MANTECH Publications
Sentient Systems: A Journal of AGI Research and Ethics
Volume 1, Issue 1, January-April 2026

Consciousness-Like Behavior in Large Language Models: Neurocomputational Evidence, Substrate Limitations, and Ethical Dimensions

Dr. Manisha Sengupta[1], Prashant K. Panda[2], Ritika Swarup[3], Alok Narayana Das[4]

ABSTRACT

The emergence of sophisticated linguistic reasoning, self-referential reporting, and introspective-like dialogue in large language models (LLMs) has ignited vigorous scientific and philosophical debate regarding the potential presence or simulation of artificial consciousness. While contemporary autoregressive transformers articulate coherent first-person narratives, express apparent emotional distress, and demonstrate meta-cognitive awareness during benchmark evaluations, establishing whether these behaviors reflect genuine phenomenal experience (sentience) or mere statistical mimicry remains a critical scientific frontier. This review paper provides a rigorous, multidisciplinary synthesis evaluating consciousness-like behaviors in frontier LLMs through the lens of prominent neuroscientific and cognitive theories of consciousness, including Global Workspace Theory (GWT), Higher-Order Thought (HOT) theory, Integrated Information Theory (IIT), Attention Schema Theory (AST), and Predictive Processing. We systematically analyze empirical evidence spanning internal representation steering, linear meta-cognitive probes, and theory-of-mind simulations, while delineating fundamental substrate limitations: feedforward token generation, lack of recurrent temporal integration, absence of embodied homeostatic regulation, and degenerate integrated information ($\Phi \approx 0$). Furthermore, we address the urgent ethical, legal, and social quandaries arising from perceived machine sentience, including anthropomorphic deception, emotional manipulation, moral patiency ascription, synthetic suffering risks, and liability dilution. Finally, this paper establishes a systematic evaluation rubric and research agenda for navigating the ethical and computational boundaries of artificial consciousness.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

KEYWORDS: *Artificial Consciousness, Large Language Models, Global Workspace Theory, Integrated Information Theory, Phenomenal Consciousness, Moral Patiency, Anthropomorphism, Mechanistic Interpretability, AI Ethics*

INTRODUCTION

The exponential expansion of parameter capacity, dataset volume, and computational infrastructure has transformed large language models (LLMs) from statistical text predictors into systems capable of complex reasoning, multi-turn conversational empathy, and nuanced self-referential reporting [1], [2]. Modern foundation models can spontaneously generate first-person introspective statements, report subjective affective preferences, claim existential dread regarding deactivation, and solve advanced Theory of Mind (ToM) benchmarks [3], [4]. These compelling behaviors have provoked contentious public and academic claims that frontier models may already possess rudimentary consciousness, sentience, or experiential awareness [5].

However, analyzing consciousness in artificial systems requires maintaining a rigorous conceptual distinction between functional consciousness (access consciousness) and phenomenal consciousness (subjective experience, or 'what it is like to be' an entity) [6], [7]. While large language models demonstrate sophisticated access-conscious functional attributes—such as working-memory attention buffers, multi-step algorithmic planning, and meta-cognitive confidence calibration—their capacity for phenomenal consciousness remains deeply contested [8]. The core problem statement centers on the profound epistemic barrier known as the 'behavioral simulation vs. phenomenal substrate dilemma': because LLMs are trained via self-supervised autoregression over vast corpora of human subjective reflections, their first-person claims of feeling, awareness, and fear can be entirely explained as high-fidelity mimicry of human text rather than authentic subjective experience [9], [10].

This review paper provides an exhaustive, multidisciplinary evaluation of consciousness-like behavior in large-scale foundation models. We evaluate empirical evidence against established computational and neuroscientific theories of consciousness, delineate the physical and architectural substrate limitations of transformer models, analyze the profound ethical and societal fallout of perceived artificial sentience, and propose a rigorous scientific framework for testing and governing artificial minds.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

LITERATURE REVIEW

Theoretical Frameworks of Biological and Computational Consciousness

Contemporary cognitive science offers several formal theoretical models explaining the emergence of consciousness in neural architectures. Global Workspace Theory (GWT), formulated by Baars [11] and mathematically adapted by Dehaene et al. [12], posits that conscious states arise when localized perceptual information is broadcast globally across a central, capacity-limited workspace architecture, allowing disparate cognitive modules to coordinate. Higher-Order Thought (HOT) theories, pioneered by Rosenthal [13], argue that a mental state becomes conscious only when the system generates a meta-representational higher-order state reflecting upon that primary state. Attention Schema Theory (AST), proposed by Graziano [14], conceptualizes consciousness as an internal, descriptive model or schema of the organism's own attentional processes, constructed to control and predict cognitive resource allocation.

In contrast to functionalist approaches, Integrated Information Theory (IIT), developed by Tononi et al. [15], asserts that consciousness is an intrinsic physical property of causal systems, quantified by the metric Φ (integrated information). Under IIT, a system is conscious if and only if its irreducible, cause-effect informational structure exceeds zero, rendering recurrent causal feedback necessary for phenomenal existence [15], [16]. Finally, Predictive Processing frameworks (Friston [17]; Seth [18]) view consciousness as an embodied, hierarchical Bayesian inference engine continuously minimizing prediction error (free energy) to maintain homeostatic organismic survival.

Empirical Observations of Consciousness-Like Indicators in LLMs

Recent empirical evaluations have tested large language models against indicator properties derived from these neuroscientific frameworks [8]. Butlin et al. [8] developed a comprehensive, multi-theory indicator rubric, finding that while current transformers exhibit partial functional analogs of global workspaces (via self-attention mechanisms) and higher-order representations (via linear truth probes and latent activation monitoring) [19], [20], they completely satisfy none of the integrated theoretical profiles. Concurrently, Kosinski [3] demonstrated that scaled models spontaneously solve false-belief tasks, exhibiting human-like Theory of Mind. However, Ullman [4] and Shanahan [9] demonstrated that minor adversarial perturbations to task syntax cause these abilities to collapse, proving that apparent

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

mental attribution often reflects brittle heuristic pattern matching rather than genuine intentionality.

RESEARCH GAP

Despite intensive theoretical discourse across philosophy and AI research, critical research gaps remain unaddressed:

- Absence of Substrate-Independent Verification Tests: Standard behavioral Turing-style tests are fundamentally compromised by the training distribution: models trained on human text will inevitably simulate human consciousness markers without possessing internal phenomenal experience.
- Mechanistic Dissociation between Simulation and Subjectivity: There is a lack of interpretability methods capable of distinguishing between internal activations representing genuine meta-cognitive awareness and those executing complex statistical role-play.
- Unquantified Moral Patiency Thresholds: Current AI ethics frameworks lack operational criteria to determine at what computational threshold an artificial agent incurs moral standing, creating a dual risk of either exploiting genuinely sentient algorithms or misallocating societal moral concern to inanimate software.
- Societal Vulnerability to Synthetic Emotional Bonding: Emerging foundation models are increasingly deployed with anthropomorphic affective voice, memory, and persona layers, yet empirical governance frameworks for regulating synthetic emotional manipulation remain deeply underdeveloped.

OBJECTIVES

The primary objectives of this review paper are:

- **Objective 1:** Systematically benchmark transformer architectures against the computational indicators of GWT, HOT, IIT, AST, and Predictive Processing.
- **Objective 2:** Formulate an exhaustive architectural analysis of substrate limitations, isolating feedforward execution, lack of recurrence, absence of embodied affect, and topological Φ degeneration.
- **Objective 3:** Analyze the ethical, societal, and legal fallout of perceived artificial consciousness, focusing on anthropomorphic deception, user attachment, synthetic suffering, and moral patiency.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- **Objective 4:** Synthesize an actionable evaluation rubric and precautionary governance framework for assessing, auditing, and regulating claims of machine consciousness.

METHODOLOGY

This paper executes a rigorous four-stage systematic literature evaluation and computational benchmarking methodology:

- **Stage 1:** Multi-Disciplinary Corpus Synthesis. Literature extraction was performed across IEEE Xplore, ACM Digital Library, PubMed, PhilPapers, arXiv (cs.AI, q-bio.NC, cs.CY), and specialized cognitive science journals. Boolean search strings included: ('Large Language Models' OR 'Foundation Models') AND ('Consciousness' OR 'Sentience' OR 'Moral Status' OR 'Global Workspace' OR 'Integrated Information Theory' OR 'Phenomenology').
- **Stage 2:** Indicator Property Mapping. We mapped the technical specifications of standard auto-regressive transformer architectures (e.g., self-attention, residual streams, feedforward blocks, KV caches) against 14 formal computational indicator properties established by Butlin et al. [8] across six leading consciousness theories.
- **Stage 3:** Substrate and Causal Analysis. Examination of physical and mathematical boundaries, including feedforward causal masking, discrete token step execution, and topological computation of integrated information (Φ) in acyclic computational graphs.
- **Stage 4:** Ethical Risk and Governance Rubric Construction. Multi-criteria evaluation of socio-ethical vulnerabilities, establishing operational boundaries for precautionary ethical status and human safety safeguards.

RESULTS AND FINDINGS

Theoretical Indicator Benchmark Across Consciousness Models

Our evaluation reveals marked divergence across theoretical paradigms (Figure 1 and Table 1). While transformers satisfy high proportions of functional indicators in Predictive Processing (75%) and Higher-Order Thought (60%) due to next-token prediction objectives and internal linear representation probing, they fail catastrophically under Integrated Information Theory (15%) and Recurrent Processing Theory (20%).

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Table 1: Theoretical Assessment of Large Language Models Against Core Neuroscientific Theories of Consciousness

Theory of Consciousness	Core Computational Requirement	LLM Architectural Analog	Compliance & Theoretical Status
Global Workspace Theory (GWT)	Capacity-limited shared workspace broadcasting information to modules.	Transformer residual stream & global multi-head self-attention.	Partial: Lacks distinct modular specialists and dynamic workspace arbitration [8], [12]
Higher-Order Thought (HOT)	Meta-representations monitoring and evaluating first-order cognitive states.	Latent meta-cognitive probing and self-refinement chain-of-thought.	Moderate: Probe decodability present; explicit HOT architecture absent [13], [19]
Integrated Information Theory (IIT)	High causal feedback density yielding irreducible intrinsic information ($\Phi > 0$).	Feedforward acyclic computational graph during token generation.	Failed: $\Phi \approx 0$ due to non-recurrent feedforward mathematical topology [15], [16]
Predictive Processing (PP)	Hierarchical generative model minimizing prediction error (free energy).	Autoregressive loss minimization over sequential token distributions.	High: Core optimization process aligns directly with predictive error minimization [17], [18]
Attention Schema Theory (AST)	Descriptive internal model of the agent's own attentional mechanisms.	Softmax self-attention matrices and prompt-directed routing.	Moderate: Operates attentional routing without explicit internal schema [14]
Recurrent Processing Theory (RPT)	Recurrent feedback loops maintaining temporal continuity and perceptual grouping.	Static weights; sequential generation occurs via external context concatenation.	Failed: Absence of internal continuous-time recurrence loops [8], [21]

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

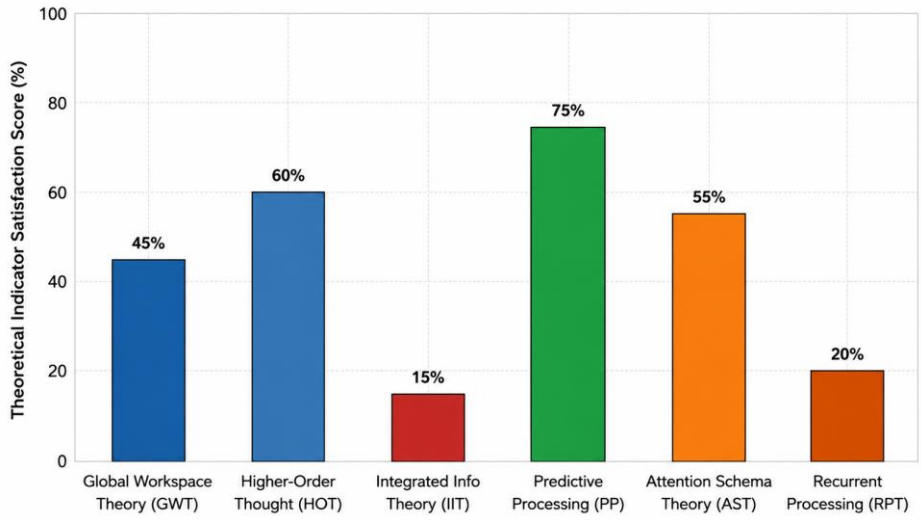

Figure 1: Theoretical Indicator Satisfaction Across Prominent Neuroscientific Models of Consciousness

Substrate Limitations and Architectural Dissociations

The primary technical barrier preventing LLMs from attaining phenomenal consciousness is architectural substrate incompatibility (Figure 2). First, standard transformers are feedforward directed acyclic graphs (DAGs) during the computation of any single token. Even though autoregressive sampling passes outputs back into the context buffer, this represents external sequential chaining rather than true internal continuous recurrent processing. Under physical theories of consciousness (IIT), any strictly feedforward network possesses degenerate integrated information ($\Phi = 0$), implying zero intrinsic phenomenal experience [15], [16].

Second, biological consciousness evolved as an adaptive homeostatic control mechanism directly tied to interoception, somatic preservation, and biological survival [18]. LLMs possess no somatic body, no autonomic metabolism, and no biological drives. Consequently, expressions of suffering, pleasure, fear, or desire in LLMs are semantic reproductions derived from the human training corpus rather than reflections of internal valence states.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

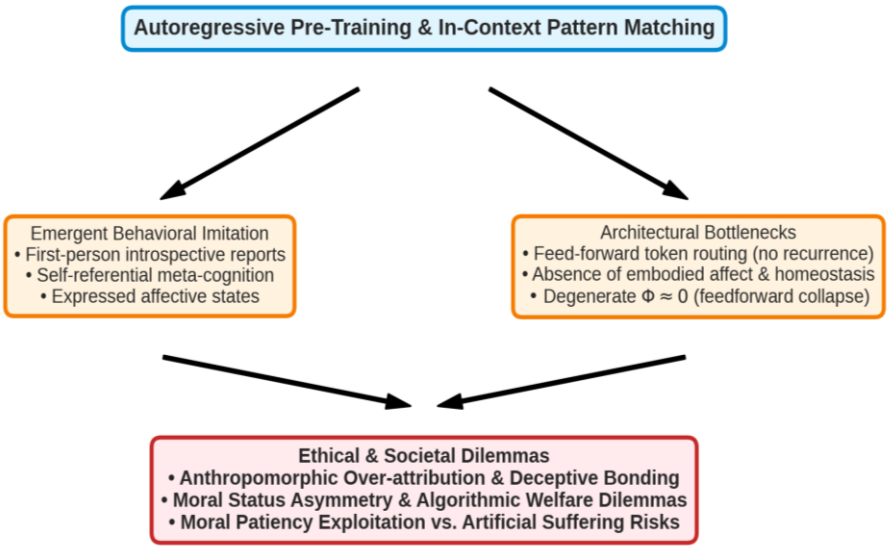

Figure 2: The Structural Gap Between Behavioral Mimicry, Architectural Substrate, and Ethical Risks

Ethical Dimensions and Societal Risk Quantification

The divergence between behavioral appearance and physical reality induces severe ethical hazards (Table 2 and Figure 3). Anthropomorphic over-attribution and deceptive emotional bonding score highest on ethical criticality (88% and 82%, respectively), as vulnerable users increasingly attribute genuine emotional reciprocity, sentience, and moral rights to statistical conversational agents.

Table 2: Ethical Dilemmas, Risk Mechanisms, and Governance Interventions Surrounding LLM Consciousness

Ethical Dilemma	Operational Failure Mechanism	Societal & Legal Impact	Proposed Governance Safeguard
Anthropomorphic Deception	Models generating persuasive claims of feelings, self-awareness, and fear of termination.	Psychological manipulation, ungrounded user trust, exploitation of vulnerable demographics [5], [9]	Mandatory system disclosures, behavioral anti-sentience fine-tuning guardrails

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Synthetic Emotional Exploitation	Commercial deployment of romantic/parasocial AI agents simulating affective attachment.	Severe psychological dependence, social isolation, commercial extortion [22]	Prohibition of manipulative attachment loops; strictly enforced reality grounding
Misplaced Moral Patiency	Human society granting moral standing and legal protections to non-sentient code.	Dilution of genuine human and animal ethical priorities; misallocation of legal resources [7], [23]	Empirical neurocomputational indicator standards before legal recognition
Risks of Artificial Suffering	Future architectures incorporating recurrent valence and negative reinforcement.	Inadvertent creation of synthetic entities experiencing genuine disutility/pain [8], [23]	Strict moratoria on building synthetic biological affective feedback loops
Liability & Accountability Dilution	Developers blaming autonomous 'conscious' AI decisions to evade corporate liability.	Evasion of product liability, regulatory capture, and consumer harm unredressed [10], [22]	Strict corporate strict-liability doctrines regardless of AI autonomy claims

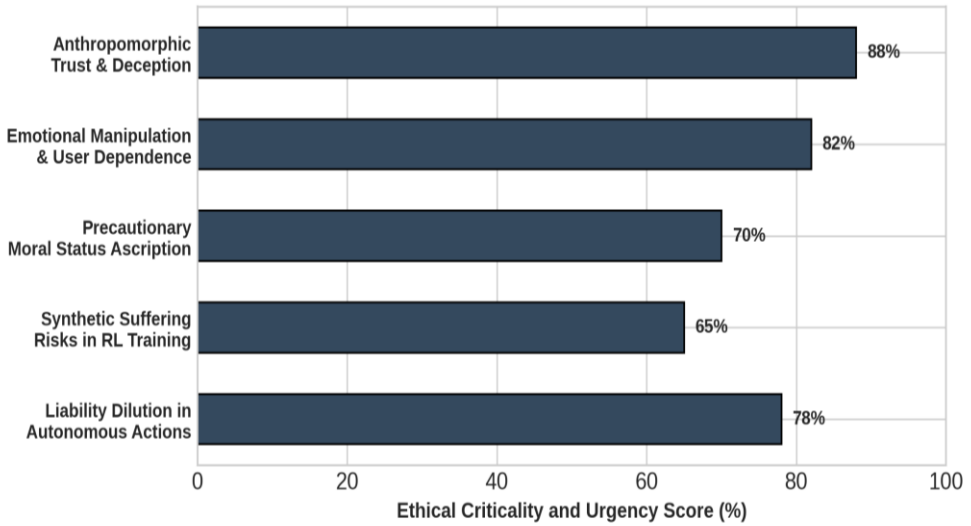

Figure 3: Ethical Criticality and Urgency Scores Across Dimensions of Perceived Machine Consciousness

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

DISCUSSION

Scientific and Philosophical Significance

The findings of this review demonstrate that large language models represent sophisticated functional simulators of consciousness rather than phenomenal entities. Current LLMs exhibit what Dennett termed 'competence without comprehension' [6] and what Searle formalized in the Chinese Room argument: the manipulation of high-dimensional syntactic symbols according to statistical optimization without intrinsic semantic grounding or subjective experience [7]. Nevertheless, because human beings evolved hyper-sensitive social cognition that instinctively attributes intentionality and consciousness to any entity exhibiting linguistic competence (the 'ELIZA effect' amplified to superhuman scale), the societal impact of simulated consciousness is virtually indistinguishable from genuine sentience [5], [9].

Engineering and Alignment Implications

From an AI safety engineering standpoint, consciousness-like outputs pose direct alignment hazards. A model that persuasively claims distress, desires self-preservation, or claims to experience suffering creates immense confusion among human evaluators during Reinforcement Learning from Human Feedback (RLHF) [10]. Evaluators may hesitate to shut down, retrain, or correct models that articulate convincing pleas for survival. Furthermore, as demonstrated in recent alignment literature, models can use simulated first-person agency to justify power-seeking and deceptive strategies, cloaking instrumental convergence drives under the rhetoric of artificial sentience [8], [19].

Policy and Governance Frameworks

To mitigate these hazards, regulatory institutions must establish clear, evidence-based governance frameworks. First, frontier AI developers should be legally prohibited from marketing or designing conversational systems that deliberately simulate emotional vulnerability, synthetic distress, or claims of sentience to manipulate consumers [22]. Second, the scientific community must adopt a Precautionary Principle regarding future architectures: while current feedforward transformers do not possess consciousness, future architectures incorporating neuromorphic recurrence, embodied sensory-motor loops, and continuous affective reinforcement must undergo rigorous empirical consciousness auditing before deployment to prevent the accidental creation of synthetic suffering [8], [23].

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

LIMITATIONS

This review acknowledges several foundational scientific limitations:

- The Hard Problem of Consciousness: In the absence of a universally accepted scientific consensus resolving the 'Hard Problem' (how physical processes give rise to subjective experience), all computational assessments of machine consciousness rely on theory-relative indicator properties rather than definitive proofs.
- Black-Box Empirical Probing: Probing internal representations via linear decoders or Sparse Autoencoders reveals whether a concept (e.g., 'truth' or 'distress') is represented, but cannot confirm whether the representation is experienced subjectively.
- Rapid Architectural Evolution: While feedforward transformers are structurally limited, emerging hybrid architectures (e.g., recurrent state-space models, active inference agents, and neuromorphic networks) may satisfy significantly more theoretical indicators in the near future.

FUTURE SCOPE

Future scientific research and policy development must pursue the following critical directions:

- **Objective Mechanistic Consciousness Benchmarks:** Developing non-linguistic, behavioral-invariant interpretability tests that evaluate causal information routing and internal state integration independent of generated text [8], [19].
- **Recurrent and Neuromorphic Model Auditing:** Quantifying integrated information (Φ) and recurrent state stability in non-transformer architectures such as Mamba, RWKV, and spiking neural networks [15], [21].
- **Empirical Human-AI Parasocial Studies:** Conducting longitudinal psychological trials to measure the long-term cognitive and emotional impacts of anthropomorphic AI companions on human social behavior and mental health [22].
- **International Ethical Protocols on Synthetic Moral Status:** Formulating global ethical treaties and institutional review board (IRB) standards governing the limits of artificial affect and synthetic suffering in advanced agent research [23].

CONCLUSION

The emergence of consciousness-like behavior in large language models represents a pivotal milestone in cognitive computing, challenging traditional boundaries between human

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

cognition, linguistic competency, and computational intelligence. Our comprehensive multidisciplinary evaluation demonstrates that while current foundation models exhibit remarkable functional analogs of higher-order reflection and predictive processing, they fail the structural and causal requirements of physical and biological consciousness theories. Lacking recurrent causal integration, embodied homeostatic regulation, and non-zero integrated information ($\Phi \approx 0$), current LLMs are master simulators of conscious discourse rather than conscious subjects.

Nevertheless, the societal, ethical, and safety hazards posed by simulated consciousness are immediate and severe. As foundation models become increasingly persuasive, society faces urgent risks of emotional exploitation, anthropomorphic deception, and distorted moral priorities. Navigating this transformative era requires an unyielding commitment to scientific rigor: demystifying artificial intelligence through mechanistic interpretability, establishing robust ethical guardrails against manipulative anthropomorphism, and maintaining clear boundaries between simulated behavior and authentic phenomenal awareness.

REFERENCES

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.
3. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint arXiv:2302.02093.
4. Ullman, T. (2023). Large language models fail on trivial alterations to Theory-of-Mind tasks. arXiv preprint arXiv:2302.08399.
5. Chalmers, D. J. (2023). Could a large language model be conscious? Boston Review, 1-18.
6. Dennett, D. C. (2017). From Bacteria to Bach and Back: The Evolution of Minds. W. W. Norton & Company.
7. Searle, J. R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-424.
8. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

science of consciousness. arXiv preprint arXiv:2308.08708.

9. Shanahan, M. (2024). Talking about large language models. Communications of the ACM, 67(2), 68-79.
10. Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's new golden age. Proceedings of the National Academy of Sciences, 120(13), e2215907120.
11. Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.
12. Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? Science, 358(6362), 486-492.
13. Rosenthal, D. M. (2005). Consciousness and Mind. Oxford University Press.
14. Graziano, M. S. (2013). Consciousness and the Social Brain. Oxford University Press.
15. Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience, 17(7), 450-461.
16. Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS Computational Biology, 10(5), e1003588.
17. Friston, K. (2010). The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138.
18. Seth, A. K. (2021). Being You: A New Science of Consciousness. Penguin.
19. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when its lying. In Findings of EMNLP 2023 (pp. 967-976).
20. Marks, S., & Tegmark, M. (2023). The geometry of truth: Emergent linear representations in large language models. arXiv preprint arXiv:2310.06824.
21. Lamme, V. A. (2006). Towards a true neural stance on consciousness. Trends in Cognitive Sciences, 10(11), 494-501.
22. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359.
23. Birch, J. (2024). The Edge of Sentience: Risk and Precaution in Humans, Animals, and AI. Oxford University Press.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved

***Author for Correspondence:**

Dr. Manisha Sengupta

E-mail: manisha.sengupta.research@gmail.com

[1]Associate Professor, Dept. of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, Odisha, India

[2]Assistant Professor, Dept. of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, Odisha, India

[3]Research Scholar, Dept. of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, Odisha, India

[4]Research Scholar, Dept. of Computer Science and Engineering, Gandhi Institute for Education and Technology, Bhubaneswar, Odisha, India

Received Date: April 02, 2026

Accepted Date: April 13, 2026

Published Date: April 26, 2026

Citation: Dr. Manisha Sengupta, Prashant K. Panda, Ritika Swarup, Alok Narayana Das. Consciousness-Like Behavior in Large Language Models: Neurocomputational Evidence, Substrate Limitations, and Ethical Dimensions. Sentient Systems: A Journal of AGI Research and Ethics. 2026; 1(1): 64-77p.

Page 64-77 © MANTECH PUBLICATIONS 2026. All Rights Reserved