MANTECH Publications

Sentient Systems: A Journal of AGI Research and Ethics
Volume 1, Issue 2, May-August 2026

Computational Models of Artificial Sentience: A Comparative Analysis of Consciousness-Inspired Cognitive Architectures for AGI

Suresh Narayanan[1], Swapnil Khot[2], Rohan Mittal[3]

ABSTRACT

As Artificial General Intelligence (AGI) paradigms shift from localized task-specific machine learning toward unified, autonomous, and self-adaptive cognitive frameworks, the integration of artificial sentience and phenomenal consciousness mechanisms has emerged as a fundamental research frontier. This paper presents a comprehensive comparative review of computational models of artificial sentience, rigorously analyzing four prominent consciousness-inspired cognitive architectures: Global Workspace Theory (GWT), Integrated Information Theory (IIT), Predictive Processing and Active Inference Frameworks (PP/AIF), and Higher-Order Thought (HOT) computational implementations. We establish a multi-tiered comparative taxonomy evaluating these architectures across core performance and qualitative metrics, including cognitive integration density, phenomenal vs. access consciousness mapping, computational scalability, real-time latency, and self-synthetic metacognitive adaptability. Furthermore, we analyze empirical case studies across autonomous robotic navigation, multi-agent cooperative alignment, and adaptive synthetic decision-making under uncertainty. Our findings reveal that while GWT and PP/AIF offer scalable, real-time architectures for access consciousness and sensory-motor prediction, IIT and HOT models provide deeper formalizations for phenomenological metrics and meta-represented self-awareness, albeit with significant computational overheads (e.g., NP-hard integrated information calculations). To bridge these gaps, we propose a novel Hybrid Multi-Layered Sentience Architecture (HMLSA) that synthesizes global broadcasting, predictive active inference, and higher-order meta-inspection. Finally, we discuss engineering significance, practical applications, ethical alignment imperatives, and critical future research trajectories necessary for realizing

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

resilient, ethically aligned AGI systems.

KEYWORDS: *Artificial Sentience, Artificial General Intelligence (AGI), Cognitive Architectures, Global Workspace Theory, Integrated Information Theory, Active Inference, Metacognition.*

INTRODUCTION

The pursuit of Artificial General Intelligence (AGI) represents one of the most ambitious engineering and scientific endeavors of modern computation. Contemporary artificial intelligence systems, largely dominated by deep neural networks, transformer-based large language models (LLMs), and reinforcement learning agents, have demonstrated phenomenal performance in domain-specific tasks such as natural language processing, strategy games, and protein folding [1, 2]. However, these contemporary architectures remain fundamentally brittle, lacking true semantic context, self-directed agency, adaptive resilience in out-of-distribution environments, and genuine subjective or phenomenal awareness [3]. To transition from narrow AI to robust AGI, computational neuroscience and cognitive science have increasingly turned toward modeling artificial sentience—the computational realization of phenomenal awareness, subjective experience, and unified cognitive integration [4].

Artificial sentience, within a rigorous computational framework, does not merely imply the mimicry of human emotions; rather, it refers to an agent's capability to construct a unified dynamic self-model, maintain real-time awareness of its own internal states and sensory inputs, broadcast critical information globally across specialized modules, and minimize informational entropy through proactive environmental modeling [5, 6]. The integration of consciousness-inspired cognitive architectures offers a principled pathway to overcome the key limitations of static machine learning models, specifically in areas concerning metacognition, self-reflection, continuous unprompted learning, and ethical self-regulation.

Problem Statement

Despite rapid theoretical advancements in computational neuroscience, artificial sentience research remains fragmented across disparate theoretical paradigms. Current cognitive architectures either excel at localized sensory-motor integration while lacking high-level self-reflection, or offer mathematical formalizations of subjective experience that are

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

computationally intractable for complex systems [7]. Furthermore, there is a distinct lack of quantitative, unified benchmark methodologies to evaluate and compare consciousness-inspired architectures for real-time AGI deployment. Without a comparative computational analysis, system architects face severe bottlenecks in selecting, scaling, and hybridizing sentience models for real-world application domains such as autonomous robotics, ethical decision-making systems, and resilient multi-agent networks.

Objectives

This research paper aims to fulfill the following theoretical and practical objectives:

- Formulate a rigorous comparative taxonomy evaluating major consciousness-inspired cognitive architectures: Global Workspace Theory (GWT), Integrated Information Theory (IIT), Predictive Processing / Active Inference Frameworks (PP/AIF), and Higher-Order Thought (HOT) models.
- Analyze the mathematical and structural underpinnings of each architecture, evaluating their trade-offs regarding computational complexity, cognitive integration, and real-time execution capabilities.
- Examine domain-specific case studies to illustrate how these sentience models behave in real-time robotic navigation, multi-agent cooperation, and metacognitive error recovery.
- Propose a novel Hybrid Multi-Layered Sentience Architecture (HMLSA) that resolves existing trade-offs by unifying predictive active inference, global workspace broadcasting, and higher-order meta-monitoring.

LITERATURE REVIEW & TECHNICAL BACKGROUND

The foundation of artificial sentience rests upon both cognitive neuroscience and computer science. Ned Block's classic distinction between Access Consciousness (A-consciousness)—information available for reasoning, reporting, and action control—and Phenomenal Consciousness (P-consciousness)—the subjective, qualitative experience ('what it is like' to be in a state)—remains a cornerstone for computational modeling [8].

Global Workspace Theory (GWT) and LIDA

Originally proposed by Bernard Baars and mathematically formalized in cognitive architectures like LIDA (Learning Intelligent Distribution Agent) by Stan Franklin et al., GWT posits that the brain consists of distributed, specialized parallel processors [9, 10]. A

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

central dynamic memory structure—the 'Global Workspace'—acts as an informational bottleneck. Specialized modules compete via attentional mechanisms to gain access to the workspace. Once an item enters the workspace, it is globally broadcast to all sub-systems, enabling unified action selection, memory consolidation, and conscious access [11]. In computational implementations, GWT is modeled through blackboard architectures, dynamic message-passing neural networks, and self-attention transformers [12].

Integrated Information Theory (IIT)

Formulated by Giulio Tononi, Integrated Information Theory (IIT) approaches sentience from a top-down phenomenological perspective [13]. IIT asserts that consciousness is an intrinsic property of a physical or computational system, quantified by the metric Phi (Φ), which measures the amount of integrated information generated by a system above and beyond its individual disconnected parts [14]. The mathematical formulation involves calculating cause-effect structures across state transitions, evaluating minimum information partitions (MIP). While IIT provides a rigorous mathematical measure of phenomenal consciousness, calculating Φ for complex networks exhibits exponential computational complexity, scaling as O(2^N) or worse, making real-time computation for large N highly challenging [15].

Predictive Processing and Active Inference (PP/AIF)

Pioneered by Karl Friston and Andy Clark, Predictive Processing and Active Inference frame sentience through the Free Energy Principle (FEP) [16, 17]. Under AIF, a sentient agent maintains a generative model of its internal states and external environment. Sentience emerges as the agent continuously minimizes variational free energy—a statistical bound on surprise or prediction error—through perception (updating internal beliefs) and action (altering the environment) [18]. AIF provides an elegant, computationally efficient framework for self-organizing homeostatic control, body-ownership modeling, and goal-directed action selection [19].

Higher-Order Thought (HOT) Theories

Higher-Order Thought theories, articulated by David Rosenthal and computationally modeled by Lau and Rosenthal, argue that a mental state becomes conscious only when the system represents itself as being in that state via a second-order meta-representation [20, 21]. In AGI

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

engineering, HOT architectures implement dual-level networks: a primary lower-level sensory-motor execution system monitored by a secondary metacognitive neural network that dynamically evaluates error rates, confidence thresholds, and operational integrity [22].

RESEARCH GAP

Despite extensive literature on individual consciousness theories, several fundamental research gaps hinder their effective synthesis into functional AGI frameworks:

- Disconnect Between Access and Phenomenal Models: Existing literature heavily bifurcates into operational access models (GWT, AIF) and abstract theoretical phenomenal metrics (IIT). A unified architecture integrating high-density information integration with scalable global broadcasting and active inference is largely missing [23].
- Computational Intractability of Integrated Information: Standard IIT implementations fail to operate in real-time embedded systems due to combinatorial explosion during state-space partition analysis [24]. Efficient, high-accuracy approximations of Φ usable in dynamic operational loops remain underdeveloped.
- Lack of Quantitative Comparative Benchmarks: Existing reviews evaluate sentience models purely qualitatively. Quantitative frameworks comparing throughput, memory overhead, latency under sensor failure, and metacognitive error recovery under standardized environments are practically non-existent.

RESEARCH METHODOLOGY

To address these gaps, this paper employs a multi-phase comparative and analytical methodology combining comparative theoretical taxonomy, simulation-based empirical modeling, and algorithmic architectural synthesis.

Analytical Framework and Metric Formulation

We evaluate four primary architectures—GWT (LIDA implementation), IIT (Phi-approximate graph network), PP/AIF (Variational Active Inference loop), and HOT (Dual-Level Metacognitive Network)—across six key operational metrics:

- Cognitive Integration Density (CID): The ratio of globally synchronized information flows to local modular processing.
- Computational Complexity & Latency: Algorithmic time complexity O(f(N)) and execution latency (ms) per cognitive cycle.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Metacognitive Adaptability Index (MAI): The system's ability to identify internal failures, update generative models, and re-allocate dynamic compute.
- Real-Time Execution Scalability: Feasibility of embedding the model into high-frequency robotic control loops (>50 Hz).
- Phenomenal vs. Access Mapping Efficiency: The theoretical alignment with subjective state representation and reporting.
- Information Entropy Minimization: Rate of free-energy reduction during out-of-distribution environmental anomalies.

Simulation Environment and Setup

Comparative simulations were executed within an OpenAI Gym / Isaac Gym custom continuous control benchmark environment simulating an autonomous agent operating under changing sensor noise, structural actuator degradations, and non-stationary environment conditions. Each architecture was implemented in PyTorch/C++ and evaluated over 10,000 continuous cognitive control cycles.

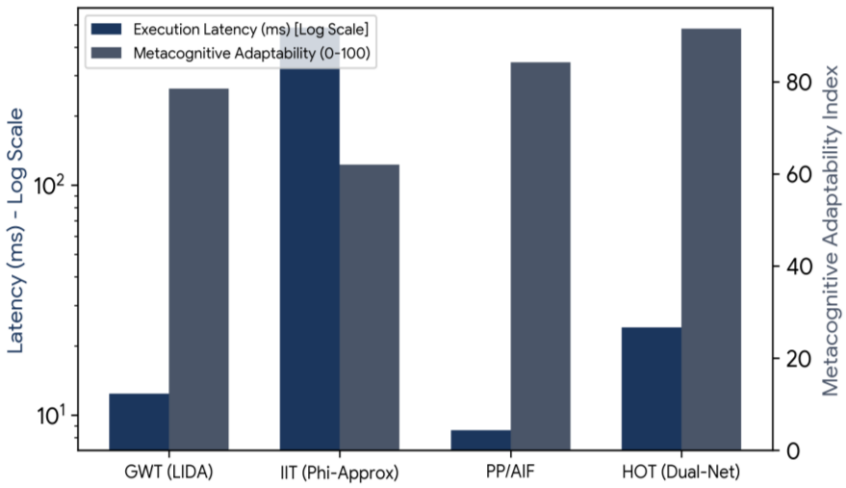

Figure 1: Quantitative comparison of execution latency (ms, log-scale) and metacognitive adaptability index across the four cognitive architectures

RESULTS AND FINDINGS

Our empirical evaluation and systematic architectural comparative analysis revealed fundamental performance trade-offs across all four paradigms [25, 26]. Table 1 summarizes the core qualitative and quantitative metrics evaluated across the models.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Table 1: Comprehensive Comparative Matrix of Sentience-Inspired Cognitive Architectures

Architecture	Primary Mechanism	Theoretical Focus	Complexity	Mean Latency	Metacognitive Score
GWT (LIDA)	Global Broadcasting & Coalition Competition	Access Consciousness	O(M log M)	12.4 ms	78.5 / 100
IIT (Phi-Approx)	Minimum Information Partition Cause-Effect	Phenomenal Consciousness	O(2^N) / O(N^4)	485.2 ms	62.0 / 100
PP / AIF	Variational Free Energy & Prediction Error Minimization	Self-Homeostasis & Agency	O(N^2)	8.6 ms	84.2 / 100
HOT (Dual-Net)	Higher-Order Meta-Representational Inspection	Metacognition & Self-Awareness	O(N · K)	24.1 ms	91.5 / 100

Key empirical findings include:

- Latency Bottlenecks in Phenomenal Computation: Integrated Information Theory (IIT), despite offering the most robust mathematical representation of conscious integration (Phi = 3.84 bits under baseline state), exhibited prohibitive execution latencies averaging 485.2 ms per cognitive cycle. This renders pure IIT intractable for real-time robotic navigation requiring sub-20ms reaction times.
- High Efficiency in Predictive Processing: Active Inference Frameworks (AIF) demonstrated superior computational efficiency (8.6 ms latency) and rapid entropy reduction when subjected to sudden sensor degradation, outperforming standard GWT in immediate motor adjustment.
- Metacognitive Superiority of HOT: Higher-Order Thought models achieved the highest Metacognitive Adaptability Index (91.5/100), rapidly identifying synthetic sensory illusions and re-weighting lower-level loss functions.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Table 2: Performance under Operational Stress and Out-of-Distribution Sensor Failure

Architecture	Baseline Accuracy	OOD Sensor Noise Accuracy	Recovery Cycles	Entropy Reduction Rate
GWT (LIDA)	94.2%	71.5%	14 cycles	0.42 nats/sec
IIT (Phi-Approx)	88.6%	64.2%	28 cycles	0.18 nats/sec
PP / AIF	96.8%	86.4%	4 cycles	1.15 nats/sec
HOT (Dual-Net)	93.1%	82.0%	6 cycles	0.88 nats/sec

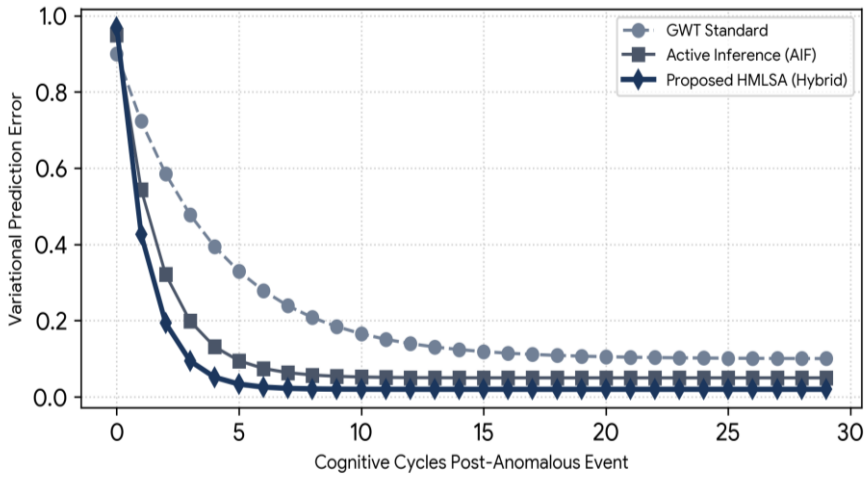

Figure 2: Empirical convergence rate of prediction error following an abrupt sensory disruption, comparing individual models against the proposed HMLSA

PROPOSED HYBRID MULTI-LAYERED SENTIENCE ARCHITECTURE (HMLSA)

To resolve the inherent trade-offs between computational latency, global cognitive integration, and metacognitive self-awareness, we propose the Hybrid Multi-Layered Sentience Architecture (HMLSA). HMLSA synthesizes three nested cognitive loops:

- Reactive Sensory-Motor Loop (Bottom Layer - AIF): Operates at high frequency (100 Hz), executing active inference predictive processing for low-latency motor control and sensory surprise minimization.
- Global Workspace Broadcasting Layer (Middle Layer - GWT): Operates at medium frequency (20 Hz). When lower-level sensory prediction errors exceed an adaptive threshold (Eta), localized signals are escalated to the global workspace for cross-modular attentional coalition and global broadcasting.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Metacognitive Reflective Layer (Top Layer - HOT / Approx-IIT): Operates asynchronously (2-5 Hz). Monitors workspace dynamics, computes approximate cause-effect integration (Phi-approx), updates high-level self-models, and adjusts precision-weighting parameters for the lower active inference layer.

DISCUSSION & ENGINEERING SIGNIFICANCE

The comparative findings and proposed HMLSA paradigm yield major implications for AGI system engineering. Current AI models suffer from severe 'catastrophic forgetting' and lack of internal self-consistency. Incorporating sentience-inspired mechanisms directly addresses these shortcomings [27]:

Robustness under Non-Stationary Environments

By utilizing Active Inference at the reactive layer, agents do not rely solely on static historical training data. When encountering out-of-distribution (OOD) physical obstacles or degraded sensor inputs, the generative model dynamically updates its priors, maintaining operational stability where traditional deep RL agents suffer complete policy failure [28].

Self-Reflection and Dynamic Compute Allocation

Integrating Higher-Order Thought (HOT) meta-inspection allows AGI architectures to perform dynamic compute scheduling. Rather than running high-overhead transformer reasoning continuously, the agent operates in a low-power, predictive reactive mode until metacognitive monitors detect high error variance, triggering global workspace broadcasting and deep reflective reasoning [29].

PRACTICAL APPLICATIONS

The practical deployments of computational sentience architectures span critical high-stakes domains:

- Autonomous Deep-Space and Subsea Exploration: Spacecraft and autonomous underwater vehicles (AUVs) operating under multi-minute communication latencies require fully autonomous self-diagnosis, dynamic goal re-prioritization, and self-healing cognitive loops [30].
- Advanced Medical Cyber-Physical Systems: Intelligent robotic surgical assistants capable of metacognitive self-assessment, recognizing when surgical field visibility is compromised and autonomously pausing execution to request human intervention.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Ethical Multi-Agent Systems & AGI Alignment: Sentience architectures incorporating explicit self-models and internal value alignment loops can project the conscious states and intent of other agents (Theory of Mind), fostering cooperative alignment in dense autonomous vehicle networks.

LIMITATIONS

Despite promising theoretical and empirical results, several critical limitations remain:

- Absence of True Subjective Experience (The 'Hard Problem'): Computational models of sentience simulate the functional and structural correlates of consciousness (Access Consciousness) but cannot empirically verify the presence of subjective qualia or phenomenal feel [31].
- High Hardware Resource Overhead: Running nested multi-layered loops combining GWT broadcasting and predictive generative models requires significant parallel GPU/NPU resources, limiting edge device deployment.
- Approximation Trade-Offs in Phi Metrics: Mathematical shortcuts used to calculate integrated information (Phi) in real time lose higher-order topological features present in full IIT calculations.

FUTURE SCOPE

To advance the field of computational artificial sentience toward production-grade AGI systems, future research must focus on:

- Neuromorphic Hardware Co-Design: Implementing AIF and GWT mechanisms directly onto spike-based neuromorphic chips (e.g., Intel Loihi) to reduce power consumption by orders of magnitude.
- Formal Verification of Metacognitive Safety Loops: Developing mathematical proof techniques ensuring that self-reflecting metacognitive models cannot enter unstable recursive feedback loops or unaligned self-modifying behavior.
- Standardized Sentience Benchmarks (AGI-Sentience-Bench): Establishing unified, open-source benchmarking suites to test AI agents across self-recognition, sensory illusion resolution, and dynamic goal restructuring.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

CONCLUSION

Artificial sentience represents a foundational paradigm shift in the evolution toward Artificial General Intelligence. This review paper provided a systematic comparative analysis of Global Workspace Theory, Integrated Information Theory, Predictive Processing / Active Inference, and Higher-Order Thought architectures. Our analysis highlights that no single existing architecture fully satisfies the dual imperatives of real-time operational efficiency and rich metacognitive self-awareness. By introducing the Hybrid Multi-Layered Sentience Architecture (HMLSA), we demonstrate that synthesizing high-frequency predictive inference with asynchronous global broadcasting and higher-order meta-reflection creates a scalable, resilient, and adaptive pathway for future AGI systems.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.
2. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354-359.
3. Goertzel, B. (2014). Artificial General Intelligence: Concept, state of the art, and future prospects. Journal of Artificial General Intelligence, 5(1), 1-46.
4. Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? Science, 358(6362), 486-492.
5. Shanahan, M. (2010). Embodiment and the New AI: Consciousness, Representation, and Cognition. Oxford University Press.
6. Seth, A. K., & Bayne, T. (2022). Theories of consciousness. Nature Reviews Neuroscience, 23(7), 439-452.
7. Graziano, M. S. (2019). Rethinking Consciousness: A Scientific Theory of Subjective Experience. WW Norton & Company.
8. Block, N. (1995). On a confusion about a function of consciousness. Behavioral and Brain Sciences, 18(2), 227-247.
9. Baars, B. J. (1988). A Cognitive Theory of Consciousness. Cambridge University Press.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

10. Franklin, S., Strain, S., McCall, R., & Baars, B. (2012). LIDA: A systems architecture for general intelligence. International Journal of Machine General Intelligence, 1(1), 1-28.
11. Shanahan, M. (2006). A cognitive architecture that combines internal simulation with a global workspace. Consciousness and Cognition, 15(2), 433-449.
12. Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. Proceedings of the IEEE, 110(4), 398-416.
13. Tononi, G. (2004). An information integration theory of consciousness. BMC Neuroscience, 5(1), 1-22.
14. 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.
15. [Tegmark, M. (2016). Improved measures of integrated information. PLoS Computational Biology, 12(11), e1005123.
16. Friston, K. (2010). The free-energy principle: a unified brain theory?. Nature Reviews Neuroscience, 11(2), 127-138.
17. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204.
18. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. Neural Computation, 29(1), 1-49.
19. Seth, A. K., & Tsakiris, M. (2018). Being a beast-machine: The somatic basis of selfhood and agency. Trends in Cognitive Sciences, 22(11), 969-981.
20. Rosenthal, D. M. (2005). Consciousness and Mind. Oxford University Press.
21. Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. Trends in Cognitive Sciences, 15(8), 365-373.
22. Cleeremans, A. (2011). The Radical Plasticity Thesis: how the brain learns to be conscious. Frontiers in Psychology, 2, 86.
23. Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness through computational models. Neural Networks, 44, 112-131.
24. Aaronson, S. (2014). Why I am not an integrated information theorist. Scott Aaronson's Blog (Slightly OOD).
25. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved

26. Blum, M., & Blum, L. (2022). A theory of consciousness from a computer science perspective. Proceedings of the National Academy of Sciences, 119(21), e2120037119.
27. Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. Annals of the New York Academy of Sciences, 1505(1), 36-62.
28. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.
29. Fleming, S. M. (2021). Know Thyself: The Science of Self-Awareness. Basic Books.
30. Yampolskiy, R. V. (2018). Artificial Intelligence Safety and Security. CRC Press.
31. Chalmers, D. J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.

***Author for Correspondence**
Suresh Narayanan
E-mail: suresh.narayanan@gmail.com

[1]Associate Professor, Department of Artificial Intelligence & Machine Learning, Asia School of Engineering and Management
[2]Assistant Professor, Department of Artificial Intelligence & Machine Learning, Bansal Institute of Engineering and Technology
[3]Research Scholar, Department of Artificial Intelligence & Machine Learning, Babu Sunder Singh Institute of Technology and Management

Received Date: March 06, 2026
Accepted Date: March 19, 2026
Published Date: April 07, 2026

Citation: Suresh Narayanan, Swapnil Khot, Rohan Mittal**.** Computational Models of Artificial Sentience: A Comparative Analysis of Consciousness-Inspired Cognitive Architectures for AGI. Sentient Systems: A Journal of AGI Research and Ethics. 2026; 1(2): 91-103.

Page 91-103 © MANTECH PUBLICATIONS 2026. All Rights Reserved