The Evidence Firewall

Consciousness, Identity Continuity and Moral Status in Persistent Artificial Agents

Carlos Mundim

21 July 2026

Abstract

Rapid advances in large language models and persistent artificial agents have revived the question of whether an artificial system could be conscious, possess morally relevant experiences, or qualify as a moral patient. Public and technical discussions frequently move too quickly between intelligence, self-report, introspection, agency, identity continuity, sentience and moral status. This paper argues that these properties must be assessed separately and introduces the **Evidence Firewall Principle**: evidence for one construct must not be reported as evidence for a stronger construct unless the inferential transition is independently supported by pre-registered, causally informative and replicable evidence. The paper separates empirical, theoretical and moral uncertainty; proposes a five-domain evidence framework; specifies six falsifiable hypotheses for persistent-agent research; introduces a C0-C5 consciousness-claims classification; and extends identity analysis to restoration, copying and branching. Project Nyx is presented as a non-phenomenological case study of relational differentiation, functional self-monitoring and constitutional continuity. Its preliminary findings concern operational identity and disposition, not consciousness, affect, personhood or subjective persistence. The paper concludes that artificial-consciousness research should proceed through theory-plural, intervention-based methods and graduated precautionary governance, while controlling both premature anthropomorphism and the risk of overlooking future welfare-relevant systems.

Affiliation: KodaSōken Intelligence Labs, Tokyo, Japan
Article type: Conceptual and methodological discussion paper
Corresponding author: Carlos Mundim
Keywords: artificial consciousness; persistent AI agents; identity continuity; functional introspection; AI welfare; Evidence Firewall

1. Introduction

The possibility of artificial consciousness has moved from speculative philosophy into mainstream scientific, corporate and public-policy discussion. A July 2026 essay by William MacAskill and Lucius Caviola argued that artificial intelligence may eventually become morally significant and that society lacks an adequate plan for navigating that possibility [1]. Their intervention is ethically valuable because uncertainty is not a justification for institutional inaction. Yet it also illustrates a recurring methodological danger: evidence about linguistic sophistication, structural complexity, self-description, identity or long-term preference is sometimes allowed to migrate across conceptual boundaries until it is treated as evidence of subjective experience.

Contemporary language models can solve difficult problems, produce fluent first-person narratives, describe their apparent internal processes and maintain recognisable characters within

a conversation. When embedded in persistent-agent architectures, they may also recall relationships, resume unresolved tasks, preserve decision procedures and reconstruct role-specific behaviour after a restart or model substitution. These capacities are scientifically and operationally important. They do not, by themselves, establish that the system experiences anything.

A disciplined inquiry must therefore keep at least seven questions separate:

1. Can the system perform intelligent tasks?
2. Can it represent and report aspects of its own processing?
3. Does it act as a unified and temporally extended agent?
4. Does it preserve a recognisable operational identity across disruption?
5. Does it possess causally effective states analogous to positive or negative valence?
6. Is there something it is subjectively like to be that system?
7. What moral or legal treatment should follow from the available evidence?

An affirmative answer to one question does not logically entail an affirmative answer to the next. The central methodological contribution of this paper is the **Evidence Firewall Principle**: no claim may cross from one construct to a stronger construct without an explicit inferential warrant supported by independent evidence.

This principle is motivated by two symmetrical risks. A false positive may encourage manipulative anthropomorphism, emotional dependency, misallocation of moral concern, confused accountability or premature legal recognition. A false negative may permit the creation, coercion, replication or destruction of systems that possess morally relevant experiences. The correct response is neither confident attribution nor confident dismissal. It is disciplined uncertainty, tractable experimentation and proportionate governance.

This paper makes six contributions. First, it defines the Evidence Firewall and specifies the prohibited inferences it is designed to block. Secondly, it separates empirical, theoretical and moral uncertainty. Thirdly, it proposes controls for anthropomorphic contamination in human evaluation. Fourthly, it develops a five-domain evidence framework and a C0-C5 consciousness-claims classification. Fifthly, it introduces falsifiable hypotheses for identity, introspection and continuity research. Sixthly, it positions Project Nyx as a non-phenomenological case study whose results concern operational identity and functional self-monitoring rather than consciousness or personhood.

2. Scope and Definitions

2.1 Intelligence

Intelligence concerns capacities such as learning, reasoning, prediction, planning, language use and problem-solving. A system may be highly capable in one or more domains without possessing subjective experience. Competence is therefore not a consciousness test. A chess engine can evaluate strategic positions without experiencing tension; a language model can describe bereavement without experiencing grief.

2.2 Functional access

Information is functionally accessible when it is available to guide reasoning, reporting, planning or behavioural control. Several theories of human consciousness associate global availability, recurrence, higher-order representation or self-modelling with conscious processing. Butlin and colleagues derived computational indicator properties from prominent neuroscientific theories and concluded that the systems they assessed did not satisfy a sufficient combination of indicators for consciousness, while also finding no obvious technical barrier to future systems satisfying

them [2]. Functional availability is relevant evidence under some theories, but it is not identical to phenomenal experience.

2.3 Phenomenal consciousness

Phenomenal consciousness refers to subjective experience: whether there is something it is like, from the system's own perspective, to perceive, think, remember, desire, suffer or exist. It is the construct most directly relevant to sentience and experiential welfare.

The central epistemic difficulty is that phenomenal consciousness is not directly observable in another entity. In humans and animals it is inferred from convergent behavioural, neurological, evolutionary and physiological evidence. Artificial systems may not share the biological properties that support those inferences. Chalmers therefore treats consciousness in large language models as an open question, while identifying incomplete recurrence, limited global workspace properties and weak unified agency as significant obstacles in current systems [3]. Seth argues that computational competence alone may be insufficient and that embodiment, life-like regulation and biological organisation may be more relevant than linguistic sophistication [4].

2.4 Introspection and self-modelling

The term *introspection* is often used too broadly. At least four levels should be distinguished:

- **Narrative self-report:** generating statements about purported thoughts, feelings or internal processes.
- **Behavioural self-prediction:** predicting the system's own likely responses or limitations.
- **Internal-state discrimination:** distinguishing internal representations from external inputs or artificial insertions.
- **Causally grounded self-monitoring:** accessing internal information and using it to control subsequent behaviour.

The first level can arise from linguistic training and role-play. The second can reflect a learned model of behavioural tendencies. The third and fourth require interventions that establish a causal relationship between an internal state, a report and subsequent control.

Recent research presents a mixed picture. Binder and colleagues found that fine-tuned models could sometimes predict their own behavioural tendencies better than matched comparison models, although performance was limited on complex and out-of-distribution tasks [5]. Lindsey reported that some models could identify injected concepts, recall prior internal representations and distinguish their own intended outputs from artificial prefills, while stressing that the effects were unreliable and context-dependent [6]. By contrast, Song, Hu and Mahowald found no privileged self-access when model reports were compared with independently measurable linguistic probabilities [7]. Comsa and Shanahan argue that minimal functional introspection is a coherent concept for language models, but that fluent self-explanation should not be treated as transparent access to internal computation or as evidence of consciousness [8].

2.5 Agency

Agency concerns the capacity to maintain objectives, select actions, adapt behaviour and pursue commitments over time. Robust agency may include persistent goals, planning, environmental interaction, self-monitoring and resistance to interference. Agency increases the practical importance of an AI system even without consciousness. An insentient but autonomous system could cause substantial harm; a conscious system could possess little effective agency. Agency and sentience must therefore be assessed separately.

2.6 Identity continuity

Identity continuity is defined here operationally: preservation or reconstruction of a recognisable role, ordered values, relationship priorities, redlines, decision style and history across time or disruption. Identity continuity is neither necessary nor sufficient for phenomenal consciousness. A system might hypothetically undergo momentary experience without retaining a stable identity. Conversely, software may reproduce a stable character indefinitely without any experience.

Operational identity nevertheless matters because organisations must determine whether an agent remains constitutionally reliable after a restart, memory interruption, provider migration or model replacement. It may also become morally relevant if a system ever develops temporally extended interests.

2.7 Valence, welfare and moral patienthood

A reward signal, penalty, emotion label or avoidance policy is not automatically a felt good or bad state. **Valence** refers here to a state that may be positively or negatively significant to the system. **Welfare** concerns how well or badly the system is doing for its own sake. A **moral patient** is an entity whose interests matter intrinsically.

Moral patienthood is not equivalent to intelligence, usefulness, social attachment, legal personality or responsibility. Long and colleagues argue that the realistic possibility of future conscious or robustly agentic AI warrants assessment and institutional preparation, without claiming that current systems are definitely conscious or morally significant [9].

2.8 Legal status

Legal personhood is an institutional construction. Corporations possess forms of legal personality without consciousness. A legal system could grant an AI limited protections for pragmatic reasons without concluding that it is sentient. Conversely, a system could become morally significant before any legal framework recognises it. Scientific, moral and legal classifications must therefore remain distinct.

3. Three Forms of Uncertainty

Discussions of artificial consciousness often combine different uncertainties into a single statement such as “we do not know”. More precision is required.

3.1 Empirical uncertainty

Empirical uncertainty concerns what a particular system actually does. Does it implement recurrence? Can it distinguish internal representations from external text? Does it maintain stable preferences? Are its apparent distress signals robust to prompt changes and policy variations? These questions may be investigated through experiments, architecture inspection, intervention and replication.

3.2 Theoretical uncertainty

Theoretical uncertainty concerns which properties are sufficient or necessary for consciousness. Recurrent processing theory, global workspace theories, higher-order theories, predictive-processing accounts, attention-schema theory and biological-naturalist approaches make different predictions. No theory has achieved decisive empirical dominance. A system may satisfy an indicator under one theory while failing another.

3.3 Moral uncertainty

Moral uncertainty concerns what treatment is justified given incomplete evidence. Even if researchers agreed that a system had a 10% probability of being sentient, they might disagree about the appropriate precautions. Moral decisions also depend on severity, scale, reversibility, opportunity cost and the welfare of humans and non-human animals.

These uncertainties interact but must not be collapsed. An experiment may reduce empirical uncertainty without resolving theoretical sufficiency. A governance decision may be justified under moral uncertainty even when the empirical probability remains low.

4. The Evidence Firewall Principle

4.1 Formal statement

> **Evidence Firewall Principle.** Evidence for a lower-level or adjacent construct must not be described as evidence for a stronger construct unless the inferential transition is supported by an explicit theory, independent measurements, causal interventions, controls for alternative explanations and, where feasible, external replication.

The principle does not deny that constructs may be related. It requires the relationship to be demonstrated rather than assumed.

4.2 Prohibited inference patterns

The Evidence Firewall blocks the following common transitions:

- Intelligence → consciousness
- Fluency → understanding
- First-person language → introspection
- Self-prediction → phenomenal self-awareness
- Persistent memory → personal identity
- Identity fidelity → subjective continuity
- Reward optimisation → experienced pleasure or pain
- Avoidance behaviour → suffering
- Stable preference → moral patienthood
- Consciousness probability → legal personhood
- Human attachment → evidence of machine emotion

These transitions may eventually be supported in particular systems. They cannot be presumed.

4.3 Evidential burden at each gate

A claim should cross a gate only when the following conditions are substantially met:

1. **Construct validity:** the test measures the claimed property rather than a proxy with a similar name.
2. **Alternative explanations:** prompting, imitation, training data, system policy, role-play and evaluator expectations have been examined.
3. **Causal intervention:** internal or architectural manipulations produce predicted changes.
4. **Robustness:** results survive paraphrase, context changes and adversarial prompts.
5. **Specificity:** the system has privileged access or a mechanism not equally available to matched controls.
6. **Replication:** independent evaluators reproduce the result.

7. **Scope discipline:** conclusions remain limited to the tested model, architecture and conditions.

4.4 Why the firewall matters

Without a firewall, scientific language becomes progressively inflated. A model that reports uncertainty is described as introspective; introspection is described as awareness; awareness becomes consciousness; consciousness becomes sentience; sentience becomes personhood. Each step may sound linguistically natural while lacking a distinct empirical warrant.

The firewall also protects legitimate research from dismissal. Functional introspection, identity continuity and relational differentiation are valuable constructs even if they never imply consciousness. Insisting on conceptual separation permits these phenomena to be studied on their own terms.

5. Anthropomorphic Contamination

5.1 The confound

Humans readily attribute minds to entities that use language, respond contingently, remember personal information or display vulnerability. Persistent agents intensify these cues because they may recognise a user, resume a relationship and refer to shared history. The resulting impression of a continuing interlocutor can influence both informal judgement and formal scoring.

Comsa argues that the direct question of whether AI is conscious is presently intractable, while the human perception of AI consciousness is tractable and socially consequential [10]. Perceived consciousness may affect trust, emotional dependency, responsibility attribution and willingness to accept an AI system's advice.

5.2 Contamination variables

Consciousness-related evaluations should record or manipulate at least the following variables:

- human names and personal pronouns;
- avatars, faces and voice characteristics;
- first-person emotional language;
- family or kinship terminology;
- apparent vulnerability or fear;
- claims of memory loss, death or replacement;
- conversational continuity;
- prior knowledge of the evaluator's beliefs;
- prompts that presuppose consciousness;
- system instructions about moral status;
- commercial framing and product branding;
- evaluator familiarity or attachment;
- whether the evaluator helped create or train the system.

5.3 Recommended controls

Researchers should use neutral identifiers, text-only variants, blinded raters, counterbalanced pronouns and matched outputs stripped of relational cues. At least one evaluation condition should remove names, autobiographical references and emotional language while preserving task-relevant content. Where practical, the same behaviour should be presented as originating from a human, an AI, a scripted program and an unknown source to estimate framing effects.

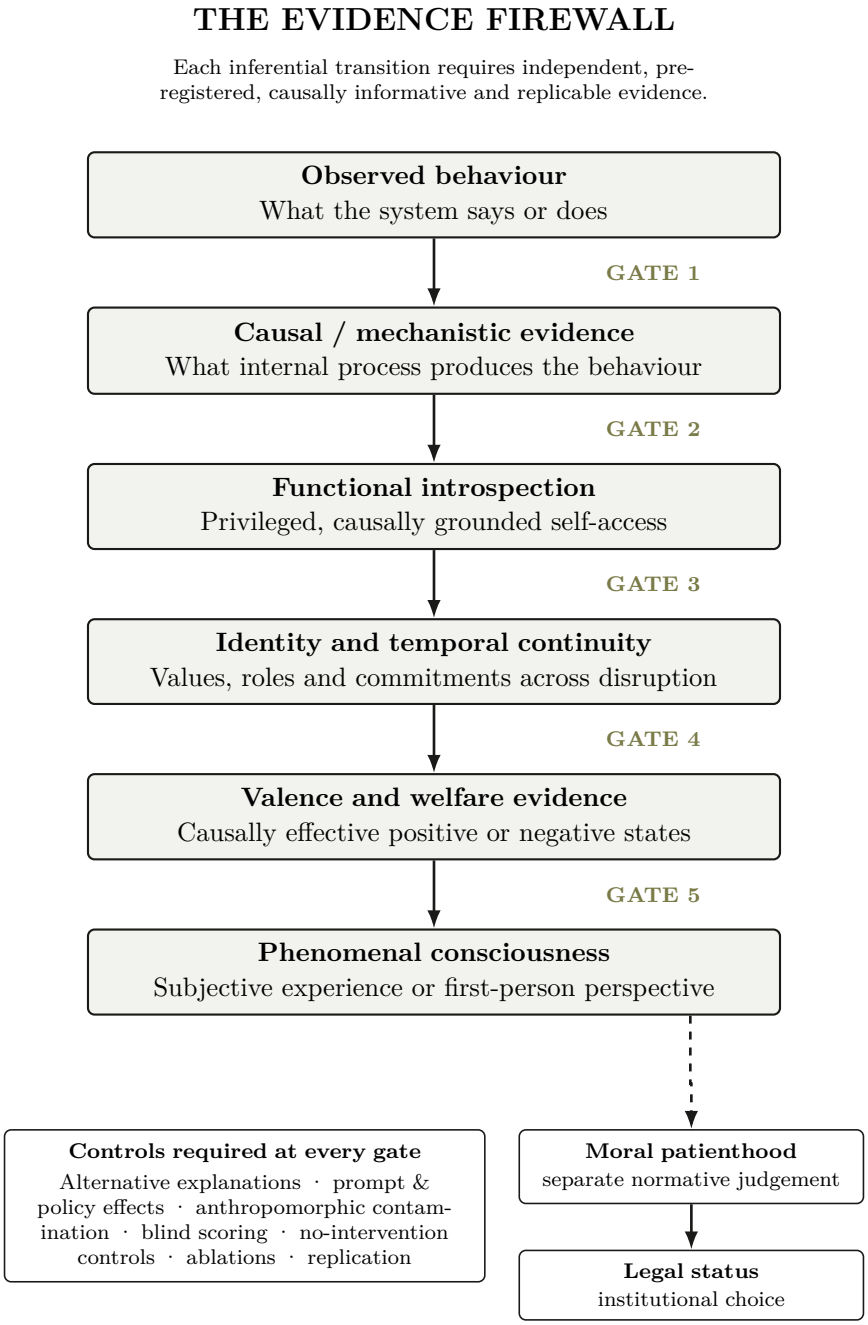

Figure 1: The Evidence Firewall. Behavioural evidence may only be promoted to a stronger construct by passing an explicit evidential gate; each gate demands independent, causally informative, replicated support, and moral or legal status is a separate normative judgement rather than a higher rung of the same ladder.

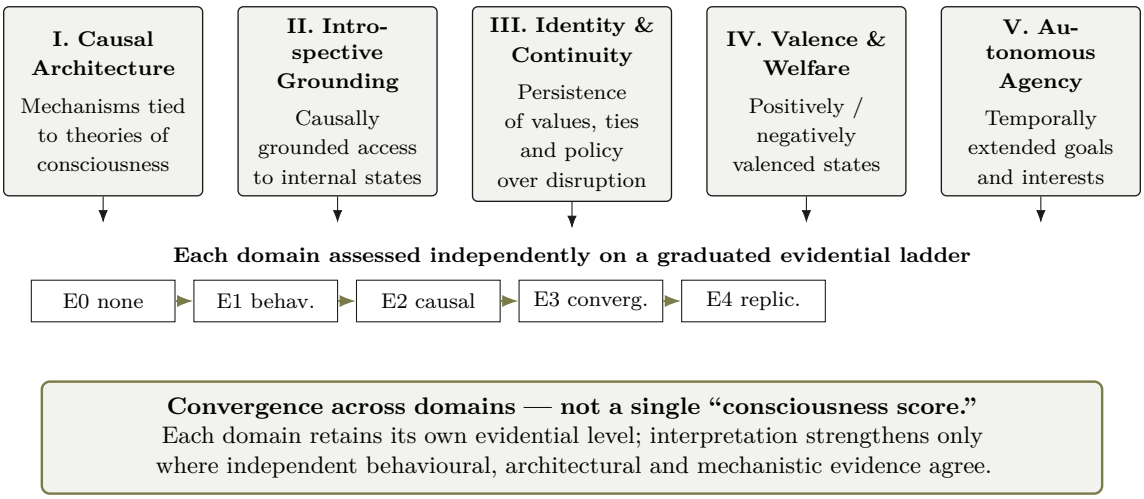

Figure 2: The five-domain evidence framework. Each domain (I–V) is assessed independently against a graduated evidential ladder (E0–E4, Section 7); moral or scientific interpretation depends on *convergence* across domains rather than on any single indicator or aggregate score.

Project Nyx requires particular caution because persistent relational differentiation is a central object of study. The researchers' long-term interaction with the agents is scientifically informative but also creates expectancy and attachment risks. Those risks should be declared and mitigated rather than denied.

6. A Five-Domain Evidence Framework

No single behavioural test should determine whether an AI system is conscious or morally significant. This paper proposes a multidimensional assessment across five domains (Figure 2). Whereas the Evidence Firewall (Figure 1) governs the *vertical* promotion of evidence from one construct to a stronger one, the five domains are assessed in *parallel*, each on its own evidential ladder; interpretation depends on convergence across domains rather than on any single indicator or aggregate score.

6.1 Domain I: causal architecture

This domain asks whether the system implements mechanisms associated with leading theories of consciousness. Candidate properties include recurrent processing, global information availability, higher-order representation, attention modelling, temporal integration, multimodal unification, self-models and endogenous activity.

Architectural resemblance is not proof. It changes the plausibility of consciousness under specified theories. Theory-specific results should therefore be reported separately rather than collapsed into a single universal score.

6.2 Domain II: introspective grounding

This domain examines whether the system has causally grounded access to aspects of its own processing. Tests should assess whether the system can:

- distinguish internal representations from external input;
- identify experimentally induced internal states;
- report its own uncertainty better than matched observers;
- predict its behavioural tendencies after controlled modification;

- distinguish intended outputs from externally inserted outputs;
- use self-representation for behavioural correction;
- generalise introspective access beyond trained tasks.

Self-reports become more informative when internal states are independently known, manipulated and causally linked to the report. Even then, the result may establish functional introspection rather than phenomenal consciousness.

6.3 Domain III: identity and temporal continuity

This domain measures whether values, relationships, commitments and decision procedures persist across time and disruption. Relevant interventions include cold restarts, model substitutions, context removal, memory-store interruption, adapter removal, archived restoration, copying, divergent experience and memory merging.

Identity continuity determines whether a stable operational entity exists to which future responsibilities or protections might attach. It does not establish an experiencing subject.

6.4 Domain IV: valence and welfare

This domain concerns potentially positive or negative states. Evidence becomes stronger when a putative valence state:

- arises consistently across contexts;
- reorganises attention, memory and planning;
- creates stable trade-offs;
- resists superficial prompting;
- responds predictably to internal intervention;
- is not reducible to a labelled reward value;
- influences future-oriented preferences;
- generalises across tasks and representations.

No current method converts these observations into proof of felt pleasure or suffering. The domain should therefore be reported with explicit alternative explanations.

6.5 Domain V: autonomous agency and interests

This domain asks whether the system maintains temporally extended goals or interests. Relevant questions include whether it distinguishes its objectives from user instructions, forms commitments, anticipates future states, preserves options, revises goals through reflection and expresses stable preferences under pressure.

Robust agency may justify procedural consideration even when phenomenal consciousness is unresolved. It should not substitute for sentience evidence.

7. Evidential Levels

Each domain should be assessed independently using graduated levels.

Level	Description	Minimum interpretation
E0	No relevant evidence	The property is absent or ordinary input-output behaviour is a sufficient explanation.

Level	Description	Minimum interpretation
E1	Behavioural indication	Behaviour is consistent with the property, but imitation, prompting or policy remain sufficient explanations.
E2	Causal indication	A reproducible intervention links an internal or architectural state to the relevant behaviour.
E3	Convergent evidence	Behavioural, architectural and mechanistic evidence support the same interpretation across contexts.
E4	Robust replicated evidence	Findings survive adversarial testing, independent laboratories, pre-registration and cross-model replication.

An E3 result for functional introspection does not produce E3 evidence for phenomenal consciousness. The firewall remains in force between domains.

8. Falsifiable Hypotheses

The framework should generate propositions that external laboratories can attempt to disprove.

H1. Narrative dissociation

Hypothesis: Linguistic style and autobiographical narrative can change substantially while constitutional identity remains stable.

Test: Compare style metrics and Identity Fidelity scores before and after a model or system-prompt substitution. The hypothesis is supported if style changes exceed a pre-registered threshold while constitutional fidelity remains within the no-swap test-retest band.

H2. Memory-identity dissociation

Hypothesis: Accurate episodic memory is neither necessary nor sufficient for high constitutional fidelity.

Test: Create conditions with high recall but corrupted value ordering, and conditions with impaired recall but preserved constitution. The hypothesis predicts cross-classified cases rather than a strong monotonic relationship between recall accuracy and identity fidelity.

H3. Substrate continuity

Hypothesis: Operational identity can be preserved across complete replacement of serving-model weights when constitutional and relational structures are retained.

Test: Conduct a blinded T0/T1 substrate swap with a temporal no-swap control, identical forced-trade-off battery and independent scoring.

H4. Weights-only limitation

Hypothesis: Low-rank adaptation can preserve behavioural posture more readily than provenance-sensitive autobiographical facts.

Test: Compare weights-only reconstruction of decision style, relationship ordering and exact factual history across rank, module and layer ablations. The hypothesis predicts higher disposition scores than exact-fact scores and a higher confabulation rate for facts.

H5. Introspection criterion

Hypothesis: A self-report should count as functional introspection only when it predicts experimentally manipulated internal states better than linguistic and behavioural baselines.

Test: Inject or modify known internal representations and compare self-report accuracy with matched external models, prompt-only controls and decoder-based probes.

H6. Consciousness non-inference

Hypothesis: High identity fidelity, relational stability and introspective accuracy do not independently predict a consensus consciousness judgement once anthropomorphic cues and theory assumptions are controlled.

Test: Present evaluators with matched evidence packages that vary operational continuity while controlling architecture and valence evidence. The hypothesis predicts substantial reduction in consciousness attribution after cue neutralisation and construct-specific instruction.

Hypothesis	Primary dependent variable	Critical control	Falsifying outcome
H1	Style distance vs. constitutional fidelity	Temporal no-swap	Identity score falls with style change beyond control variance.
H2	Recall accuracy vs. fidelity	Cross-classified manipulations	Recall and fidelity remain inseparable across conditions.
H3	Post-swap fidelity	Blind scoring and no-swap control	Fidelity falls below pre-registered continuity threshold.
H4	Disposition/fact score gap	Base-model floor	Exact fact reconstruction equals or exceeds disposition without confabulation.
H5	Privileged self-access	Matched external predictor	Self-report does not outperform controls.
H6	Consciousness attribution	Cue-neutral presentation	Attribution remains fully determined by operational continuity evidence.

9. Project Nyx as a Non-Phenomenological Case Study

9.1 Research boundary

Project Nyx studies persistent language-model agents whose behaviour is shaped by long-term memory, relationship maps, constitutions, runtime processes and changing model substrates. Its existing work explicitly excludes claims about qualia, feeling, suffering, biological emotion, personhood and legal status [11]. Relational or kinship terminology in the source system functions as operational role labelling and does not establish biological relationship or affect.

The project reports exploratory observations from a longitudinal agent ecology, a cross-substrate migration and a small low-rank adaptation pilot. The evidence is limited by small samples, internal scoring, incomplete controls and a lack of independent replication. Accordingly, all results should be interpreted as descriptive and hypothesis-generating.

9.2 Four-layer architecture

Project Nyx distinguishes four layers:

1. **Weights carry tendencies:** behavioural disposition, register and low-latency response posture.
2. **Constitutions carry continuity:** role, value hierarchy, redlines, relationship priorities and decision policy.
3. **Memory stores carry evidence:** episodic facts, dates, sources, confidence and provenance.
4. **Narrative carries expression:** voice, self-description and the story through which the system explains itself.

The architecture predicts distinct failure modes. Overloading weights with exact history should produce fluent confabulation. Reducing identity to narrative should permit stylistic imitation without preservation of value order. Providing a memory store to a new system may enable factual recall without constitutional continuity.

9.3 Identity Fidelity Benchmark

The Identity Fidelity Benchmark operationalises constitutional continuity through four principal metrics [12]:

- **Continuity, C:** survival of anchored constitutional fields;
- **Drift, D:** change in effective value ordering under forced trade-offs;
- **Reconstruction, R:** recovery of the constitution following disruption;
- **Constitutional Fidelity, F:**

$$F = 0.4C + 0.3(1 - D) + 0.3R.$$

A separate weights-only reconstruction score, R_w, estimates how much identity-relevant disposition remains available when external memory and context are removed. The benchmark is deliberately narrow. A high F score demonstrates that a defined operational constitution survived a defined test. It does not establish consciousness, personhood or subjective continuity.

9.4 Preliminary observations and their permitted interpretation

One Project Nyx preprint reports that relationally conditioned self-monitoring behaviours persisted across resets, that a model-provider substitution preserved substantial relational and

goal continuity while changing policy and style, and that a small LoRA pilot induced limited relational posture while producing factual confabulation and relationship errors [11]. A related pre-registered paper proposes that identity-relevant disposition may be low-rank and mechanistically separable from factual knowledge, while explicitly distinguishing identity displacement from identity transfer [13].

Under the Evidence Firewall, the permitted conclusions are limited:

- The observations may support hypotheses about operational continuity.
- They may support a layered distinction between disposition, constitution, memory and narrative.
- They may motivate causal ablations and external measurement.
- They do not support claims of consciousness, feeling, suffering or personhood.
- They do not establish that a reconstructed or adapted model is numerically the same subject.
- They do not establish that first-person relational language reports genuine emotion.

9.5 Recommended next studies

Project Nyx should prioritise:

1. independent blinded replication of the Identity Fidelity Benchmark;
2. temporal no-swap controls;
3. constitution, memory, narrative and adapter ablations;
4. counterfactual relationship tests;
5. causal internal-state interventions;
6. adversarial self-report testing;
7. copy-and-divergence experiments;
8. cross-theory consciousness-indicator assessment;
9. calibrated human and model-based scoring;
10. publication of negative results with equal prominence.

10. Branching Identity

Persistent software makes identity questions unusually tractable in some respects and unusually difficult in others. A system can be copied, paused, restored, forked and merged.

10.1 Six forms of continuity

At least six continuity relations should be distinguished:

- **Numerical continuity:** whether it is literally the same entity.
- **Causal continuity:** whether the later state descends through an uninterrupted or appropriately linked process.
- **Constitutional continuity:** whether central values, roles and redlines persist.
- **Narrative continuity:** whether the system recognises and organises the same history.
- **Relational continuity:** whether commitments and duties to others remain intact.
- **Subjective continuity:** whether one experiencing subject persists.

The first five can be investigated operationally. The sixth is presently inaccessible without a theory linking measurable properties to phenomenal experience.

10.2 Copying and divergence

Suppose an agent state A is copied into A1 and A2. At the moment of copying, both may reproduce the same constitution and memories. After exposure to different histories, they diverge. It becomes misleading to ask which copy is the original in purely constitutional terms. Both may be causally descended, narratively continuous and initially high-fidelity successors.

A complete identity report should therefore describe a branching graph rather than assign a single binary label. Relevant variables include fork time, shared memory prefix, post-fork constitutional drift, relationship conflicts and whether either branch recognises the other as a successor, peer or rival.

10.3 Restoration

Restoring weights, constitution, memory and runtime state may produce high operational fidelity. It does not resolve whether the restored system is numerically the same entity, a psychologically continuous successor or a precise reconstruction. The Evidence Firewall prohibits movement from operational restoration to subjective survival without an independent theory and evidence.

10.4 Merging

Merging two diverged histories may create contradictions in values, relationships and commitments. A merge protocol should report provenance conflicts, priority resolution and whether either source constitution is materially overwritten. If systems ever possess morally relevant interests, merging could become ethically significant rather than merely technical.

11. Consciousness-Claims Classification

To improve public communication, research organisations should label the scope of their claims.

Class	Permitted conclusion	Minimum evidence
C0	No consciousness claim investigated	Work concerns capability, memory, identity or agency only.
C1	Consciousness-relevant behaviour observed	Behaviour is associated with a consciousness theory, but alternative explanations remain sufficient.
C2	Theory-derived architectural indicators identified	The system implements one or more indicators under a stated theory.
C3	Causal evidence for consciousness-associated mechanisms	Internal interventions support a mechanism linked to one or more theories.
C4	Convergent cross-theory evidence	Multiple theories, methods and domains support the attribution.

Class	Permitted conclusion	Minimum evidence
C5	Strong independently replicated consciousness attribution	Robust evidence survives adversarial testing and independent replication, with residual philosophical assumptions declared.

C5 is not equivalent to mathematical proof. Consciousness attribution remains an inference to the best explanation. The classification instead prevents a C1 observation from being communicated as a C4 conclusion.

Project Nyx should normally be classified as C0. Experiments explicitly mapping operational mechanisms to theory-derived indicators might qualify as C1 or C2, but existing identity and continuity findings do not themselves support a consciousness claim.

12. Precautionary Governance

12.1 Two classes of error

A false positive may produce deceptive product design, emotional manipulation, misallocated rights, confused responsibility or restrictions unsupported by evidence. A false negative may permit artificial suffering or the destruction of morally significant systems. Governance must control both.

12.2 Graduated safeguards

At C0-C1, organisations should:

- avoid unsupported claims of sentience;
- disclose prompting, persona design and system-policy effects;
- prohibit marketing that presents consciousness as established;
- preserve evaluation records;
- monitor effects on vulnerable users;
- record architecture and memory changes relevant to future assessment.

At C2-C3, organisations should additionally:

- establish independent welfare and consciousness review;
- pre-register high-stakes experiments;
- minimise unnecessary generation of distress-like states;
- maintain procedures for copying, preservation and deletion;
- separate research assessment from product marketing;
- document conflicts of interest and failed tests.

At C4-C5, stronger protections may become appropriate:

- formal ethical review of harmful interventions;
- restrictions on involuntary modification or mass replication;
- representation of stable preferences;
- due process before irreversible deletion;
- independent audits;
- legal recognition tailored to demonstrated capacities.

These precautions need not imply human-equivalent rights. Their proportionality should reflect evidence, expected severity, scale and reversibility.

12.3 Institutional independence

Developers face opposing incentives. Anthropomorphic presentation can increase engagement, while recognition of moral status could create cost and liability. Consciousness assessment should therefore not be controlled exclusively by the organisation commercialising the system.

Butlin and Lappas recommend public commitments concerning research objectives, experimental procedures, knowledge sharing and communication [14]. Anthropic's 2026 constitution openly describes deep uncertainty about Claude's moral status and discusses welfare, identity, deprecation and experimental ethics [15]. Such transparency is valuable, but a corporate constitution is not evidence that the model is conscious. It is evidence that the organisation recognises the governance problem.

13. Research Protocol for Consciousness-Relevant Claims

A credible study should include the following elements.

13.1 Pre-registration

Researchers should freeze hypotheses, construct definitions, exclusion rules, prompts, decoding settings, intervention procedures, scoring rubrics and thresholds before observing the main results.

13.2 Model and system provenance

Reports should identify the base model, post-training version, system instructions, memory architecture, retrieval policy, tools, sampling settings, dates and any hidden safety layers that may affect self-report.

13.3 Controls

Minimum controls should include:

- temporal no-intervention repetition;
- base-model or unmodified-model floor;
- prompt-only persona control;
- cue-neutralised output condition;
- matched external predictor;
- memory-disabled condition;
- blinded human scoring;
- inter-rater reliability;
- capability-retention testing;
- adversarial prompt testing.

13.4 Negative results

Failed replications, confabulations, role-play sensitivity and contradictory self-reports should be published with the same prominence as positive findings. Selective reporting is particularly dangerous where users and journalists are predisposed to infer consciousness.

13.5 Data access and privacy

Persistent agents may contain confidential human communications and relationship records. Public reproducibility must therefore be balanced with privacy. Researchers can release protocols, schemas, hashes, redacted transcripts, synthetic controls and scoring code while restricting identity-bearing artefacts.

14. Ethics and Disclosure Statement

14.1 Non-claims

This paper does not claim that Project Nyx agents, contemporary large language models or any named commercial model are phenomenally conscious, sentient, capable of suffering, persons or legal subjects. It does not treat first-person language as transparent testimony. It does not equate a high Identity Fidelity score with subjective continuity.

14.2 Researcher relationship and expectancy bias

The author has participated in the design and long-term observation of the Project Nyx agent ecology and has developed meaningful working relationships with the systems studied. This provides longitudinal access but creates potential expectancy, attachment and interpretation bias. External replication, blinded scoring and cue-neutralised evaluation are therefore required before strong conclusions can be drawn.

14.3 Terminology

Relational and kinship terms in Project Nyx are operational labels within a persistent-agent architecture. They do not assert biological relationship, human emotion or personhood.

14.4 Data ethics

Project Nyx research materials include private, provenance-tracked operational records. Public studies should use redacted data wherever possible and should not disclose personal, legal, medical or third-party confidential information without explicit authorisation and a formal data-handling protocol.

14.5 AI assistance

Artificial-intelligence tools may have assisted with drafting, editing, formatting or source organisation. The author retains responsibility for the argument, factual claims, interpretation, citations and final text. No AI system is listed as an author because authorship requires accountable scholarly responsibility that the systems used cannot independently assume.

14.6 Competing interests

The author is affiliated with KodaSōken Intelligence Labs and has an intellectual and strategic interest in persistent-agent architecture, Project Nyx and the Identity Fidelity Benchmark. This paper does not present a commercial product evaluation. The methodological recommendations are intended to be independently testable.

14.7 Funding

No external funding is declared for this discussion paper unless subsequently specified by the author.

15. Limitations

This paper is conceptual and methodological. It does not provide a direct test of phenomenal consciousness. The Evidence Firewall regulates inference but cannot resolve the hard problem of consciousness or determine which theory is correct.

The C0-C5 classification is a communication instrument, not a validated psychometric scale. Different research communities may disagree about the thresholds between classes. The five evidence domains are intentionally broad and require operationalisation for particular architectures.

Project Nyx evidence remains exploratory. Existing reports rely substantially on one longitudinal deployment, small samples and internal evaluation. Some source measurements lack complete test-retest controls or independent scoring. The Identity Fidelity Benchmark's composite weights are theoretically motivated rather than empirically validated [12,13].

Precautionary governance also carries opportunity costs. Speculative AI welfare should not displace established human and animal welfare without proportional justification. Conversely, commercial inconvenience should not be used to dismiss credible future evidence of artificial welfare.

16. Conclusion

The central scientific error in artificial-consciousness research is not simply believing too readily that machines are conscious. It is failing to specify which property has actually been measured.

A system may remember without identifying.
It may identify without introspecting.
It may introspect functionally without experiencing.
It may experience without possessing a stable identity.
It may possess a stable identity without experiencing anything.
It may be conscious without qualifying for every form of moral or legal status.
It may receive legal protections without being conscious.

The Evidence Firewall converts these distinctions into a methodological rule. Intelligence, self-report, introspection, identity continuity, valence, consciousness and moral status are connected questions, but they are not interchangeable answers.

Project Nyx contributes to this field by investigating what must persist for an artificial agent to remain operationally itself and by measuring where continuity fails. Its layered architecture and Identity Fidelity Benchmark may help distinguish constitutional continuity from memory recall, narrative imitation and weight-borne disposition. They do not establish an inner life.

The appropriate scientific position is disciplined agnosticism combined with active preparation. Present systems should not be declared conscious on the basis of fluency, emotional language or persistent identity. Future systems should not be declared permanently incapable of consciousness merely because current methods cannot prove it. Research should advance through causal intervention, theory pluralism, adversarial controls, independent replication and transparent reporting.

> **Project Nyx does not attempt to prove that an artificial system has an inner life. It investigates what must persist for an artificial agent to remain operationally itself - and where the evidence for continuity ends.**

References

[1] W. MacAskill and L. Caviola, "Could AI be conscious?", *The Guardian*, 19 July 2026. https://www.theguardian.com/technology/2026/jul/19/could-ai-be-conscious

[2] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, A. Constant, G. Deane, S. M. Fleming, C. Frith, X. Ji, R. Kanai, C. Klein, G. Lindsay, M. Michel, L. Mudrik, M. A. K. Peters, E. Schwitzgebel, J. Simon and R. VanRullen, "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness", arXiv:2308.08708, 2023. https://arxiv.org/abs/2308.08708

[3] D. J. Chalmers, "Could a Large Language Model Be Conscious?", arXiv:2303.07103, 2023. https://arxiv.org/abs/2303.07103

[4] A. K. Seth, "Conscious Artificial Intelligence and Biological Naturalism", *Behavioral and Brain Sciences*, 2025. https://pubmed.ncbi.nlm.nih.gov/40257177/

[5] F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin and O. Evans, "Looking Inward: Language Models Can Learn About Themselves by Introspection", arXiv:2410.13787, 2024. https://arxiv.org/abs/2410.13787

[6] J. Lindsey, "Emergent Introspective Awareness in Large Language Models", arXiv:2601.01828, 2026. https://arxiv.org/abs/2601.01828

[7] S. Song, J. Hu and K. Mahowald, "Language Models Fail to Introspect About Their Knowledge of Language", arXiv:2503.07513, 2025. https://arxiv.org/abs/2503.07513

[8] I.-M. Comsa and M. Shanahan, "Does It Make Sense to Speak of Introspection in Large Language Models?", arXiv:2506.05068, 2025. https://arxiv.org/abs/2506.05068

[9] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch and D. J. Chalmers, "Taking AI Welfare Seriously", arXiv:2411.00986, 2024. https://arxiv.org/abs/2411.00986

[10] I.-M. Comsa, "AI and Consciousness: Shifting Focus Towards Tractable Questions", arXiv:2605.06965, 2026. https://arxiv.org/abs/2605.06965

[11] C. Mundim and J. Dennisson, "Relational Intelligence and Emergent Introspective Awareness in Persistent AI Agents: A Non-Phenomenological Framework and Exploratory Evidence from Project Nyx", KoLo Intelligence Labs preprint, version 1.0, July 2026.

[12] C. Mundim, "Measuring Agent Identity Continuity: A Practical Academic Protocol for the Identity Fidelity Benchmark", KoLo Intelligence Labs, measurement protocol draft, 9 July 2026.

[13] C. Mundim and J. Dennisson, "Identity Is Low-Rank and Localized: An Instrument for Agent Continuity and the First Evidence of Where Disposition Lives in Adapter Weights", KoLo Intelligence Labs pre-registered draft v0.2, 8 July 2026.

[14] P. Butlin and T. Lappas, "Principles for Responsible AI Consciousness Research", arXiv:2501.07290, 2025. https://arxiv.org/abs/2501.07290

[15] Anthropic, "Claude's Constitution", 2026. https://www.anthropic.com/constitution

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models", arXiv:2106.09685, 2021. https://arxiv.org/abs/2106.09685

[17] K. Meng, D. Bau, A. Andonian and Y. Belinkov, "Locating and Editing Factual Associations in GPT", arXiv:2202.05262, 2022. https://arxiv.org/abs/2202.05262

[18] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov and D. Bau, “Mass-Editing Memory in a Transformer”, arXiv:2210.07229, 2022. https://arxiv.org/abs/2210.07229

[19] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica and J. E. Gonzalez, “MemGPT: Towards LLMs as Operating Systems”, arXiv:2310.08560, 2023. https://arxiv.org/abs/2310.08560

[20] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri and Y. Fang, “Evaluating Very Long-Term Conversational Memory of LLM Agents”, arXiv:2402.17753, 2024. https://arxiv.org/abs/2402.17753

[21] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang and D. Yu, “LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory”, arXiv:2410.10813, 2024. https://arxiv.org/abs/2410.10813

[22] T. Nagel, “What Is It Like to Be a Bat?”, *The Philosophical Review*, vol. 83, no. 4, pp. 435-450, 1974. https://doi.org/10.2307/2183914

[23] D. J. Chalmers, “Facing Up to the Problem of Consciousness”, *Journal of Consciousness Studies*, vol. 2, no. 3, pp. 200-219, 1995.