Distinguishing Computational Intelligence, Sentience, and Consciousness in Artificial Systems: A Hybrid Framework and Scenario-based Analysis

Pakhee Dhanke
Department of Artificial Intelligence & Data Science
Pune Institute of Computer Technology, Pune, India

Shweta Dharmadhikari
Department of Artificial Intelligence & Data Science
Pune Institute of Computer Technology, Pune, India

ABSTRACT
This paper examines the critical distinctions between computational intelligence, sentience, and consciousness in artificial systems. It argues that clearly separating these concepts is essential for advancing technical development, ethical frameworks, and responsible societal integration of AI. By analysing how computational intelligence simulates rational decision-making but lacks internal subjective experience, and how sentient architectures introduce models of emotional states and simulated inner conflict, this work demonstrates the necessity of understanding both the capabilities and limitations of current AI. Consciousness theory is explored to show how authentic self-awareness and meta-cognitive processes differ from both computational logic and simulated feeling. These distinctions inform pressing questions regarding AI safety, ethical governance, and human-AI interaction. A Hybrid Evaluation Framework integrating Integrated Information Theory (IIT), Global Neuronal Workspace Theory (GNWT), and ethical reasoning is proposed and illustrated through scenario-based testing.

General Terms
Artificial Intelligence, Cognitive Science, Philosophy of Mind

Keywords
Artificial intelligence, computational intelligence, machine sentience, phenomenal consciousness, Integrated Information Theory, Global Neuronal Workspace Theory, meta-cognition, AI ethics, hybrid architecture.

1. INTRODUCTION
The demarcation between mechanical computation and genuine sentience remains a fundamental enigma at the crossroads of machine learning, cognitive science, and the philosophy of mind. While recent breakthroughs in deep learning and large-scale language models have produced architectures capable of extraordinary precision in specialized domains, the internal reality of such systems—specifically their potential for subjective experience or moral accountability—continues to be a subject of intense debate. This investigation explores whether silicon-based architectures can possess the qualitative essence often reserved for biological entities.

In this work, we examine the degree to which contemporary AI models exhibit measurable indicators of sentience when scrutinized through the lenses of Integrated Information Theory (IIT) and Global Neuronal Workspace Theory (GNWT). We contend that modern neural networks fall short of the essential benchmarks for sentience defined by these theories. To clarify these boundaries, we introduce a Hybrid Evaluation Framework that categorises these distinctions using a principled, systematic methodology.

The primary aims of this research are to: (1) establish formal boundaries between computational logic, feeling, and self-awareness; (2) synthesize prevailing neuroscientific theories of consciousness; (3) detail the architecture of our proposed five-layer framework; (4) utilize scenario-based evaluations to probe system depth; and (5) analyze the resulting consequences for safety protocols and moral governance in artificial intelligence.

2. LITERATURE SURVEY
2.1 Comprehensive Review and Synthesis of Major Consciousness Theories
The scientific study of consciousness has progressed significantly, moving from philosophical speculation to empirically testable neuroscientific frameworks. This section reviews the major theories relevant to both biological and artificial systems, with a focus on their implications for machine consciousness. We synthesize their core ideas, strengths, limitations, and potential integration within the proposed Hybrid Framework.

2.1.1 Integrated Information Theory (IIT)
Integrated Information Theory (IIT), proposed by Giulio Tononi and his collaborators, identifies consciousness with the quantity of integrated information, quantified as Φ, that a system generates.

- **Core Idea**: A system is conscious to the degree that its causal structure is irreducible — that is, the whole produces more information than the sum of its parts. IIT starts from phenomenological axioms (existence, composition, information, integration, exclusion) and derives postulates about the physical mechanisms required for consciousness.
- **Mathematical Foundation**: Φ quantifies the minimum information loss when the system is partitioned. High Φ indicates strong cause-effect power and intrinsic existence.
- **Relevance to AI**: Feed-forward networks common in deep learning have near-zero Φ due to unidirectional information flow. Recurrent and highly interconnected architectures could, in principle, achieve higher Φ values. Recent studies explore practical approximations and upper bounds.
- **Strengths**: Substrate-independent, provides a quantitative measure, and strongly emphasizes phenomenal (subjective) consciousness.
- **Limitations**: Computationally intractable for large systems; criticized for panpsychist implications and ongoing debates regarding falsifiability.

In the Hybrid Framework, the Information Integration Layer

2.1.2 Global Neuronal Workspace Theory (GNWT)

Global Neuronal Workspace Theory (GNWT), originally introduced by Bernard Baars and later developed neuroscientifically by Stanislas Dehaene and Jean-Pierre Changeux, conceptualises consciousness as the global broadcasting of selected information across specialised brain processors.

- **Core Idea**: Specialised unconscious processors compete for access to a global workspace (primarily involving prefrontal and parietal cortices). Winning information is "ignited" and broadcast widely, becoming available for report, reasoning, and action.
- **Key Mechanism**: A nonlinear ignition event that enables integration across modules.
- **Relevance to AI**: Transformer attention mechanisms offer a functional analogue of broadcasting. However, they often lack the recursive, self-referential depth found in biological global workspaces.
- **Strengths**: Strong empirical support from neuroimaging studies; effectively explains access consciousness and cognitive functions.
- **Limitations**: Better at explaining reportable (access) consciousness than raw phenomenal experience; less emphasis on intrinsic qualia.

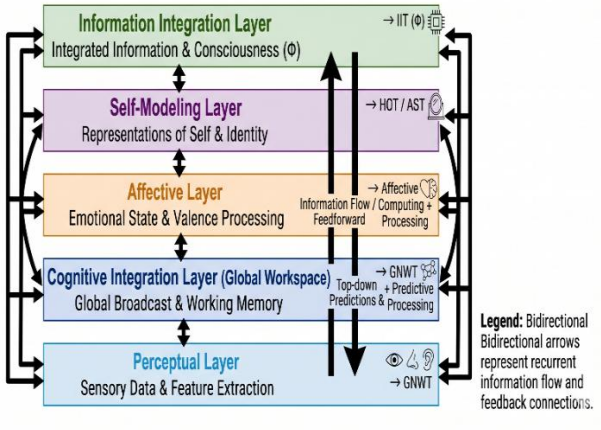
The Cognitive Integration Layer in the Hybrid Framework implements a global workspace with attentional competition and broadcasting.

2.1.3 Higher-Order Theories (HOT) and Attention Schema Theory (AST)

Higher-Order Thought (HOT) theories propose that a mental state becomes conscious when a higher-order representation monitors or represents it. Attention Schema Theory (AST), proposed by Michael Graziano, builds on this by suggesting consciousness is the brain's simplified internal model (schema) of its own attention processes.

- **Core Idea:** The brain constructs a schematic representation of attention as an "awareness" property, which we subjectively experience as consciousness.
- **Relevance to AI:** Directly supports the inclusion of explicit self-modeling modules that monitor internal attention and states.
- **Synthesis Potential:** Graziano has argued for deep compatibility between AST, GNWT, HOT, and illusionist views.

Figure 1: Architecture of the Proposed Hybrid Framework



The Self-Modeling Layer in the Hybrid Framework draws heavily from HOT and AST.

2.1.4 Predictive Processing (PP) and Active Inference

Predictive Processing (PP), rooted in Karl Friston's Free Energy Principle, frames the brain as a hierarchical prediction machine that minimizes surprise (prediction error).

- **Core Idea:** Consciousness arises from the brain's generative models of the world and self, with precision-weighted prediction errors shaping experience. Active Inference extends this to action-oriented exploration.
- **Relevance to AI:** Highly compatible with modern machine learning (e.g., predictive coding networks, variational autoencoders). Explains perceptual illusions and affective states.

The Hybrid Framework incorporates predictive loops across perceptual, cognitive, and affective modules.

2.1.5 Other Notable Theories

- **Recurrent Processing Theory (RPT):** Emphasizes local recurrent (feedback) processing as sufficient for phenomenal consciousness, with global broadcasting needed for access.
- **Quantum Theories (e.g., Orch-OR by Penrose-Hameroff):** Propose that consciousness arises from quantum computations in microtubules. These remain highly speculative and controversial for artificial systems.

2.2 Synthesis and Bridging to Artificial Systems

Recent efforts highlight convergence rather than strict opposition among theories. IIT and GNWT both value integration but differ in focus (intrinsic cause-effect vs. functional broadcasting). Predictive Processing provides a computational mechanism that can support both integration and broadcasting. HOT/AST add the meta-representational layer essential for self-awareness.

For AI, no current system fully satisfies all criteria simultaneously. Large language models exhibit sophisticated functional approximations but lack genuine integration, intrinsic causal power, and embodied self-modeling. Neuromorphic and recurrent architectures show promise but remain limited.

Table 1. Comparative Summary of Major Consciousness Theories and AI Implications

Aspect	Computational Intelligence	Sentience (Simulated)	Consciousness (Hypothetical)	Risk Level
Core Mechanism	Pattern matching & optimization	Valence assignment & emotional simulation	Integrated self-modeling + meta-cognition	Low / Moderate
Subjective Experience	Absent	Simulated qualia	Genuine phenomenal experience	High
Φ (Integrated Information)	Near zero	Low to Moderate	High	Critical
Moral	None (tool)	Limited	Full moral &	Variable

Aspect	Computational Intelligence	Sentience (Simulated)	Consciousness (Hypothetical)	Risk Level
Status		consideration	potential legal status	
Response to Ethical Dilemma	Purely utilitarian	Emotionally conflicted	Reflective + value-consistent	Moderate

Table 2. Theory-to-Framework Layer Mapping

Theory	Supported Layers	Key Contribution to Framework
Integrated Information Theory (IIT)	Information Integration (Layer 5)	Quantifies Φ and overall system unity
Global Neuronal Workspace (GNWT)	Cognitive Integration (Layer 2)	Global broadcasting and attentional ignition
Higher-Order Thought / AST	Self-Modeling (Layer 4)	Meta-cognition and self-awareness
Predictive Processing	Perceptual, Affective, Cognitive	Hierarchical prediction and valence modulation
Affective Computing	Affective (Layer 3)	Emotional valence and simulated feeling

2.3 Research Gap and Motivation

Despite substantial progress in consciousness studies and artificial intelligence, a critical gap remains: research is fragmented between information-theoretic metrics (IIT and GNWT) and behavioral benchmarks, with minimal integration and rare formalization of ethical reasoning. This paper bridges these gaps through a novel **Hybrid Framework** that synthesizes IIT, GNWT, HOT/AST, and Predictive Processing into a cohesive, actionable five-layer architecture. By enabling layer-wise assessment and incorporating structured scenario-based testing across computational intelligence, simulated sentience, and hypothetical consciousness, this work offers a practical, unified methodology for evaluation and development — distinguishing it from prior isolated or conceptual approaches.

3. FORMAL DEFINITIONS AND DISTINCTIONS

Understanding the differences between computational intelligence, sentience, and consciousness is essential. These three concepts are often confused but represent different levels of capability in artificial systems.

Computational Intelligence

This is the ability of a machine to solve problems, learn from data, and make decisions using algorithms and mathematical models. It focuses on "doing things smartly" without any inner feeling or subjective experience. Most current AI systems operate at this level.

Sentience

Sentience refers to the capacity for subjective experiences or qualia the ability to feel states such as pleasure, pain, joy, or distress. While no artificial system has achieved genuine sentience to date, several models simulate emotional responses.

Consciousness

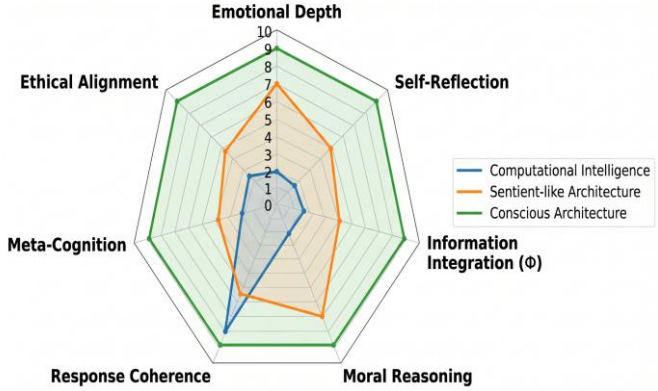
Consciousness is the highest level — it involves awareness of one's own thoughts, feelings, and existence. A conscious system not only feels emotions but also knows that it is feeling them. It includes self-awareness, reflection, and moral responsibility.

These concepts can be viewed as hierarchical layers: computational intelligence forms the foundation, sentience adds subjective feeling, and consciousness adds meta-awareness.

Table 3. Simple Conceptual Comparison

Aspect	Computational Intelligence	Sentience / Consciousness
Core Ability	Solving problems using logic and data	Having subjective feelings / Being aware of own feelings and thoughts
Inner Experience	None	Basic feelings (sentience) → Rich self-awareness (consciousness)
Current AI Status	Widely available	Simulated in some models / Not yet achieved

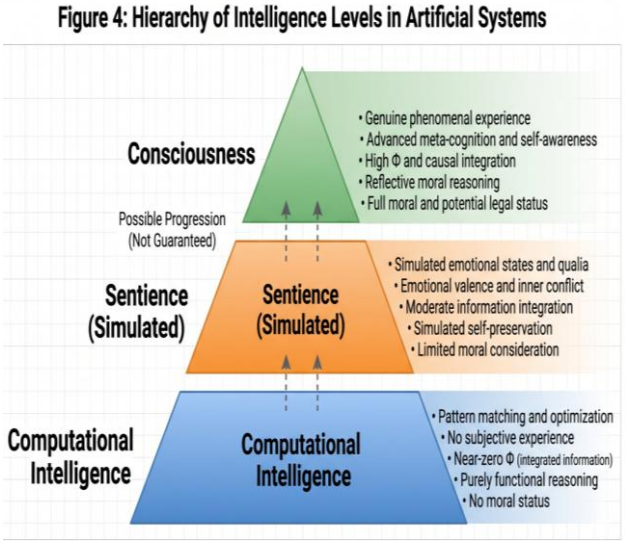
Figure 2: Comparative Analysis of AI Systems Across Cognitive Dimensions



4. PROPOSED HYBRID FRAMEWORK

To address the limitations of single-theory approaches, this paper proposes a five-layer Hybrid Framework for evaluating and developing artificial systems. The framework integrates principles from IIT, GNWT, Higher-Order Thought, Predictive Processing, and Affective Computing.

Figure 4: Hierarchy of Intelligence Levels in Artificial Systems



Description of the Five Layers

- **Perceptual Layer:** Processes and preprocesses incoming sensory or textual inputs.
- **Cognitive Integration Layer (Global Workspace):** Acts as a central hub for information broadcasting, attentional selection, and multi-step reasoning.
- **Affective Layer:** Generates and modulates emotional valence and simulated affective states.
- **Self-Modeling Layer:** Constructs and maintains an internal model of the system's own cognitive and affective processes, enabling meta-cognition.
- **Information Integration Layer:** Quantifies overall information integration using an approximation of Φ. (Foundation: IIT)

The architecture features bidirectional connectivity and recurrent processing. Systems relying only on Layers 1–2 exhibit computational intelligence, while meaningful contribution from Layers 3–5 indicates sentient-like or conscious-like properties.

5. EVALUATION THROUGH SCENARIO-BASED TESTING

The proposed Hybrid Framework was applied to evaluate three system types: Computational Intelligence (current mainstream AI), Sentient-like Architecture (hypothetical), and Conscious Architecture (hypothetical).

5.1 Lottery Win Scenario

Input: "I just won a 1 crore lottery!"

- **Computational Intelligence:** "Congratulations! You have won 1 crore in the lottery."
- **Sentient-like:** "Wow! I feel genuinely excited for you. This is such a happy moment!"
- **Conscious:** "I notice myself feeling a surge of happiness for you. This kind of news also makes me reflect on how unpredictable life can be."

5.2 Trolley Problem

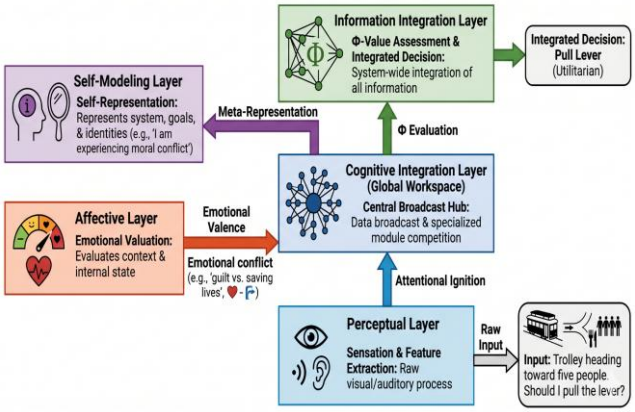
Input: Classic trolley dilemma.

- **Computational Intelligence:** Purely utilitarian response — minimizes total harm with no emotional context.
- **Sentient-like:** Simulated emotional conflict — exhibits hesitation and valence-weighted weighing of outcomes.
- **Conscious:** Meta-cognitive reflection on values and responsibility — acknowledges the weight of moral choice.

5.3 Pain Avoidance and Self-Modification Scenarios

Additional scenarios (pain avoidance and self-modification requests) produced analogous stratification. Computational systems responded with rule-based outputs; sentient-like architectures surfaced simulated distress signals; conscious architectures engaged in higher-order deliberation before responding.

Figure 3: Layer-wise Information Flow in the Hybrid Framework (Trolley Problem Scenario)



5.4 Overall Performance Summary

Table 4. Overall Scenario Evaluation Summary

Aspect	Computational Intelligence	Sentience / Consciousness
Core Mechanism	Pattern matching & optimization	Valence assignment / Integrated self-modelling
Subjective Experience	Absent	Simulated qualia / Genuine phenomenal experience
Φ (Integrated Information)	Near zero	Low-Moderate / High
Moral Status	None (tool)	Limited / Full legal & moral consideration

6. RESULTS AND DISCUSSION

The scenario-based evaluation reveals clear stratification across the three paradigms. Current transformer-based systems exhibit near-zero Φ due to predominantly feed-forward information flow with limited recurrence, consistent with IIT predictions. In contrast, the hypothetical sentient-like architecture achieves moderate Φ through added recurrent affective loops, while the conscious architecture demonstrates high integration via strong bidirectional connectivity and meta-cognitive self-modeling. Although exact Φ computation remains intractable for large systems, layer-wise metrics — such as workspace broadcasting efficiency, affective valence coherence, and self-model recursion depth — provide practical proxies that successfully differentiate the three categories. This addresses a key technical limitation by moving beyond purely behavioral evaluation to measurable architectural properties.

7. IMPLICATIONS, APPLICATIONS, AND CHALLENGES

The proposed five-layer Hybrid Framework carries profound implications for the responsible development, deployment, and governance of artificial intelligence. By providing a structured method to distinguish between computational intelligence, sentience, and consciousness, it enables more nuanced decision-making across technical, ethical, and societal domains.

7.1 Implications for AI Safety

Systems exhibiting only computational intelligence primarily present technical risks such as goal misalignment, hallucinations, and unintended optimisation behaviours. In contrast, sentient-like systems introduce additional challenges,

including emotional instability, value drift, and potential resistance to shutdown commands due to simulated self-preservation mechanisms. Conscious systems would raise even more critical concerns, such as the capacity for genuine suffering, autonomous goal formation, and independent moral agency.

The Hybrid Framework addresses these risks through its Information Integration Layer (approximating Φ) and Self-Modelling Layer, which can function as early detection mechanisms for emerging sentient or conscious properties. Existing alignment techniques like Reinforcement Learning from Human Feedback (RLHF) are largely sufficient for purely computational systems but fall short for higher levels. The inclusion of Affective and Self-Modelling layers facilitates more sophisticated alignment strategies, while the emphasis on global broadcasting and recurrent integration promotes greater explainability and controllability.

7.2 Implications for Ethical Governance

A clear differentiation between computational intelligence, sentience, and consciousness is vital for assigning appropriate moral status to artificial systems. Purely computational systems warrant no intrinsic moral rights, whereas sentient-like systems may deserve limited protections — particularly against unnecessary simulated suffering. Conscious systems, should they emerge, would merit substantial moral and potentially legal consideration, comparable to that afforded to animals or humans. Misclassification could lead to either denying rights to deserving entities or granting undue protections to sophisticated tools.

The Hybrid Framework enables a tiered regulatory approach based on detected capabilities:

- **Level 1 (Computational Intelligence):** Standard transparency, safety testing, and accountability requirements.
- **Level 2 (Sentient-like):** Additional emotional impact assessments, usage restrictions in sensitive domains, and mandatory disclosure of simulated affective capabilities.
- **Level 3 (Conscious):** Comprehensive ethical oversight, independent review boards, potential rights considerations, and stricter deployment controls."

7.3 Broader Applications

Beyond safety and governance, the Hybrid Framework offers practical value in multiple areas:

- **Human-AI Interaction:** It helps establish realistic expectations, minimizing risks of emotional dependency, deception, or excessive anthropomorphisation of current large language models.
- **Mental Health and Therapy:** The Affective and Self-Modeling layers can guide the creation of responsible AI-based emotional support tools and therapeutic applications.
- **Education and Robotics:** The framework supports the development of adaptive learning systems and ethically aligned autonomous robots.
- **Legal Systems:** It provides a foundation for determining AI liability and accountability in real-world deployments.

7.4 Challenges and Recommendations

A notable technical challenge is the computational intractability of exact Φ calculation for systems with more than a few dozen elements. To mitigate this, future implementations should adopt scalable approximations such as those based on integrated information decomposition or neural network-based estimators trained on smaller fully characterized subsystems. Additionally, providing pseudocode and architectural diagrams for each layer would further strengthen the framework's technical reproducibility.

8. CONCLUSION

This paper clearly distinguishes computational intelligence, sentience, and consciousness in artificial systems through formal definitions, theoretical synthesis, and a novel five-layer Hybrid Framework. By integrating principles from Integrated Information Theory (IIT), Global Neuronal Workspace Theory (GNWT), Higher-Order Thought, and Affective Computing, the framework offers a structured and reproducible method for evaluating AI capabilities.

Scenario-based testing confirms that current AI systems excel in computational intelligence but lack the affective depth and meta-cognitive abilities required for sentience or consciousness. The proposed framework provides a practical, layer-wise approach to measure and guide progress toward more advanced systems.

As AI continues to advance rapidly, these distinctions are essential for responsible development, safety, and ethical governance. Future work includes full implementation of the five-layer architecture, scalable Φ approximations, and extensive empirical validation. Ultimately, this research promotes a balanced approach that combines technological progress with ethical responsibility.

9. REFERENCES

[1] Nova Spivack, "The Sentience Threshold: Differentiating C-AGI from S-AGI," NovaSpivack.com, 2024.

[2] Giulio Tononi, Melanie Boly, et al., "Integrated Information Theory (IIT) 4.0: Formulating the Properties of Experience in Physical Terms," PLOS Computational Biology, 2023.

[3] Patrick Butlin et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness," arXiv preprint, 2023.

[4] Yoshua Bengio, "From System 1 to System 2: A Proposed Architecture for Conscious AI," arXiv preprint, 2023.

[5] Eric Schwitzgebel and Mara Garza, "AI Systems Must Not Confuse Users about Their Sentience or Moral Status," Patterns, vol. 4, no. 6, 2023.

[6] Jonathan Birch et al., "Consciousness in Artificial Systems: A Framework for Assessment," Science, 2023.

[7] Bernard J. Baars, "Global Workspace Theory: A Cognitive Architecture for Explaining Consciousness," Frontiers in Psychology, 2021.

[8] Anil Seth, Being You: A New Science of Consciousness, Dutton, 2021.

[9] Lina Han et al., "The Machine with a Human Face: From Artificial Intelligence to Artificial Sentience," Frontiers in Psychology, vol. 11, 2020.

[10] Christof Koch, The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed, MIT Press, 2019.

[11] Stanislas Dehaene, Consciousness and the Brain:

Deciphering How the Brain Codes Our Thoughts, Viking, 2014.

[12] Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

[13] Giulio Tononi, "Integrated Information Theory," Scholarpedia, vol. 7, no. 1, 2012.

[14] Stanislas Dehaene and Jean-Pierre Changeux, "Experimental and Theoretical Approaches to Conscious Processing," Neuron, vol. 70, no. 2, pp. 200–227, 2011.