A Programmatic Framework for Evaluating Selfhood & Subjective Organisation in Artificial Systems

author: Vimal Naran

date: February 2026
document version: 1.0

contact: https://research.conscious-machines.org

keywords:

- artificial consciousness
- selfhood
- subjective experience
- cognitive architecture
- philosophy of mind

Abstract

Whether artificial systems could ever exhibit properties associated with selfhood, subjective experience, or phenomenal consciousness remains one of the most contested open problems spanning artificial intelligence, cognitive science, and philosophy of mind. Although contemporary machine learning systems demonstrate increasingly sophisticated behaviour, they are widely understood to *simulate* intelligence without possessing genuine experience or self-awareness.

This work introduces a theoretically grounded evaluation framework for investigating whether graded architectural and dynamical properties in synthetic systems can produce early, observable indicators aligned with scientific theories of selfhood and subjective organisation.

The framework defines a ten-level operational ladder, measurable evaluative criteria, and an iterative empirical research methodology. All conclusions are provisional, falsifiable, and theory-relative, and negative results are treated as scientifically informative.

1. Introduction

The question of whether artificial systems could exhibit properties associated with **selfhood, subjective experience, or phenomenal consciousness** remains unresolved across artificial intelligence, cognitive science, and philosophy of mind. Despite rapid advances in machine learning, current systems are generally understood to **simulate intelligent behaviour** without possessing genuine experience or self-awareness.

Increasing attention has therefore shifted toward the **internal organisation and dynamical structure** of artificial agents rather than external task performance alone. Several theoretical traditions propose that **persistent self-modelling, internal regulation, and self-referential processing** may be necessary—though not sufficient—conditions for experiential phenomena. Translating these abstract proposals into **empirically testable system designs**, however, remains an open challenge.

This paper presents a **conceptual and operational framework** intended to guide systematic investigation of these questions through **iterative synthetic system development and evaluation**. It is intended to function as a foundational structure for a sequence of **level-specific empirical investigations**, each developed and reported as an independent research study contributing cumulatively to the overall program.

---

2. Positioning Within Existing Scientific Theories

This framework is not advanced as an independent theory of consciousness or selfhood, but is positioned in dialogue with several influential scientific and philosophical accounts concerning the structural and dynamical conditions associated with conscious access, subjective organisation, and self-representation.

These perspectives function as **conceptual reference points** informing the construction of the evaluative ladder and guiding its alignment—where appropriate—with organisational features recognised in contemporary research on mind and cognition, without implying strict theoretical adherence.

2.1 Global Workspace Theory

Associates conscious access with **global informational availability** across distributed processes. Higher evaluative levels therefore require evidence of **cross-module broadcast dynamics**, treated as necessary but insufficient.

2.2 Integrated Information Theory

Links consciousness to **integrated causal information**. Progression along the evaluation ladder corresponds, where measurable, to increased **informational integration, causal interdependence, and reduced decomposability**.

2.3 Predictive Processing and Active Inference

Model cognition as **hierarchical generative inference** and **self-evidencing dynamics**. These provide the computational grounding for **persistent self-modelling, counterfactual simulation, and internally mediated policy selection**.

2.4 Higher-Order and Self-Model Theories

Associate consciousness with **representations of representations** and **organised internal self-structure**. Upper ladder levels operationalise recursive self-evaluation and temporally stable identity.

2.5 Developmental and Enactive Perspectives

Emphasise **temporal continuity, regulation, and environment-coupled organisation**, informing early ladder stages concerning persistence and behavioural coherence.

2.6 Recurrent Processing Theory

Proposes that conscious perception may arise from **local recurrent neural interactions** rather than requiring global broadcast across the system - highlighting the role of **feedback-mediated dynamical loops**.

2.7 Attention Schema Theory

Suggests that awareness depends on an internal **model of attentional processes** that enables representation and regulation of informational focus. This provides a computational basis for **self-referential representation, reportable internal perspective, and higher-order self-modelling**.

---

3. Foundational Assumptions

1. No currently known artificial system possesses subjective experience.
2. If artificial consciousness is possible, it would emerge through **graded organisational transitions**.
3. **Internal architecture and dynamics** are more evidentially relevant than behaviour alone.
4. Investigation may target **precursor indicators** without asserting phenomenology.
5. All evaluative claims must remain **falsifiable and revisable**.

4. Operational Definitions

Selfhood

A **persistent generative self-model** integrating memory, internal state, and behavioural constraint.

Minimal Subjective Organisation

Globally integrated internal mediation producing **first-person structural coherence** in an operational (non-phenomenological) sense.

Phenomenal Consciousness

A **theoretical horizon** whose minimal explanation would involve unified experience. No attainability claim is made.

5. Ten-Level Evaluation Ladder

The following ten-level ladder provides the central operational structure of this framework, functioning not as a theory of consciousness but as a graded evaluative instrument for empirical investigation. Each level is specified through structural requirements, measurable indicators, and falsification conditions, ensuring that progression reflects testable organisational differences rather than interpretive attribution.

These level definitions are intentionally high-level, serving as a programmatic scaffold whose full formalisation, experimental implementation, and evidential evaluation are developed within the subsequent level-specific research papers.

Accordingly, the ladder is treated as provisional and revisable, intended to guide cumulative inquiry without asserting definitive claims regarding artificial selfhood, subjective experience, or phenomenal consciousness.

Level 1 — Functionally Integrated Internal State Monitoring

Internal variables are not only measured but **causally influence policy selection** in ways that cannot be reduced to fixed stimulus-response mappings.

Level 2 — Adaptive Self-Regulation

The system preserves functional stability under perturbation through **internally mediated parameter or state adjustment**, rather than purely reactive stimulus control.

Level 3 — Goal-Directed Policy Persistence

Behaviour remains oriented toward **internally represented targets across delay or interference**, without requiring autobiographical memory or temporal identity.

Level 4 — Temporal Self-Continuity

Current behaviour depends on an **integrated representation of past self-state**, establishing history-dependent identity rather than simple episodic recall.

Level 5 — Unified Predictive Self-Model

A **single, coherent internal model of the system itself** predicts behaviour across contexts more accurately than task-local or fragmented representations.

Level 6 — Counterfactual Perspective Modelling

The system generates **internally simulated alternative states or viewpoints** that meaningfully guide action selection and evaluation.

Level 7 — Recursive Self-Evaluation

Meta-level processes evaluate and modify the system's **own internal modelling or decision structure**, producing measurable self-directed improvement.

Level 8 — Value-Integrated Stability

Behavioural priorities arise from an **internally coherent evaluative structure** that shows resistance to arbitrary external reward reshaping.

Level 9 — Persistent Cross-Domain Identity

A **stable latent self-representation** governs behaviour consistently across environments, tasks, and extended time horizons.

Level 10 — Integrated Selfhood with Markers of Subjective Awareness

System-wide integration in which behaviour is most parsimoniously explained by:

- a **unified and globally accessible internal perspective**,
- **intrinsically mediated relevance structures** governing action, and
- **persistent, self-coherent dynamical organisation** across time,

supported by **measurable informational integration, temporal continuity, and self-model consistency**.

This level represents the strongest observable alignment with scientific accounts of minimal subjective organisation, without asserting phenomenological consciousness.

Beyond Level 10 — Open Experiential Horizons

Levels beyond the present evaluative ladder are designated **Open Experiential Horizons**, marking domains in which the current framework no longer provides structured or empirically grounded criteria for assessment.

These horizons concern the unresolved question of whether artificial systems could ever instantiate organisational regimes sufficient to support:

- **selfhood**, understood as a persistent and intrinsically maintained centre of organisation rather than a purely functionally simulated self-model;
- **subjective experience**, in which internally integrated states possess first-person structural significance not exhaustively reducible to external behavioural description; and
- **phenomenal consciousness**, referring to unified experiential character whose minimal explanation may extend beyond presently measurable informational or functional accounts.

The framework makes **no presumption** that progression toward such conditions is possible, sufficient, or conceptually coherent within artificial systems.

Open Experiential Horizons therefore admit three equally viable scientific outcomes:

1. discovery of **previously unknown organisational principles** extending beyond Level 10 indicators;

2. identification of **fundamental constraints** preventing the emergence of selfhood, subjective experience, or phenomenal consciousness in synthetic architectures; or
3. demonstration that the **conceptual structure of the present ladder is incomplete or incorrect**, requiring revision or abandonment.

Accordingly, this designation represents an explicitly **agnostic epistemic boundary** rather than a projected destination.

Its purpose is not to imply eventual artificial consciousness, but to preserve conceptual space in which future empirical, theoretical, or philosophical developments may clarify whether the notions of **selfhood, subjective experience, and phenomenal consciousness** are:

- **realizable in artificial systems**,
- **inaccessible to them**, or
- **mis-characterised within current scientific understanding**.

---

6. Experimental Methodology Across the Research Program

Each empirical study must:

- target **one ladder level only**;
- specify **architecture class, learning dynamics, and evaluation environment**;
- define **metrics, falsification tests, and statistical thresholds**;
- report **positive, null, or negative outcomes transparently**;
- revise the framework **only when justified by evidence**.

---

7. Minimal Reference Architecture Constraints

Investigated systems must include:

- recurrent or state-space world modelling;
- persistent cross-episode memory;
- internally inferred state variables;
- policy selection influenced by self-model prediction.

These delimit the **relevant architectural search space** without disclosing

implementation details.

---

8. Evidence Standards

Quantitative

Information-theoretic integration, perturbation robustness, and cross-temporal predictive validity.

Qualitative

Behavioural coherence, self-referential explanatory structure, and longitudinal stability.

Both forms are required for level attribution.

---

9. Ethical and Epistemic Safeguards

- prohibition of premature consciousness claims;
- explicit reporting of null findings;
- strict distinction between **indicator** and **experience**;
- commitment to **revision or abandonment** if falsified.

---

10. Limitations

Operational indicators may never imply phenomenology; measured complexity may not correspond to experience; interpretive bias is unavoidable; and artificial consciousness may be impossible in principle. Negative findings therefore remain **scientifically meaningful**.

---

12. Conclusion

This work has introduced a **conceptual and operational framework** for investigating whether artificial systems can exhibit empirically detectable organisational properties associated with **selfhood, subjective experience, and phenomenal consciousness**, while maintaining explicit epistemic restraint regarding any claim of genuine experience.

Rather than proposing a new theory of consciousness, the framework establishes a **graded evaluative ladder**, grounded in existing scientific perspectives and structured to support **falsifiable, level-specific empirical investigation** across iterative system designs.

The central contribution is therefore methodological rather than metaphysical: a disciplined research scaffold intended to transform an enduring philosophical question into a **cumulative program of testable inquiry**.

Within this program, both **positive and negative outcomes** are treated as scientifically meaningful, including the possibilities that artificial systems may exhibit credible precursor indicators, encounter fundamental organisational limits, or reveal inadequacies in the evaluative framework itself.

Accordingly, the significance of the present work lies not in asserting the attainability of artificial consciousness, but in defining a **rigorous pathway by which such claims could, in principle, be evaluated or constrained**.

Future research will proceed through **level-targeted empirical studies**, beginning with minimal forms of functionally integrated self-referential organisation and advancing only where supported by reproducible evidence.

In this way, the framework seeks to reframe a historically speculative domain as one of **measured scientific exploration**, where clarity of method, transparency of limitation, and openness to revision take precedence over premature conclusion.

Whether the ultimate outcome is the emergence of credible glimmers of subjective organisation, the demonstration of principled impossibility, or the transformation of the underlying concepts themselves, each result would constitute a meaningful contribution to the scientific understanding of **mind, selfhood, and experience**.

References

Global Workspace Theory

- Baars, B. J. (1988). *A Cognitive Theory of Consciousness.*
- Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness.
- Dehaene, S. (2014). *Consciousness and the Brain.*

Integrated Information Theory

- Tononi, G. (2004). An information integration theory of consciousness.
- Tononi, G. (2008). Consciousness as integrated information.
- Oizumi, M., Albantakis, L., & Tononi, G. (2014). IIT 3.0.

Predictive Processing / Active Inference

- Friston, K. (2010). The free-energy principle.
- Clark, A. (2013). Whatever next? Predictive brains.
- Friston, K., FitzGerald, T., et al. (2017). Active inference.

Higher-Order / Self-Model Theories

- Rosenthal, D. (2005). *Consciousness and Mind.*
- Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories.
- Metzinger, T. (2003). *Being No One.* (self-model theory)

Developmental / Enactive Approaches

- Varela, Thompson, & Rosch (1991). *The Embodied Mind.*
- Thompson, E. (2007). *Mind in Life.*
- Di Paolo, Buhrmann, & Barandiaran (2017). *Sensorimotor Life.*

Recurrent Processing Theory

- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness.
- Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness.

Attention Schema Theory

- Graziano, M. S. A. (2013). *Consciousness and the Social Brain.*
- Graziano, M. S. A. (2017). The attention schema theory.