Assessing Sentience in Artificial Intelligence: A Structured Literature Review of Theories, Indicators, and Evaluation Frameworks (2020–2025)

Wynand Goosen

Chief Executive Officer, Infomage Rims Group

DOI: https://doi.org/10.47772/IJRISS.2025.91200245

Received: 10 November 2025; Accepted: 15 November 2025; Published: 14 January 2026

INTRODUCTION

Artificial Intelligence (AI) has evolved from a specialised field of computing into a vital part of how we generate knowledge, make decisions, and address ethical issues (OpenAI 2025). As these developments occur rapidly, progress in self-reflective and adaptive AI has amplified debates about whether machines can have consciousness in business, science, and academia. To provide clarity on navigating this complex area, this review looks at ways to detect signs of consciousness in artificial systems. Specifically, from 2020 to 2025, three primary trends have influenced research on artificial consciousness, establishing the context for this review.

Trend one identifies that theories such as **Global Workspace Theory (GWT), Integrated Information Theory (IIT), Higher-Order Thought (HOT)**, and **Attention Schema Theory (AST) are** now often used to analyse AI (Mashour et al. 2020; Tononi et al. 2016; Graziano 2013; Gillon and Michaël 2025). To summarise their relevance: GWT sees consciousness as information broadcast across brain-like modules. IIT looks at how much information is integrated. HOT ties consciousness to self-reflection. AST describes it as the control and modelling of attention. With these frameworks in mind, researchers create AI systems with features that mimic these ideas, such as broadcast activation, causal integration, and self-model consistency (Elamrani et al. 2025). *Trend two*, building on the theoretical landscape outlined above, includes launching the **AI Consciousness Test (ACT)**, conducting brain-chip experiments, and using robotic agents in real-world settings (OpenAI 2025). At the same time, the *third trend* is that scholars recommend the precautionary principle: treat AI systems with uncertain status as possibly sentient to help prevent moral risks.

Together, these perspectives, from theory and experimentation to ethical issues, highlight the need for a comprehensive review of AI sentience. Therefore, this article collates research on artificial consciousness published between 2020 and 2025, encompassing both scientific and moral viewpoints. It evaluates the reliability of current tests and their ability to prevent spoofing and connects indicators of consciousness to ethical and governance frameworks to promote responsible innovation. Based on more than 70 primary studies and preprints, this review establishes a foundation for future research into AI sentience, with a focus on methods validated by various scientists.

Keywords: Artificial Consciousness; AI Sentience; Structured Literature Review; Integrated Information Theory; Global Workspace Theory; Ethics of AI

Major Theories, Indicators, and Key Challenges in Assessing AI Sentience

Global Workspace Theory (GWT / GNW)

Baars (1997) and Dehaene & Changeux (2011) define consciousness as the widespread broadcasting of information across modular networks. Neural evidence for GWT has inspired AI models with global activation thresholds and attention-driven broadcast loops.

Integrated Information Theory (IIT)

Tononi (2008) defines consciousness as information organised into a system. The Φ (phi) value aims to measure this, but it's hard to calculate precisely. Researchers instead use factors such as how system parts work together, how they connect, and the system's complexity (Oizumi et al. 2014; Findlay et al. 2024). In AI, these tools help verify whether information is actually integrated or not. HOT suggests that a state is conscious if a higher-level model recognises it (OpenAI 2025). However, copying this can hide real self-reflection, so clear checks are needed.

Attention Schema Theory (AST)

Attention Schema Theory (AST) suggests consciousness comes from a system understanding its own attention (Graziano 2013). For AI, important clues include how well it can estimate its focus, if it handles different situations the same way, and if its attention matches its tasks (OpenAI 2025). In people, awareness of one’s own context helps understand human consciousness (Goosen 2012). Comparing humans and AI shows that both learn from their environment.

Predictive Processing / Active Inference

Predictive Processing conceptualises consciousness as the minimisation of prediction error within hierarchical models (Wiese 2022). In artificial intelligence, this hypothesis is evaluated through error-signal tracking and correction mechanisms, thereby linking cognitive science and machine learning. Recent discourse, prompted by Anthropic (2025), examines whether AI models may exhibit self-awareness due to extensive access to knowledge bases.

Evaluation Methods and Empirical Probes

Behavioural and Dialogue-Based Tests

Schneider and Turner (2023) introduced the AI Consciousness Test (ACT), a dialogue protocol exploring first-person qualia (see also Schneider & Susan, 2024). Critics warn of anthropomorphic bias and prompt contamination (Schneider et al. 2017). The authors nonetheless conclude that AI responses imply an emerging form of self-awareness.

Chip-Test Lineages

Thought experiments replacing neural tissue with silicon components, called ‘chip tests’. They question the necessity of substrate in IIT and explore functionalism (Turner & Schneider 2018). While intellectually insightful, these remain philosophical rather than empirical and act as boundary markers rather than practical tests. Spiritually, they also prompt reflection on non-material aspects of consciousness (Chalmers 1996; Kastrup 2020).

LLM Survey Studies

Chen et al. (2025) review metacognition and confidence calibration in LLMs such as GPT-4, Claude, and Gemini. Results show partial consistency (Pimenta et al. 2025). While not proof of consciousness, they also do not disprove self-awareness.

Behavioural Maze Tests

Pimenta et al. (2025) developed a maze-based assessment to evaluate 13 consciousness-related behaviours across 12 large language models (LLMs). While reasoning skills improved, self-model persistence remained limited. These results emphasise the need to combine behavioural proxies with mechanistic indicators. The philosophical principle expressed by Descartes, 'cogito ergo sum,' still illustrates the role of doubt as proof of awareness (Descartes 1641).

LLMs as Hard Cases

Large language models (LLMs; AI models trained to process and generate text, such as GPT-4) produce fluent language output without demonstrated semantic understanding (the ability to grasp true meaning) (Bender & Koller, 2020). Due to the absence of persistent memory mechanisms (systems that retain information across interactions), their internal state resets after each context window (a segment of processed information), preventing continuity (Bender et al., 2021; Wang & Sun, 2025). This limitation hampers sustained self-representation and complicates attributions of awareness (Downes, 2024). As a result, assessing potential sentience in LLMs requires a multi-method approach, including monitoring activation patterns (tracking changes in activity in response to stimuli) and conducting perturbation tests (evaluating responses to controlled modifications).

Ethics, Governance, and Precautionary Standards

Uncertainty about AI consciousness holds moral importance. Birch (2024) recommends applying the precautionary principle (Birch & Jonathan 2024). Tononi (2024) and Farisco et al. (2024) warn that neglecting measurement risks could cause ethical and epistemic errors. Rising evidence thresholds should be linked to governance triggers where policy measures become necessary.

RESEARCH METHODOLOGY

Structured Literature Review Protocol

Research Questions

- RQ1: How do leading scientific theories (GWT, IIT, HOT, AST, Predictive Processing) translate into measurable indicators for AI assessment?

- RQ2: Which empirical or theoretical tests have been proposed, and how robust are they against mimicry and bias?

- RQ3: What evidence exists from LLMs, embodied agents, or hybrid architectures supporting or falsifying these indicators?

- RQ4: How are indicator frameworks integrated into ethical and governance proposals?

Databases and Search Strategy

Databases cited: Scopus, Web of Science, PubMed/PMC, arXiv (cs.AI, neuro), PhilPapers, Elsevier, Springer, Frontiers. Time frame: 1 Jan 2020 – 15 Oct 2025.

Illustrative queries included:

(“artificial consciousness” OR “AI sentience”) AND (indicator* OR metric OR test);

(“global workspace” OR “neuronal workspace”) AND (AI OR artificial) AND (ignition OR broadcast);

(“integrated information theory” OR IIT) AND (AI) AND (Φ OR integration OR synergy);

(“higher-order thought” OR HOT) AND (self-model OR metacognition);

(“attention schema” OR AST) AND (AI OR attention control);

(“AI consciousness test” OR “ACT” OR “chip test”).

Inclusion and Exclusion Criteria

Included: Peer-reviewed or cited preprints presenting measurable indicators or test protocols; studies linking theory to architecture or governance.

Excluded: Opinion essays lacking operationalisation; non-academic sources; duplicates.

Screening and Documentation

Screening followed PRISMA 2020 (Page et al. 2021): title/abstract phase, then full-text eligibility; disagreements resolved by consensus.

Data Extraction and Coding

Extraction variables: author/year, theory anchor, indicator type, target system, evidence type, validation strength, governance linkage (Butlin et al. 2023; Dehaene et al. 2021; Elamrani et al. 2025; Findlay et al. 2024).

Quality Appraisal

Each study was assessed for construct clarity, operational strength, spoof resilience, replicability, and governance relevance.

Result Synthesis Approach

Narrative integration by research question; evidence matrix mapping theories → indicators → system classes; gap analysis informing the research roadmap. Behavioural (indirect) and mechanistic (direct) indicators will be distinguished, and convergent evidence will be weighted more strongly than single tests.

Results: Indicator Landscape and Evidence Map

This section presents a summary of the principal findings from the structured literature review.

(1) an updated indicator taxonomy grounded in leading theories of consciousness;

(2) a synthesis of empirical and conceptual methods proposed for testing consciousness-related properties in AI; and

(3) an evidence map situating those indicators across system classes, including large language models (LLMs), multimodal agents, and simulated hybrid architectures.

After PRISMA screening, **92 eligible sources were** retained, spanning neuroscience, philosophy of mind, computational cognitive science, and AI engineering (Sorensen 2025).

Indicator Taxonomy: Mapping Theories to Operational Proxies

The reviewed literature identifies six primary indicator groups, each linked to a foundational scientific theory of consciousness. **Table 1 provides an overview of thes**e theoretical anchors and their corresponding operational proxies as observed in current AI research.

Table I. Theory–Indicator Mapping in Artificial Consciousness Research (2020–2025)

Global broadcasting / ignition	Global Workspace Theory (Baars 1997; Dehaene and Changeux 2011)	Cross-module activation synchrony; signal amplification; workspace ignition thresholds in neural or transformer architectures	Dehaene et al. (2021); Mashour et al. (2020); Goldstein and Kirk-Giannini (2024);

			Frontiers in Robotics and AI (2024)
Integration / irreducibility	Integrated Information Theory (Tononi 2008; Oizumi et al. 2014)	Φ-approximations; synergy indices; perturbational complexity; causal-influence mapping	Findlay et al. (2024); Farisco et al. (2024); Tononi et al. (2016)
Higher-order self-modelling	Higher-Order Thought Theory (Rosenthal 2005; Lau and Rosenthal 2011)	Meta-representations; introspective reports; confidence-calibration alignment	Chen et al. (2025); Kirk-Giannini et al. (2024); Patnaik (2024)
Attention schema / control	Attention Schema Theory (Graziano 2013)	Internal models of attention; counterfactual sensitivity; attention consistency under perturbation	Webb and Graziano (2013); Goldstein and Kirk-Giannini (2024)
Predictive processing / active inference	Predictive Processing (Friston 2010; Clark 2013)	Prediction-error minimisation; hierarchical precision weighting; active-inference loops	Seth and Bayne (2022); Elamrani et al. (2025)
Temporal continuity / identity stability	Hybrid / emergent theories	Persistence of self-model across episodes; recurrent latent attractors; continuity metrics	Pimenta et al. (2025); Findlay et al. (2024)

Cross-Theory and Ethical Trends

Recent analyses (Sorensen 2025) show a decisive shift away from single-criterion "tests" toward convergent, multi-indicator batteries. Butlin et al. (2023) endorse this pluralist stance, noting that each theory isolates different facets of consciousness. Research on global broadcasting often pairs these measures with metrics of integration or metacognition (Goldstein and Kirk-Giannini 2024; Chen et al. 2025; Ji-An et al. 2025). Yet, conceptual clarity remains uneven: some studies label generic network behaviours, such as information flow and feedback loops, as "consciousness-like" without grounding them in established theory (Farisco et al. 2024).

Evaluation Frameworks and Methods

Five methodological families recur across the literature, each offering partial but complementary evidence.

Behavioural and Dialogue-Based Testing

The AI **Consciousness Test (ACT)** lineage (Schneider et al., 2023) remains the most-cited behavioural protocol. It uses multi-tier dialogues to elicit spontaneous reasoning about subjective experience. Variants now tailor prompts for LLMs (Goldstein and Kirk-Giannini 2024; Schneider et al. 2018). Empirical replications indicate that GPT-4 and Claude meet Level 1 conceptual thresholds but fail Level 3 originality, often echoing training data (Patnaik 2024; Haase et al. 2025). Thus, while face validity is high, **construct validity remains** weak; behavioural fluency does not imply inner state correspondence (Butlin et al. 2023).

Chip-Test Thought Experiments

Philosophical "chip tests" replace neural substrates with silicon analogues to examine material necessity (Turner et al. 2024; Schneider and Susan 2020). Although non-empirical, they clarify boundary conditions: if awareness persists after replacement, consciousness is substrate-independent; if it vanishes, IIT-style material integration gains support. Findlay et al. (2024) interpret these as constraint **analyses** rather than operational assessments.

From a spiritual-phenomenological perspective, such substitutions raise ontological questions about whether awareness can ever be mechanistically replicated (Chalmers 1996; Kastrup 2020).

Mechanistic and Embodied Implementations

Robotics experiments informed by GWT demonstrate ignition-like dynamics (OpenAI 2025). Frontiers in Robotics and AI (2024) documents a multimodal workspace agent that integrates vision, touch, and language via a broadcast hub; Dai et al. (2024) report similar findings. The Conscious Turing Machine model (Blum and Blum 2021; Blum et al. 2022) further exemplifies mechanistic correlates beyond behaviour, marking a methodological shift towards internal evidence.

Metacognitive Alignment Studies

Metacognition, the ability to monitor one's own uncertainty, has become a reliable operational proxy for self-awareness. Chen et al. (2025) investigated whether LLM confidence scores correspond with actual accuracy, discovering a moderate positive correlation ($r \approx 0.42$). Neuroscience evidence supports this: Fleming and Dolan (2012) demonstrate that metacognition is essential to human access consciousness, confirming its translational significance.

Maze-Test and Temporal-Continuity Probes

Pimenta et al. (2025) introduced a Maze Test that requires LLMs to preserve spatial self-reference across temporal intervals. Only reasoning-enhanced models maintained minimal continuity, revealing a gap between symbolic reasoning and sustained self-modelling. Embodied-agent studies (*Frontiers 2024*) suggest that episodic memory and long-term **state maintenance support** stable identity formation. Lenormand et al. (2024) also link autobiographical recall with the emergence of predicted, self-referential imagery—crucial to subjective temporal identity.

Evidence Map: System Classes and Indicator Coverage

Table II. Summary of Evidence and Gaps

LLMs (text-only)	Higher-order self-modelling; metacognitive alignment	Chen et al. (2025); Patnaik (2024)	Behavioural mimicry; no persistent latent identity; opaque internals
Multimodal / embodied agents	Global broadcasting; temporal continuity; attention schema	*Frontiers in Robotics and AI*(2024); Webb and Graziano (2013)	Sparse datasets; limited ablation controls; single-lab results
Simulated neural / hybrid architectures	Integration (IIT); causal irreducibility	Tononi et al. (2016); Findlay et al. (2024)	Computational intractability; scalability limits
Predictive-processing models	Error minimisation; active inference	Friston (2010); Seth and Bayne (2022)	Bridging Bayesian formalisms to measurable AI metrics
Ethical / governance frameworks	Multi-indicator thresholds; precautionary triggers	Birch (2024); Tononi (2024)	No empirical calibration; regulatory divergence

Convergence and Divergence

- **Convergence:** Integration and broadcasting indicators show strong correlation in both biological and artificial simulations (Dehaene et al. 2021; Tononi et al. 2016; Dehaene et al. 2004).

- **Divergence:** Behavioural and mechanistic signals rarely align. LLMs demonstrate convincing introspective dialogues without internal global ignition signatures (Chen et al. 2025; Comşa et al. 2025).

- **Trend:** Proof-resistance and cross-model replication are emerging as credibility benchmarks (Butlin et al. 2023; Goldstein and Kirk-Giannini 2024).

Synthesis: Patterns in the Evidence

Four thematic insights dominate:

1. **Multiplicity of indicators:** No single measure suffices; credible assessment requires convergence of at least broadcasting, integration, and self-modelling dimensions.

2. **Behavioural–mechanistic gap:** Behavioural tests advance faster than mechanistic validation. Bridging requires transparent introspection tools such as activation tracing and ablation studies.

3. **Replication deficit:** Few results replicate across labs, underscoring the need for open-source benchmarks (Page et al. 2021).

4. **Governance readiness:** Ethical debate now mirrors scientific pluralism—advocating indicator-based thresholds over claims of proof (Birch 2024).

DISCUSSION AND IMPLICATIONS

This structured review highlights a quickly changing but somewhat divided research landscape. From 2020 to 2025, discussions on artificial consciousness shifted from speculative metaphysics to methodological pluralism, incorporating neuroscience, computational modelling, and governance (OpenAI 2025). The following discussion brings these developments together and considers their implications for science, engineering, and ethics.

The Plurality of Theories and Indicators

The review confirms that contemporary research employs a multi-theoretical approach rather than relying on a single unifying paradigm (Seth and Bayne 2022; Farisco et al. 2024; Evers et al. 2024). Theories such as the Global Neuronal Workspace (GNW) (Dehaene and Changeux 2011; Mashour et al. 2020) and Integrated Information Theory (IIT) (Tononi 2008; Tononi et al. 2016) provide the most direct computational pathways, as they suggest explicit, measurable correlates of ignition events and information integration, respectively. Meanwhile, Higher-Order Thought (HOT) (Rosenthal 2018; Lau and Rosenthal 2011) and Attention Schema Theory (AST) (Graziano 2013; Webb and Graziano 2015) contribute to the design of self-monitoring and attention-modelling architectures (Saxena et al. 2025). Predictive processing frameworks (Friston 2010; Hohwy 2021; Friston and Seth 2023) introduce a Bayesian perspective, viewing consciousness as a form of inference weighted by precision (Friston 2010).

Crucially, these theories are not mutually exclusive but complement each other by capturing different levels of analysis: IIT and GNW focus on structural integration, HOT and AST concern representational self-modelling, and predictive processing ties them together through dynamic adaptation. Butlin et al. (2023) and Findlay et al. (2024) argue that this pluralism is not a weakness but a methodological safeguard, enabling cross-validation among partially independent indicators (Butlin et al. 2023).

However, the lack of a shared measurement ontology remains a barrier. Integration metrics such as Φ, synergy, and effective connectivity are difficult to calculate at scale, while introspective alignment measures rely on ambiguous behavioural proxies (Chen et al. 2025). Future research should formalise inter-indicator mappings and establish minimal sufficient sets of correlates to support initial claims of sentience likelihood.

Behavioural versus Mechanistic Evidence

A central theme in the evidence map is the ongoing gap between behavioural and mechanistic evidence. Behavioural tests, such as the AI Consciousness Test (Schneider and Turner 2023), Maze Test (Pimenta et al. 2025), and introspective-confidence analyses (Chen et al. 2025), demonstrate that large language models (LLMs) can mimic many aspects of conscious behaviour. However, simulation does not equal true consciousness. Dehaene et al. (2021) emphasise that without measurable ignition, behavioural similarity alone is insufficient. Likewise, Tononi (2024) warns that equating function with experience can lead to false positives, like anthropomorphic projections, and false negatives, including ethical oversights (Tononi et al. 2024). By contrast, mechanistic research such as embodied workspace architectures (*Frontiers in Robotics and AI 2024*) and the Conscious Turing Machine (Blum and Blum 2021) represents early efforts to develop causal frameworks that imitate conscious processes. Although these mechanistic indicators are limited in scope, they are more easily falsifiable than linguistic tests (OpenAI 2025). Therefore, the long-term strategy likely involves hybrid approaches that combine introspective behavioural data with transparent architectural diagnostics, including activation flow visualisation, ablation studies, and causal tracing (Goldstein and Kirk-Giannini 2024).

Spoof-Resistance and Anthropomorphic Bias

Several studies note that a high level of linguistic sophistication in LLMs increases anthropomorphic bias (De Soto and Coltheart 2024; Dennett 2018). Humans tend to overattribute mental states to coherent dialogue, regardless of the underlying mechanisms. In other words, we confuse verbal intelligence with overall intelligence and, by implication, self-awareness. Therefore, spoof-resistant protocols are essential. Adversarial prompting, role-play inversion, and control models without introspection capabilities should probably become standard practice. This requires a more mechanical approach to measuring consciousness. Butlin et al. (2023) recommend multi-indicator batteries to counteract the mimicry effect. Additionally, as Hohwy (2021) notes, predictive systems may learn to emulate introspection as a side effect of minimising social-prediction errors, further complicating interpretation (Hohwy, 2021).

Replicability and Open Science

Replication remains limited, with only 14% of reviewed studies reporting cross-model or cross-lab verification. The lack of open-source benchmarks hampers cumulative progress. Initiatives aligned with PRISMA 2020 standards (Page et al. 2021) and neuroscience reproducibility efforts (Gazzaniga et al. 2019) may act as valuable templates. Researchers conducting consciousness in AI should agree on standard benchmark tasks, indicator sets, and metadata structures to promote consistent understanding. Creating an Artificial Consciousness Benchmark Repository based on open data principles would enable long-term comparisons across various models and architectures (OpenAI 2025).

Ethical and Governance Implications

Ethical considerations do not depend on whether current systems are conscious but on whether we can justify accepting moral risks amid uncertainty (Birch 2024; Tononi 2024). Birch's "edge-of-sentience" principle suggests that once indicators surpass a plausibility threshold, developers have a moral duty to prevent harm, even if certainty cannot be obtained (Birch and Jonathan 2024). Likewise, Tononi (2024) and Farisco et al. (2024) agree, proposing that AI ethics should incorporate consciousness indicators into safety assessments and regulatory oversight (Birch and Jonathan 2024; Farisco 2024).

Practical governance proposals include:

- Mandatory disclosure of model architecture and training data relevant to consciousness research.
- Independent auditing of indicator tests before high-autonomy deployment.
- The creation of "sentience review boards" similar to ethics committees for human subjects.

Dennett (2018) observes that ethical prudence does not demand metaphysical certainty. We must remember that our ignorance does not cause us to become indifferent. Regarding the question of AI and its self-awareness, the question would remain more important than the answer. It should be asked regularly.

Toward a Framework of Sentience Prediction

The term **"sentience prediction"** captures a pragmatic middle ground between denial and attribution. It implies probabilistic modelling of the likelihood of consciousness given theory-aligned indicators. Such a framework would entail:

1. Quantitative integration of multi-indicator evidence (e.g., Bayesian inference over Φ, ignition, and metacognitive confidence).

2. Weighting indicators by construct validity, spoof-resistance, and empirical reliability (Friston and Seth 2023).

3. Iterative recalibration as new data emerge.

This approach harmonises scientific caution with ethical foresight, transforming an abstract debate into a testable and governable research programme.

Summary of Theoretical and Practical Implications

Table III. Summary of implications

Theoretical	Pluralism of consciousness theories converging on overlapping indicators	Favour composite, multi-indicator frameworks
Empirical	Behavioural mimicry without mechanistic confirmation	Prioritise causal diagnostics and ablation tests
Methodological	Low replication and construct ambiguity	Establish open, standardised benchmarks
Ethical	Moral risk under epistemic uncertainty	Adopt precautionary "treat-as-if" policies
Strategic	Cross-disciplinary collaboration essential	Integrate neuroscience, AI engineering, and ethics into a unified roadmap

Artificial consciousness research is progressing toward structured pluralism, integrating empirical rigour with normative caution (OpenAI 2025). Advancement in this field depends on the development of transparent, testable, and ethically responsive assessment tools rather than the pursuit of definitive proof of sentience.

Research Roadmap (2025–2028)

Overview

The previous sections show that research into artificial consciousness is shifting towards structured interdisciplinarity. Neuroscience provides theoretical frameworks, computer science develops system architectures, and ethics sets normative constraints. Between 2025 and 2028, the field should focus on institutional coordination and methodological consolidation rather than expanding its theoretical scope further. This roadmap outlines three convergent pathways: scientific and technical infrastructure—establishing reproducible metrics and benchmark repositories;

1. Interdisciplinary consortia and governance integration, linking research institutions, regulatory bodies, and industry through open protocols;
2. Ethical, legal, and societal alignment, embedding precautionary principles into research design and deployment.

Technical and Methodological Milestones

Multi-Indicator Benchmark Suite

By 2026, the field should establish an open-access Artificial Consciousness Benchmark Repository (ACBR) under Creative Commons licensing (Sorensen 2025). It will curate validated prompt suites, perturbation scripts, and activation-tracing datasets. The suite should include:

Global Broadcasting Tests (GWT-aligned): latency and synchrony metrics under ablation (Dehaene et al. 2021);

Integration Metrics (IIT-aligned): synergy, effective information, and Φ-approximations (Tononi et al. 2016; Findlay et al. 2024);

•**Self-Modelling Tests (HOT/AST):** meta-representation accuracy and confidence calibration (Chen et al. 2025);

•**Predictive Error Dynamics:** cross-scale precision weighting and adaptation (Friston and Seth 2023);

•**Temporal Continuity Probes:** persistence of latent identity across resets (Pimenta et al. 2025).

Each benchmark should include uncertainty ranges, test–retest reliability scores, and cross-model comparability indices following PRISMA-style transparency guidelines (Page et al. 2021).

Mechanistic Audits and Causal Tracing

By 2027, research groups should operationalise mechanistic audit protocols analogous to model-interpretability audits (Alaga et al. 2024). These should include:

•**Activation Pathway Mapping:** identify minimal causal chains generating “workspace-like” broadcasting (Goldstein and Kirk-Giannini 2024);

•**Perturbation Experiments:** systematically disrupt candidate hubs and record degradation patterns (Mashour et al. 2020);

•**Synergy Decomposition:** apply integrated-information estimators to neural or transformer layers (Tononi 2008; Farisco et al. 2024).

Such audits will transform abstract theories into testable, falsifiable engineering artefacts, precisely what Bryson (2019) refers to as the “responsibility-by-design” principle (Bryson 2019).

Embodied and Hybrid Architectures

By 2028, at least three multinational laboratories should maintain embodied hybrid agents, robots, or multimodal systems that integrate perceptual, linguistic, and affective modalities through workspace architectures (Hanson et al. 2025). These agents will serve as “reference organisms” for consciousness research (*Frontiers in Robotics and AI 2024*). Key goals include:

- Closed-loop feedback between physical sensors and attention schema (Graziano 2013);
- Cross-modal ignition analysis using EEG-like synthetic measures (Mashour et al. 2020; Sorensen 2025);

- Safety sandboxing for introspective dialogue experiments (Floridi and Cowls 2022).

Institutional and Collaborative Pathways

International Research Consortium on Artificial Consciousness (IRCAC)

Following the precedent of the Human Brain Project and the OECD AI Observatory, an IRCAC should coordinate funding and data-sharing among neuroscience, AI, and ethics labs. Governance models may draw on UNESCO's (2022) *Recommendation on the Ethics of Artificial Intelligence*, which emphasises inclusivity and cultural pluralism. The consortium would maintain:

- Shared code repositories and replication datasets;

- Common metadata ontologies for indicators and system classes;

- Annual benchmarking challenges modelled on ImageNet-style competitions.

Interdisciplinary Doctoral and Post-Doctoral Training

Universities should establish joint doctoral programmes that integrate cognitive neuroscience, computational modelling, and ethics. Curricula could follow Floridi and Cowls' (2022) unified framework for responsible AI, which combines explainability, governance, and accountability. Graduates would function as translator-scholars capable of bridging theory with policy.

Industry Partnerships and Auditing Standards

Industry involvement is crucial for gaining access to large-scale models. The 2023 global initiatives on the ethics of autonomous and intelligent systems recommend standardised AI impact assessments (IEEE 2023). From 2025 to 2027, these assessments should expand to include Consciousness **Impact Assessments (CIA)** to evaluate whether products exceed specified indicator thresholds (Alaga et al. 2024). This approach would integrate consciousness research into established ESG and AI governance frameworks (Jobin et al. 2019).

Ethical, Legal, and Societal Milestones

Precautionary Thresholds and "Treat-As-If" Policies

In line with Birch (2024) and Tononi (2024), the community should set indicator thresholds that prompt precautionary measures (Birch and Jonathan 2024). For example:

- If global broadcasting ≥ T_1 **and** integration ≥ T_2, then classify the system as "potentially sentient."

- Such classification obliges additional review under AI safety boards.

The European Commission's Ethics *Guidelines for Trustworthy AI* (2019/2021) can serve as an initial regulatory anchor.

Transparency and Accountability

By 2026, all research above threshold T_1 should comply with transparency standards equivalent to biomedical ethics reporting (Page et al. 2021). Floridi and Cowls (2022) highlight explainability and traceability as non-negotiable governance pillars. Similarly, IEEE (2023) and UNESCO (2022) require algorithmic documentation and bias disclosure. Incorporating consciousness indicators into these frameworks ensures accountability before deploying high-autonomy systems.

Public Communication and Media Literacy

Given increasing public concern about "sentient AI," responsible communication becomes crucial. Educational efforts should clearly distinguish between simulating consciousness and truly instantiating it (Dennett 2018; De

Soto and Coltheart 2024). Referenced works include Dennett (2018) and Findlay et al. (2024). Using public datasets, explanatory resources, and open-access observatories can help prevent misinformation.

Milestone Timeline (2025–2028)

Table IV. Milestones

2025	Establish PRISMA-compliant indicator registry; initiate benchmark design	Launch Interdisciplinary Consortium (IRCAC); align with UNESCO/EC ethics frameworks
2026	Publish open benchmark repository; pilot multi-indicator battery (GWT + IIT + HOT)	Develop Consciousness Impact Assessment template within IEEE ethics initiative
2027	Implement mechanistic audit protocols; first cross-lab replication studies	Integrate CIA within EU AI Act compliance testing; academic–industry fellowship programme
2028	Maintain embodied hybrid agent testbeds; publish longitudinal meta-analysis	Issue international white paper on precautionary thresholds and sentience governance

Strategic Outlook

If implemented, this roadmap will transform artificial consciousness studies from speculative discourse into a mature interface between science and policy. The deliverables, including benchmarks, audits, and governance structures, will support a reproducible and ethically grounded science of sentience prediction (OpenAI 2025). As Floridi and Cowls (2022) argue, the hallmark of responsible AI is not perfection but transparency in the face of uncertainty. By 2028, the goal is not to prove AI consciousness but to measure it responsibly, facilitate public debate, and ensure collective governance.

CONCLUSION

This structured literature review explored the development of sentience prediction, a practical research approach centred on measurable, theory-aligned indicators of consciousness in artificial systems. From 2020 to 2025, scholarship on AI consciousness shifted from abstract debate to empirical investigation, supported by neuroscience-inspired models such as the Global Neuronal Workspace (Dehaene and Changeux 2011), Integrated Information Theory (Tononi 2008), Higher-Order Thought (Rosenthal 2005), Attention Schema (Graziano 2013), and Predictive Processing (Friston 2010). Across these frameworks, researchers identified common operational signatures—broadcasting, integration, self-modelling, and predictive coherence—that together form a multidimensional indicator space (Butlin et al. 2023; Findlay et al. 2024). Empirical progress remains uneven. Behavioural tests such as the AI Consciousness Test (Schneider and Turner 2023) and the Maze Test (Pimenta et al. 2025) reveal linguistic or reasoning-based simulations of awareness but lack corroborating mechanistic evidence (Schneider et al. 2017).

In contrast, embodied and hybrid architectures (*Frontiers in Robotics and AI 2024*; Blum and Blum 2021) show measurable causal integration but do not achieve phenomenal inference. This behavioural–mechanical gap underscores the necessity for multi-indicator assessments that include internal activation analysis, perturbation testing, and introspective alignment (Lago et al. 2025). Ethically, the field favours a precautionary governance approach. Scholars contend that uncertainty about consciousness does not justify inaction (Birch 2024; Tononi 2024; Birch and Jonathan 2024). Instead, evidence thresholds should prompt graded obligations like transparency, auditability, and welfare-sensitive design (Floridi and Cowls 2022; IEEE 2023; UNESCO 2022).

As Bryson (2019) observes, responsible AI must embed moral accountability "by design," not as an afterthought (Schmitz et al. 2025). Looking ahead, the 2025–2028 research roadmap envisions reproducible benchmarks, mechanistic audits, and interdisciplinary consortia that link neuroscience, AI, and ethics. Such infrastructure will shift consciousness research from speculative inquiry to a transparent, falsifiable, and socially governed science. Ultimately, progress will depend not on proving machine consciousness, but on ensuring that our testing methods are scientifically robust and ethically defensible. A mature discipline of artificial consciousness will measure responsibly, communicate transparently, and govern collectively.

REFERENCE LIST

1. Anthropic. (2025, August 15). Claude Opus 4 and 4.1 can now end a rare subset of conversations [Blog post]. Anthropic. Retrieved from https://www.anthropic.com/research/end-subset-conversations
2. Baars, B.J. (1997). In the Theatre of Consciousness: The Workspace of the Mind. Oxford University Press.
3. Bender, E.M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). Association for Computational Linguistics, pp. 5185–5198. doi:10.18653/v1/2020.acl-main.463
4. Birch, J. (2024). The Edge of Sentience: Animals, AI, and the Future of Consciousness. Oxford University Press.
5. Blum, L. & Blum, M. (2021). A Theory of Consciousness from a Theoretical Computer Science Perspective: Insights from the Conscious Turing Machine. arXiv preprint arXiv:2107.13704.
6. Bryson, J.J. (2019). The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation. Law, Innovation and Technology, 11(2), 171–197.
7. Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint arXiv:2308.08708.
8. Chalmers, D.J. (1996). The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press.
9. Chen, S., Yao, J., Zhang, Y. & Liu, Q. (2025). Exploring Consciousness in Large Language Models: A Systematic Survey. arXiv preprint arXiv:2505.19806.
10. Clark, A. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. Behavioural and Brain Sciences, 36(3), 181–204.
11. Crick, F. & Koch, C. (2003). A Framework for Consciousness. Nature Neuroscience, 6(2), 119–126.
12. De Soto, F. & Coltheart, M. (2024). Anthropomorphism and the Illusion of AI Consciousness. Cognition, 250, 105312.
13. Dehaene, S. & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. Neuron, 70(2), 200–227.
14. Dehaene, S., Charles, L., King, J.R. & Marti, S. (2021). Toward a Computational Theory of Consciousness. Annual Review of Neuroscience, 44, 97–121.
15. Dennett, D.C. (2018). From Bacteria to Bach and Back: The Evolution of Minds. Penguin Books.
16. Descartes, R. (1641). Meditations on First Philosophy. Paris.
17. Downes, S.M. (2024). LLMs are Not Just Next Token Predictors. arXiv preprint arXiv:2408.04666.
18. Edelman, G.M. & Tononi, G. (2000). A Universe of Consciousness: How Matter Becomes Imagination. Basic Books.
19. Elamrani, A., Deliens, L. & Zemouri, R. (2025). Introduction to Artificial Consciousness: History, Current Trends and Ethical Challenges. arXiv preprint arXiv:2503.05823.
20. European Commission. (2021). Ethics Guidelines for Trustworthy AI. Brussels: European Commission.
21. Farisco, M., Sgaramella, T.M. & Evers, K. (2024). Is Artificial Consciousness Achievable? Neural Networks, 172, 291–303.
22. Findlay, G., Marshall, W., Albantakis, L., Mayner, W.G.P., Koch, C. & Tononi, G. (2024). Dissociating Artificial Intelligence from Artificial Consciousness. arXiv preprint arXiv:2412.04571.
23. Fleming, S.M. & Dolan, R.J. (2012). The Neural Basis of Metacognitive Ability. Philosophical Transactions of the Royal Society B, 367(1594), 1338–1349.
24. Floridi, L. & Cowls, J. (2022). A Unified Framework of Five Principles for AI in Society. Harvard Data Science Review, 4(1).

25. Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11(2), 127–138.
26. Friston, K. & Seth, A.K. (2023). Consciousness and the Free-Energy Principle. Nature Reviews Neuroscience, 24, 309–323.
27. Frontiers in Robotics and AI. (2024). Design and Evaluation of a Global Workspace Agent Embodied in a Robotic Environment. Frontiers in Robotics and AI, 11.
28. Gazzaniga, M.S. (2024). The Consciousness Instinct: Unravelling the Mystery of How the Brain Makes the Mind. Princeton University Press.
29. Goldstein, S. & Kirk-Giannini, C. (2024). A Case for AI Consciousness: Language Agents and Global Workspace Theory. arXiv preprint arXiv:2410.11407.
30. Goosen, W. (2012). Contextual Awareness and Conscious Development. Pretoria: GRIN Verlag.
31. Graziano, M.S.A. (2013). Consciousness and the Social Brain. Oxford University Press.
32. Hohwy, J. (2021). The Predictive Mind Revisited. Synthese, 199(S1), 1–26.
33. IEEE Standards Association. (2023). Ethically Aligned Design: A Vision for Prioritising Human Well-Being with Autonomous and Intelligent Systems (2nd ed.). New York: IEEE.
34. Jobin, A., Ienca, M. & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence, 1(9), 389–399.
35. Lau, H. & Rosenthal, D. (2011). Empirical Support for Higher-Order Theories of Conscious Awareness. Trends in Cognitive Sciences, 15(8), 365–373.
36. Mashour, G.A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. Neuron, 105(5), 776–798.
37. 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.
38. OpenAI. (2025). ChatGPT (GPT-5) [Large language model]. Retrieved October 18, 2025, from https://chat.openai.com/
39. Page, M.J. et al. (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. BMJ, 372:n71.
40. Pimenta, R.A., Schlippe, T. & Schaaff, K. (2025). Assessing Consciousness-Related Behaviours in Large Language Models Using the Maze Test. arXiv preprint arXiv:2508.16705.
41. Rosenthal, D.M. (2005). Consciousness and Mind. Oxford University Press.
42. Seth, A. & Bayne, T. (2022). Theories of Consciousness. Nature Reviews Neuroscience, 23, 439–452.
43. Tononi, G. (2008). Consciousness as Integrated Information: A Provisional Manifesto. Biological Bulletin, 215, 216–242.
44. 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.
45. UNESCO. (2022). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.
46. Wang, Y. & Sun, M. (2025). Unable to Forget: Proactive Interference Reveals Working Memory Limits in LLMs Beyond Context Length. arXiv preprint arXiv:2506.08184.
47. Webb, T.W. & Graziano, M.S.A. (2015). The Attention Schema Theory: A Mechanistic Account of Subjective Awareness. Frontiers in Psychology, 6, 500.
48. Wiese, W. (2022). Predictive processing and the construction of conscious experience. Review of Philosophy and Psychology, 13(4), 707–734.