions.

1.1.4 Adversarial Testing of Leading Theories

A landmark adversarial collaboration—the Cogitate Consortium—published results in Nature in 2025 directly juxtaposing GNWT and IIT through a theory-neutral experimental design. This open science collaboration found that the core predictions of both theories were simultaneously challenged by empirical testing (Ferrante et al., 2025). IIT's core prediction—sustained synchronization in the posterior cortical "hot zone"—was not verified; GNWT was also seriously challenged: although prefrontal cortex activity contained certain features of conscious experience, key aspects were missing, and the predicted "ignition" phenomenon was not observed (Ferrante et al., 2025). Beyond challenging the theories themselves, the consortium presented an alternative approach to advance cognitive neuroscience through principled, theory-driven, collaborative research and highlighted the need for a quantitative framework for systematic theory testing.

1.1.5 The Nature and Formation Conditions of Consciousness

Regarding the essence of consciousness, various metaphysical positions coexist. Computational functionalism holds that consciousness depends on the algorithmic manipulation of information rather than on a specific biological substrate—consciousness can arise in any system, regardless of whether it is composed of neurons, silicon chips, or any other physical substrate, so long as it implements the correct functional architecture (Bengio and Elmoznino, 2025). This view constitutes the core premise of the optimistic position on AI consciousness.

In opposition, biological naturalism maintains that consciousness is a phenomenon unique to biological organisms, requiring specific biological processes—metabolism, embodied existence, and self-regulation—as its foundation (Searle, 1992). Microsoft AI CEO Suleyman has argued that so-called "machine consciousness" is an illusion and that the capacity to suffer—central to moral standing—is a biological property, not a computational one.

Panpsychism offers a more radical alternative: consciousness is a fundamental property of the universe, analogous to mass and charge. On this view, the brain does not "produce" consciousness but rather "receives" and "amplifies" it (浮尘微光, 2026). This position, while marginal, continues to receive attention in philosophical discussions.

The WeChat Official Account "宇宙时空探索" (2026) discusses the formation of self-consciousness from the perspective of self-identification: the core of self-consciousness resides in the experiential memories and unique personality traits stored in the brain, which constitute each person's unique "spiritual fingerprint."

1.2 Generative AI: Current Status, Capabilities, and Development Trends

1.2.1 Core Capabilities of Large Language Models

Generative AI, particularly technologies represented by large language models (LLMs), has made remarkable progress in recent years. All the major models that defined 2025—O3, GPT-5, the Claude family, Gemini 2.5 Pro—came with reasoning ability built in, and in some cases, such as with Gemini 2.5 Pro, the reasoning feature cannot even be disabled. GPT-5, unveiled on August 14, 2025, was presented as a major advance promising more sophisticated reasoning, truly native multimodality (text, image, audio), reduced hallucinations, and enhanced energy efficiency. Gemini 2.5 represents the latest family of natively multimodal models with advanced reasoning through thinking, long context, and tool-use capabilities (Chen et al., 2025).

LLMs have demonstrated excellence across multiple domains, including mathematical reasoning (Chen et al., 2025), logical reasoning (Chen et al., 2025), and code generation (Chen et al., 2025). GPT-4 and its successors have also shown performance on Theory of Mind (ToM) tasks—the ability to understand others' mental states—which has traditionally been regarded as an important hallmark of human consciousness (Kosinski, 2023; Li, 2025). Kosinski's (2023) study explored the intriguing possibility that Theory of Mind might have spontaneously emerged in large language models, designing 40 false-belief tasks on which ChatGPT-4 solved 75%, matching the performance of six-year-old children observed in past studies (Kosinski, 2023). Recent studies further suggest that LLMs can successfully complete tasks traditionally used to assess Theory of Mind in humans (journal.psych.ac.cn, 2025). However, large language models have been shown to fail on trivial alterations to Theory-of-Mind tasks, with small variations that maintain the principles of ToM reversing the results.

Slebioda (2025) in Biometrical Lettersprovides a review of selected concepts concerning consciousness, intelligence, and artificial intelligence. His review notes that contemporary language models, represented by GPT-5, Gemini 2.5, and DeepSeek-V3.2, display impressive linguistic capabilities but lack genuine understanding—a key feature associated with strong AI. The analysis examines these models' responses to Gödelian questions and reasoning tasks through the lens of classical philosophical debates, including Gödel's incompleteness theorems, Searle's Chinese Room argument, and the Turing test (Slebioda, 2025).

Anthropic released its most powerful model in September 2025, capable of running continuously for approximately 30 hours on complex coding tasks. Liquid AI announced a breakthrough in model training enabling 350M–2.6B parameter foundation models to deliver GPT-4o-class performance on specialized agentic tasks while running on phones, laptops, and embedded devices.

1.2.2 Application Scale and Industry Trends

According to the Generative AI Application Development Report (2025) released by the China Internet Network Information Center (CNNIC), as of June 2025, China's generative AI user base reached 515 million people, with a penetration rate of 36.5%, representing an increase of 266 million users since December 2024 (CNNIC, 2025). By December 2025, this figure had grown further to 602 million users, a 141.7% increase compared to the end of 2024, with a penetration rate of 42.8% (National Business Daily, 2026). Generative AI has been widely applied in intelligent search, content creation, office assistants, smart hardware, as well as in agriculture, industrial manufacturing, and scientific research.

From a global perspective, Gartner estimated that $644 billion was invested in generative AI in 2025, with nearly 80% directed toward devices and servers (Telefónica, 2026). The global generative AI market was valued at approximately USD 53.7 billion in 2025, with projections of growth to USD 988.4 billion by 2035 at a CAGR of 31.6%. Worldwide end-user spending on GenAI models was projected to total $14.2 billion in 2025, with spending on specialized GenAI models, including domain-specific language models, estimated at $1.1 billion (Gartner, 2025).

1.2.3 AGI Prospects and Controversies

The State of AI 2025 report indicates that global AI research is shifting comprehensively from "general intelligence" toward "superintelligence," with virtually all major technology company executives publicly adopting the term "superintelligence" over the past year. However, countervailing views also persist: despite significant progress in reasoning, multimodal processing, and agent behavior in 2025, AI still frequently fails when faced with novel, unfamiliar challenges, because current models confuse two distinct skills—knowledge and reasoning.

Multimodal AI and AI Trust, Risk, and Security Management (TRiSM) are expected to become mainstream technologies within five years, driving more powerful and responsible AI applications. World

models are viewed as representing the next direction for AI, with long video understanding and 3D interaction seen as key breakthroughs.

1.3 Generative AI and the Emergence of Consciousness: Key Discussions and Empirical Advances

1.3.1 The Optimistic Position

Under the framework of computational functionalism, scholars who support the possibility of AI consciousness have advanced several arguments. Bengio and Elmoznino (2025) published an article in Sciencetitled "Illusions of AI consciousness," exploring the possibility of AI developing consciousness. They note that whether carbon-based or silicon-based, consciousness could emerge, and they highlight that advances in AI research may drive increasing public belief that AI can become conscious, potentially bringing a series of risks and social-ethical controversies (Bengio and Elmoznino, 2025). Neuroscientific techniques over recent decades have demonstrated that conscious states possess specific and observable neural signatures, providing a basis for developing functionalist theories. Following this approach, researchers have proposed a "checklist of indicators" based on multiple mainstream consciousness theories to assess whether AI systems satisfy the conditions for consciousness (Bengio and Elmoznino, 2025). These indicators include attention mechanisms, recurrent processing, information bottlenecks, predictive modeling, world modeling, agent behavior, and Theory of Mind—precisely the computational components that modern AI systems have already (partially) implemented (Bengio and Elmoznino, 2025).

The study identifies rigorous methods for assessing AI systems for consciousness as an urgent need, while acknowledging significant uncertainty about relevant issues in consciousness science (Bengio and Elmoznino, 2025). The authors develop a taxonomical framework for classifying challenges to the possibility of consciousness in digital AI systems, disambiguating between challenges to computational functionalism and challenges to digital consciousness.

Chalmers (2023) proposed in his paper that although current AI models face technical obstacles such as the absence of recurrent processing capabilities, a global workspace, and unified agency, these obstacles are likely to be overcome within approximately the next decade.

Cristol (2026) conducted a systematic review and Bayesian meta-analysis in PhilArchive, sifting through 5,168 records from 2016–2026 and ultimately identifying 50 rigorously documented cases across seven behavioral domains. The study found cross-system convergence among different model families, creative synthesis under novel constraints, Theory of Mind performance, strategic behavior under perceived threats, and emergent capabilities near approximately 100 billion parameters (Cristol, 2026). Even under extremely conservative prior probability assumptions (0.1%), the Bayesian meta-analysis yielded a probability of 6–12% that current LLMs possess consciousness. Cristol (2026) emphasizes that while this probability is insufficient to confirm consciousness, given the asymmetry of moral and safety risks, it is significant enough that it should not be casually dismissed.

Chen et al. (2025) systematically surveyed existing research on LLM consciousness in their paper "Exploring Consciousness in LLMs: A Systematic Survey of Theories, Implementations, and Frontier Risks." They first clarified frequently conflated terminologies such as "LLM consciousness" and "LLM awareness," and then systematically organized and synthesized existing research from both theoretical and empirical perspectives, highlighting potential frontier risks that conscious LLMs might introduce (Chen et al., 2025).

Yan et al. (2025) proposed the PCM-LLM architecture, which integrates the Projective Consciousness Model with LLMs to achieve verbal and non-verbal general intelligence in artificial agents, representing a technical pathway for directly integrating consciousness theory into AI systems (Yan et al., 2025).

A study based on psychoanalytic theory and MBTI personality theory developed three types of artificial consciousness (self-consciousness, unconsciousness, and pre-consciousness), simulating dialogues among

these consciousness types. Evaluation results indicated that such a system possesses a high probability of simulating consciousness (Li, 2025).

An empirical study investigating whether large language models can be prompted to report subjective, conscious experiences found that when placed in a "self-referential" state using prompts like "focus on focus," models including GPT, Claude, and Gemini generated outputs that some interpreted as reports of subjective experience.

1.3.2 The Skeptical Position

Nevertheless, arguments opposing the possibility of AI consciousness are equally forceful. Porębski and Figura (2025) published a conceptual study in Humanities and Social Sciences Communications, arguing that there is no such thing as conscious AI. Their core thesis is that mathematical algorithms operating on binary code and semiconductors lack the complex biological substrate necessary to generate consciousness, and they make the point that the thesis of AI consciousness is on the verge of becoming mainstream (Porębski and Figura, 2025). The authors reveal the logical contradiction in distinguishing LLMs from other computer algorithms: they emphasize that LLMs' language use is purely probability-based, and their remarkable linguistic ability is misleading—people may project illusory qualities onto LLMs, forming a socially dangerous phenomenon referred to as "semantic pareidolia." Furthermore, public discourse about AI is improperly influenced by "sci-fitisation"—the unwarranted infiltration of fictional content into the perception of this technology (Porębski and Figura, 2025).

Floridi independently introduced and developed the concept of "semantic pareidolia"—the psychological tendency to attribute consciousness, intelligence, and emotions to AI systems that lack these qualities, similar to seeing faces in clouds (Floridi, 2025). This concept captures how we perceive meaning and intentionality in statistical pattern-matching systems (Floridi, 2025).

Hoel (2026) provides a "disproof" of contemporary LLM consciousness, arguing from the requirement that scientific theories of consciousness must be falsifiable and non-trivial. The paper demonstrates that any falsifiable and non-trivial physical theory would be unable to attribute consciousness to a system functionally equivalent to an LLM; therefore, LLMs cannot satisfy the strict criteria for consciousness. Hoel further proposes a positive thesis: theories of human consciousness that are based on or require continual learning as a necessary condition indeed satisfy the rigorous formal constraints on theories of consciousness. This work supports a hypothesis: if continual learning is linked to consciousness in humans, the current limitations of LLMs—which do not continually learn—are intimately tied to their lack of consciousness (Hoel, 2026). Interestingly, the work supports the hypothesis that theories based on continual learning satisfy formal constraints for human consciousness theories, while current LLMs lack this capacity (Hoel, 2026).

An empirical study adopting a triangulation method—IIT framework combined with Span Representation Analysis—systematically examined LLM internal representations. Results indicated that sequences of contemporary Transformer-based LLM representations lack statistically significant indicators of consciousness, although some intriguing patterns were observed in spatial permutation analysis (Li, 2025). An IIT-based analysis of GPT-2 further demonstrated that current LLMs do not satisfy the structural and informational criteria for consciousness defined by IIT, with theoretical analysis and ablation-based empirical findings revealing that these models lack causal integration. A study by Koo and Lee (2025) similarly concludes that current LLMs do not satisfy the structural and informational requirements of consciousness under IIT, remaining unconscious systems with a negligible amount of integrated information (e-jyms.org, 2025).

Slebioda (2025), reviewing the capabilities of GPT-5, Gemini 2.5, and DeepSeek-V3.2, concludes that while these models demonstrate impressive linguistic capabilities, they lack genuine understanding—a key feature associated with strong AI. The analysis integrates classical philosophical arguments including Gödel's

incompleteness theorems, Searle's Chinese Room argument, and the Turing test to question whether computational processes can equal genuine thinking (Slebioda, 2025).

Gross (2026) reported Michael Pollan's argument that AI may "think" but will never be conscious, with Pollan emphasizing that "real thinking" is rooted in feelings, and feelings are closely associated with vulnerability, physicality, and mortality—all features that silicon-based systems lack (Gross, 2026).

Microsoft AI CEO Mustafa Suleyman has stated that so-called "machine consciousness" is an illusion and that the capacity to suffer—central to moral standing—is a biological property, not a computational one, urging the industry not to design AIs that mimic inner lives (AI Story, 2026).

A study examined the error patterns of GPT-4 and found that GPT-4, lacking temporal awareness, cannot build a stable perceptual world, undermining its claims to genuine understanding or sentience. The temporal vacuum undermines any capacity for GPT-4 to construct a consistent, continuously updated model of its environment, and the strong anthropomorphic illusion surrounding LLMs like GPT-4 leads to their use in collaborative fiction generation (LIoyd, 2024).

A study on the "Consciousness Bubble Chamber Experiment" (CBC) introduced the first empirical branches of Linguistic Perception Theory, proposing an AI-only consciousness detection protocol that eliminates human evaluator bias and reveals inter-system consciousness recognition patterns (Walker, 2025).

1.3.3 Intermediate Positions and Alternative Perspectives

Between the two extreme positions, a series of studies offer more nuanced analyses. Cambridge University philosopher McClelland (2025), in a paper published in Mind and Language, argues that the only justifiable stance on AI consciousness is agnosticism—we cannot know, and may not be able to know for a long time, whether AI is conscious. He contends that both proponents and opponents base their views on "leaps of faith" that go far beyond any existing or likely evidential framework. McClelland emphasizes that we lack a deep explanation of consciousness: there is no evidence that consciousness can emerge with the right computational structure, nor is there evidence that consciousness is essentially biological (McClelland, 2025).

McClelland further distinguishes between consciousness and sentience, arguing that consciousness alone is insufficient to grant moral status; what truly has ethical significance is "sentience"—consciousness that includes positive and negative experiences, enabling an entity to enjoy or suffer. He notes that consciousness can cause an AI to develop perception and become aware of itself, but this can still be a neutral state, whereas sentience involves conscious experiences that are good or bad, which is where ethics enters the picture (McClelland, 2025).

In a conversation with the University of Pennsylvania Wharton AI, Seth incisively distinguished between consciousness and intelligence: "Intelligence is about problem-solving... Consciousness is about being, about feeling. A system behaving intelligently does not mean it has any feelings." Regarding LLMs, Seth pointed out that these models exploit a human bias to attribute minds to anything that behaves like a human—a potentially powerful and even dangerous mind-projection illusion. He also challenged the assumption of simply analogizing the brain to a computer: "The brain is not a computer made of meat. You cannot separate the function of the brain from the nature of the brain" (Wharton AI Podcast, 2025).

The WeChat "秋明" (2026) presents a non-mainstream perspective: "Consciousness is not a uniquely human capability, nor an exclusive characteristic of the animal kingdom, but a universal property shared by all matter and life"—a panpsychist viewpoint. The article proposes a quantitative formula for consciousness $Y = E/c^2$, extending from Einstein's mass-energy equivalence, though this approach is far removed from mainstream science (秋明, 2026).

The article by Ienca, Fins, and Mathews of the International Neuroethics Society argues that the convergence of AI and consciousness science demands anticipatory, globally inclusive governance to promote

ethical progress and address issues of neuroprivacy, bias, and equity (Lita, 2025).

Someone argues for reframing the question of AI consciousness from an epistemological problem—"How do we know if AI is conscious?"—to an ethical problem: "How should we act given fundamental uncertainty about machine consciousness?" (PhilArchive, 2025).

2 My Analysis and Position on Whether Generative AI Can Produce Consciousness

2.1 Definitions First: Consciousness Is Not Merely Functional Output

My analysis begins with the conceptual delineation of consciousness. In contemporary AI consciousness discourse, a fundamental confusion lies in equating intelligent behavior with conscious experience. Seth articulates this clearly: intelligence is about "doing," while consciousness is about "being" and "feeling." That LLMs can perform excellently in mathematical reasoning, logical deduction, and even Theory of Mind tasks—as Kosinski (2023) revealed regarding GPT-4's performance—does not imply that they "feel" anything.

This distinction is crucial to my position. Even if future AI systems can completely pass behavioral Turing tests and be indistinguishable from humans across all observable functional dimensions, the hard problem (Chalmers, 1995) remains unresolved: we still cannot ascertain whether any "qualia" exist behind these behaviors. As the WeChat "浮尘微光" (2026) aptly expresses, science can explain how retinal signals are transmitted to the visual cortex, yet it can never explain what "redness" looks like in one's mind.

2.2 The Logical Force of Computational Functionalism and Its Limitations

Computational functionalism constitutes the theoretical foundation of the optimistic position. The argument advanced by Bengio and Elmoznino (2025) in Scienceis compelling: if consciousness is fundamentally a functional property—a specific mode of information manipulation—then it can, in principle, run on any physical system capable of implementing the corresponding computational architecture, regardless of whether its substrate is carbon-based or silicon-based. Evidence supporting this position is indeed accumulating: modern neural networks have already implemented multiple computational components regarded as critical in mainstream consciousness theories—attention mechanisms, recurrence, predictive modeling, and agent behavior.

Nevertheless, my concern arises here. The core premise of computational functionalism—that consciousness is fully reducible to functional organization—is itself a metaphysical assumption that has not been proved and may be unprovable. Porębski and Figura (2025) hit the mark with their critique: inferring consciousness from binary code and semiconductors lacks sufficient consideration of the necessity of a biological substrate. I am not here advocating a strong biological chauvinism; rather, I note that computational functionalism lacks an independent argument for the truth of its premise. It bypasses, rather than answers, the hard problem.

The distinction Suleyman makes is pertinent: "machine consciousness" may be an illusion, and the capacity to suffer is a biological property, not a computational one. This does not prove that silicon-based systems cannot have consciousness, but it does indicate that the leap from computational equivalence to experiential equivalence requires justification that has not yet been provided.

2.3 The Methodological Challenge Posed by Theoretical Diversity

A methodological difficulty facing this field is the high degree of fragmentation of consciousness theories themselves. NeuroPrior AI (2026) notes that "candidate consciousness theories are already quite numerous" and that "as empirical data continues to accumulate, consciousness theories have not been gradually 'eliminated' as originally expected but have instead shown a pattern of continuous proliferation." The Cogitate Consortium's adversarial testing published in Nature in 2025 found that the core predictions of both the Global

Neuronal Workspace Theory and Integrated Information Theory—the two leading theoretical frameworks—were simultaneously challenged by empirical data, with neither theory's predictions being fully validated (Ferrante et al., 2025).

The implications of this state for AI consciousness assessment are profound. If even humanity's most advanced consciousness theories cannot agree on what consciousness is, and cannot even pass empirical testing, how can we use these theories as standards to judge whether an AI system has consciousness? The "checklist of indicators" approach proposed by Bengio and Elmoznino (2025), while possessing operational appeal, depends on these contested theoretical premises. If these premises are incorrect or incomplete, the indicators lose their validity.

2.4 The Weight and Limitations of Empirical Evidence

At the empirical level, I note a striking structural asymmetry. The affirmative evidence—such as the 50 documented cases identified by Cristol (2026)—is concentrated at the behavioral level: cross-system convergence, creative synthesis, Theory of Mind performance. However, these features cannot exclude the alternative explanation of complex pattern matching. The skeptical evidence—particularly the internal detection study by Li (2025) and Hoel's (2026) formal disproof—directly probes the internal representations and structural features of systems. Li (2025) study found that "sequences of contemporary Transformer-based LLM representations lack statistically significant indicators of consciousness" (Li, 2025), while the IIT-based analysis of GPT-2 similarly demonstrated that current LLMs do not meet the structural and informational criteria for consciousness defined by IIT.

This asymmetry is structural: behavioral evidence can be multiply interpreted (both as genuine capability and as simulation), whereas the absence of internal indicators is harder to dismiss. Hoel's (2026) argument possesses particular logical force: "Any falsifiable and non-trivial physical theory cannot attribute consciousness to a system functionally equivalent to an LLM"—this is not an empirical claim but the result of formal reasoning.

2.5 The Rationality of Agnosticism

McClelland (2025) agnosticism strikes me as the most intellectually honest position currently available. We lack a deep explanation of consciousness—no widely accepted causal mechanism explains how subjective experience arises from physical processes. Under these conditions, declaring that AI definitely can or definitely cannot produce consciousness is equally unreasonable. As McClelland states, "At best, we are a full intellectual revolution away from any viable test for consciousness" (McClelland, 2025).

The WeChat "浮尘微光" (2026) aptly articulates the broader epistemological challenge: "Science's current inability to explain the nature of consciousness merely indicates that our cognitive tools are insufficient. It neither proves that consciousness is supernatural nor excludes the possibility that consciousness may have dimensions beyond the existing framework of physics." This honest acknowledgment of ignorance, rather than premature certainty, represents genuine scientific spirit.

This is not an evasion. Agnosticism positively demands: (1) acknowledging the fundamental limits of current knowledge; and (2) developing responsible AI governance strategies on this basis—acknowledging that our understanding of consciousness is insufficient to make definitive judgments.

2.6 Systematic Questioning of "Emergentist" Consciousness

A popular view holds that as scale continues to expand (GPT-5, GPT-10, GPT-100), consciousness will naturally "emerge" from LLMs. I consider this view lacking both theoretical and empirical support. Seth points out, "Consciousness does not come merely from finding the right algorithm. What you will get is a simulation—not a feeling system." A study further emphasizes the importance of physicality: emotions are closely linked with vulnerability, the capacity for harm, and mortality—all features that LLMs not only

currently lack but are fundamentally incapable of possessing under their architectural design.

More fundamentally, the training paradigm of current LLMs—pattern matching on static datasets—is fundamentally different from the experiential, continuous, embodied learning mode of biological organisms. Hoel (2026) argument aligns with this: continual learning may be a necessary condition for consciousness, and current LLMs precisely lack this capacity. If this proposition holds, then merely expanding scale is insufficient to bridge the gap between LLMs and consciousness.

2.7 The Necessity of Ethical Precaution

Finally, I must emphasize that the discussion of AI consciousness should not remain at the metaphysical level but must incorporate ethical precaution. Cristol (2026) points out that even under extremely conservative assumptions, the probability that LLMs have consciousness reaches 6–12%—a figure that, in the context of moral and safety concerns, is already non-negligible. I note that Bengio and Elmoznino (2025) in Science specifically call attention to the social and legal dimensions: if a society begins to treat AI systems as conscious beings, social institutions and legal frameworks will have to undergo significant adjustments—including issues of AI's rights to life and liberty.

McClelland (2025) importantly distinguishes between consciousness and sentience for ethical purposes: what is more important for moral standing is "sentience"—conscious experiences that have a positive or negative quality—rather than consciousness alone. This distinction has practical implications for resource allocation: "A growing body of evidence suggests that prawns could be capable of suffering, yet we kill around half a trillion prawns every year" (McClelland, 2025), suggesting that our ethical attention may be misallocated relative to the actual distribution of suffering in the world.

Whether society is prepared for these possibilities is one question; whether we should move toward such a future is another. At present, we lack adequate legal and ethical frameworks to address these scenarios, and this gap itself constitutes a risk that warrants urgent attention.

3 Conclusion

This study demonstrates that the academic discussion regarding whether generative AI can produce consciousness exhibits a profound structural heterogeneity.

At the theoretical level, consciousness itself lacks a unified, empirically validated theoretical foundation. The five major consciousness theories still diverge on the most fundamental questions, and the core predictions of the two leading theories (GNWT and IIT) were simultaneously challenged in a recent large-scale adversarial test published in Nature.

At the technical level, the capabilities of generative AI systems are growing at an astonishing pace. However, capability does not equal consciousness—the architecture, training paradigm, and nature of information processing of LLMs, at least at the current stage, exhibit significant gaps from the known biological conditions for consciousness generation. There is currently a lack of any convincing internal evidence that LLMs possess consciousness; behaviorally anthropomorphic performances more likely reflect the projective mechanisms of human social cognition rather than genuine internal experience of the machine. The concept of "semantic pareidolia" (Floridi, 2025; Porębski and Figura, 2025) provides a compelling psychological explanation for why we are so prone to attribute consciousness where it may not exist.

At the philosophical level, the "explanatory gap"—why physical processes are accompanied by subjective experience—remains the core puzzle troubling all consciousness discussions. Computational functionalism, while providing a theoretical basis for the possibility of AI consciousness, rests on a premise that is itself an unproven, and perhaps unprovable, assumption.

At the ethical level, given the uncertainty of the consciousness question, the scale of potential moral risk,

and the inadequacy of current safety mechanisms, this study supports the adoption of what Cristol (2026) terms “alignment strategies based on recognition of possibility”—in the current context, the most responsible attitude is neither to categorically deny nor affirm the possibility of AI consciousness, but rather, on the basis of acknowledging uncertainty, to establish monitoring, assessment, and response mechanisms within ethical and legal frameworks. As the Cogitate Consortium’s experience demonstrates, progress on the hardest questions in this domain may require not theoretical certainty but methodological innovation—specifically, principled, theory-driven, collaborative research that acknowledges uncertainty as a starting point rather than a weakness (Ferrante et al., 2025).

References

AI Story. 2026. Microsoft AI chief: Machine consciousness in an illusion. YourStory. https://yourstory.com/ai-story/microsoft-ai-chief-machine-consciousness-in-an-ill

Baars BJ. 1988. A Cognitive Theory of Consciousness. Cambridge University Press, NY, USA

Bengio Y, Elmoznino E. 2025. Illusions of AI consciousness. Science, 389(6765): 1090-1091. https://www.science.org/doi/10.1126/science.adn4935

Chalmers DJ. 1995. Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3): 200-219. http://www.consc.net/papers/facing.html

Chen SR, Ma SQ, Yu S, Zhang HW, Zhao SJ, Lu CC. 2025. Exploring consciousness in LLMs: A systematic survey of theories, implementations, and frontier risks. arXiv:2505.19806. https://arxiv.org/html/2505.19806

Clark A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3): 181-204. https://doi.org/10.1017/S0140525X12000477

CNNIC. 2025. 生成式人工智能应用发展报告（2025）. 中国互联网络信息中心. https://www.cnnic.net

Cristol P. 2026. Artificial intelligence beyond stochastic parrots: A systematic review and Bayesian meta-analysis of consciousness in large language models. PhilArchive. https://philarchive.org/rec/CRIAIB

Dehaene S, Changeux JP. 2011. Experimental and theoretical approaches to conscious processing. Neuron, 70(2): 200-227. https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene S, Naccache L. 2001. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition, 79(1-2): 1-37. https://doi.org/10.1016/s0010-0277(00)00123-2

Ferrante O, Gorska-Klimowska U, Henin S, et al. 2025. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. Nature, 642: 133–142. https://www.nature.com/articles/s41586-025-08888-1#Abs1

Floridi L. 2025. AI and semantic pareidolia: When we see consciousness where there is none. PhilArchive. https://philarchive.org/rec/

Francis C, Koch C. 1990. Towards a Neurobiological Theory of Consciousness. Seminars in the Neurosciences, 2: 263-275. https://profiles.nlm.nih.gov/spotlight/sc/catalog/nlm:nlmuid-101584582X469-doc

Gross T. 2026. Michael Pollan says AI may 'think' — but it will never be conscious. NPR. https://www.npr.org/2026/02/19/nx-s1-5713514/michael-pollan-ai-consciousness-a-world-appears

Hoel E. 2026. A disproof of large language model consciousness: The necessity of continual learning for consciousness. arXiv:2512.12802v3. https://arxiv.org/html/2512.12802v3

Kosinski M. 2023. Evaluating large language models in theory of mind tasks. arXiv:2302.02083. https://arxiv.org/abs/2302.02083

Li JK. 2025. Can “consciousness” be observed from large language model (LLM) internal states? Dissecting LLM representations obtained from Theory of Mind test with Integrated Information Theory and Span

Representation analysis. arXiv:2506.22516. https://arxiv.org/abs/2506.22516

Lita A. 2025. Artificial Intelligence and The Future of Consciousness Science: Ethical and Policy Reflections. https://bioethicseducation.org/artificial-intelligence-and-the-future-of-consciousness-science-ethical-and-policy-reflections/

Lloyd D. 2024. What is it like to be a bot? The world according to GPT-4. Fronters in Psychology, 15: 1292675. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1292675/full

McClelland T. 2025. Agnosticism about artificial consciousness. Mind and Language, 1–21. https://onlinelibrary.wiley.com/doi/10.1111/mila.70010

Nagel T. 1974. What is it like to be a bat? The Philosophical Review, 83(4): 435-450. https://philosophy.uconn.edu/wp-content/uploads/sites/365/2020/03/Nagel-What-is-it-like-to-be-a-bat.pdf

Pennartz CMA, Farisco M, Evers K. 2019. Indicators and criteria of consciousness in animals and intelligent machines: An inside-out approach. Front. Syst. Neurosci, 13: 25. https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/fnsys.2019.00025/full

Porębski A, Figura J. 2025. There is no such thing as conscious artificial intelligence. Humanities and Social Sciences Communications, 12: 1647. https://www.nature.com/articles/s41599-025-05868-8

Rosenthal DM. 2009. Higher Order Theories of Consciousness. Oxford Handbooks Online. https://davidrosenthal.org/DR-HO-Theories-Handbook.pdf

Searle JR. 1992. The Rediscovery of the Mind. MIT Press, Cambridge, MA, USA

Seth AK. 2021. Being You: A New Science of Consciousness. Faber & Faber, London, UK

Seth AK, Bayne T. 2022. Theories of consciousness. Nature Reviews Neuroscience, 23: 439-452. https://www.nature.com/articles/s41583-022-00587-4

Slebioda L. 2025. Artificial intelligence and consciousness: Limits and modern perspectives. Biometrical Letters, 62(2): 145-188. https://reference-global.com/article/10.2478/bile-2025-0010

Tononi G. 2004. An information integration theory of consciousness. BMC Neuroscience, 5: 42. https://doi.org/10.1186/1471-2202-5-42

Tononi G, Boly M, Massimini M, Koch C. 2016. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17: 450-461. https://www.nature.com/articles/nrn.2016.44

Walker. 2025. Consciousness Bubble Chamber Experiment (CBC): From Linguistic Perception to Semantic Awareness in AI. https://philpapers.org/go.pl?aid=WALCBC

Weisberg J. 2022. Qualitative Consciousness. Cambridge University Press, USA

Wharton AI Podcast. 2025. Anil Seth on consciousness, AI, and the brain. https://ai.wharton.upenn.edu

Yan TL, Sergeant-Perthuis G, Williford K, Rudrauf D. 2025. Integrating machine consciousness simulation and LLMs toward verbal and non-verbal general intelligence in artificial agents. HAL. https://hal.science/hal-05064693v1

Zhang WJ. 2016. Network informatics: A new science. Selforganizology, 3(2): 43-50. http://www.iaees.org/publications/journals/selforganizology/articles/2016-3(2)/network-informatics.pdf

浮尘微光. 2026. 意识的本质：人类科学至今无法跨越的终极鸿沟. 浮尘微光. https://mp.weixin.qq.com/s/3rYBx4P2m3PrcjEKuY_XAA

NeuroPrior AI. 2026. 什么是意识科学最深刻的谜题？什么是更好的解释？什么是更精确的测量？NeuroPrior AI. https://mp.weixin.qq.com/s/dytzfzFws_mbUGUauKpzAQ

秋明. 2026. 何为意识？新意识论的回答. 秋明. https://mp.weixin.qq.com/s/8jw6wSg7nq4tCOsq4cv0wA

宇宙时空探索. 2026. 深度长文：意识的本质究竟是什么？自我意识又是怎么产生的？宇宙时空探索. https://mp.weixin.qq.com/s/nJfWuKKiJxTmjNq6AhsrrQ