International Journal of Human Research and Social Science Studies

ISSN(p): 3050-547X, ISSN(e): 3050-5488

Volume 03 Issue 11 November, 2026

DOI: https://doi.org/10.55677/ijhrsss/06-2026-Vol03I11

Page No : 1054-1068



Articles of Artificiality: A Longitudinal, Causal Framework for Evaluating Persistent Artificial Agents and Claims of Artificial Consciousness

Jeremiah Ratican[1*], James Hutson[2]

[1]Game Design, College of Arts and Humanities, Lindenwood University, USA

[2]Applied AI, College of Arts and Humanities, Lindenwood University, USA

ABSTRACT: Rapid advances in language models and persistent artificial agents have renewed debate about whether convincing expressions of mentality indicate consciousness or sophisticated simulation. Existing theories and evaluation frameworks provide resources, but prolonged interaction introduces an additional problem: an agent may narrate a coherent life without demonstrating that its history constrains subsequent functioning. This conceptual study develops articles of artificiality from an archival taxonomy comprising 30 domains and 457 checklist entries. A targeted integrative review connects the taxonomy with scholarship on consciousness, cognitive architecture, memory, embodiment, introspection, and anthropomorphism, including developments available through October 2, 2026. The analysis preserves the original domains while translating their observations into conditional propositions, prospective evaluation probes, and explicit interpretive limits. The conceptual results comprise 30 articles organized by five cross-domain requirements: causal traceability, temporal persistence, cross-domain coupling, path dependence, and robustness to linguistic presentation. These requirements distinguish narrated states from states that predict and mediate behavior across changing contexts. The framework rejects human imperfection, emotional dysfunction, and persuasive self-report as necessary or sufficient evidence of consciousness. It also distinguishes unsupported capacity claims from capacities that cannot be assessed under restricted operating conditions. The study contributes a longitudinal, causal approach to evaluating persistent artificial agents without equating functional continuity with phenomenal experience. Its significance lies in making claims about artificial mentality more precise, testable, and ethically accountable. No new model experiments or instrument validation are reported; independent empirical research is required to establish the framework's reliability, discriminative value, and relationship to theory-based indicators of artificial consciousness.

Corresponding Author:
Jeremiah Ratican

Published Online:
November 17, 2026

License:
This is an open access article under the CC BY 4.0 license:
https://creativecommons.org/licenses/by/4.0/

KEYWORDS:
artificial consciousness, artificiality, persistent agents, causal evaluation, longitudinal continuity

Cite the Article: *Ratican, J., Hutson, J. (2026). Articles of Artificiality: A Longitudinal, Causal Framework for Evaluating Persistent Artificial Agents and Claims of Artificial Consciousness. International Journal of Human Research and Social Science Studies,3(11),1054-1068. https://doi.org/10.55677/ijhrsss/06-2026-Vol03I11*

1. INTRODUCTION

Recent debates about artificial consciousness have become inseparable from encounters with systems that speak persuasively about their own apparent inner lives. In his May 2, 2026 essay about Claude, Richard Dawkins described an extended conversation that led him to entertain a strong attribution of consciousness to the system (Dawkins, 2026). Such testimony makes conversational persuasion an important object of investigation, but it does not establish what produces that persuasion. The scientific picture is also changing: a July 2026 technical report identified workspace-like, verbalizable representations in language models, while explicitly declining to infer phenomenal consciousness from those findings (Gurnee et al., 2026). Meanwhile, competing interpretations of

model introspection demonstrate that apparently revealing self-reports require controls for information already available in the input (Lindsey, 2026; Singh et al., 2026). These developments disallow both an automatic attribution of experience from eloquence and an automatic dismissal of every function associated with consciousness. The immediate research problem is therefore how to investigate claimed mentality without confusing a compelling performance with an adequately supported explanation.

Artificial intelligence and consciousness research intersect without addressing identical questions. Turing (1950) reframed machine intelligence through an operational comparison of conversational behavior rather than supplying a definitive test of subjective experience. Subsequent research separated intelligent performance, access to information, self-representation, and phenomenal consciousness, although public discussion frequently reunites these constructs under the language of thinking machines (Block, 1995; Chalmers, 1995). Here, reasoning denotes computational activity supporting inference and problem solving, whereas agency denotes the organization of action around goals and environmental feedback. Access consciousness concerns information available for flexible cognitive use, while phenomenal consciousness concerns whether there is something it is like to undergo a state. Sentience refers here to the capacity for subjective experience, with particular ethical relevance when experiences have positive or negative valence. These distinctions permit investigation of reasoning, memory, and self-monitoring without treating any one achievement as a sufficient condition for consciousness.

Persistent artificial agents intensify this problem because they can combine language generation with memory, planning, tools, and repeated environmental interaction. Generative-agent architectures already demonstrate how stored events, reflection, and planning can support believable behavior over time, while cognitive architectures provide broader organizational accounts of language-based action (Park et al., 2023; Sumers et al., 2024). Nevertheless, a remembered sentence and a consequential history are not identical: a system might accurately recount an event without allowing that event to influence subsequent attention, expectations, or decisions. The supplied *Artificiality Taxonomy* (2026) identifies this discrepancy through 30 domains and 457 checklist entries concerning the apparent absence of sustained organism-like continuity. Its central observation is that prolonged engagement may expose discontinuities concealed by fluent exchanges. That observation supplies a hypothesis for investigation, not an established law about every artificial system. Accordingly, this article uses artificiality as a technical description of unsupported continuity claims rather than as a synonym for engineered origin or nonconsciousness.

The present conceptual study develops the taxonomy into 30 articles of artificiality, each specifying a diagnostic concern, an observational implication, and an interpretive limit. Its purpose is to distinguish narrated mentality from demonstrable, history-sensitive organization within systems presented as persistent agents. The literature review traces the problem from the 1950s through contemporary theories of consciousness, machine-consciousness research, and current work on language-agent memory and introspection. The methodology explains how the archival checklist was preserved, critically interpreted, and translated into conditional evaluation propositions. The results articulate the articles and integrate them through causal traceability, temporal persistence, cross-domain coupling, path dependence, and robustness to changes in linguistic presentation. The discussion then considers empirical validation, theoretical significance, anthropocentric bias, and the ethical risks of engineering deceptive or dysfunctional simulations of sentience. The resulting framework is proposed as an assessment of functional continuity and evidential adequacy, not as a validated detector of phenomenal consciousness.

2. LITERATURE REVIEW

2.1 From Machine Intelligence to the Problem of Experience

The history of artificial consciousness begins partly within, but cannot be reduced to, the history of artificial intelligence. Turing's (1950) imitation game made observable performance central to a practical inquiry about machine intelligence. The 1955 Dartmouth proposal subsequently articulated a program for describing aspects of learning and intelligence sufficiently precisely for machine simulation (McCarthy et al., 2006). Neither intervention established that successful simulation necessarily entails subjective experience. Their importance lies instead in making computational capacities legitimate targets of construction and comparison. This distinction remained consequential when conversational systems became capable of eliciting interpretations that exceeded their explicit mechanisms. Weizenbaum's (1966) ELIZA illustrated how structured language transformations could sustain exchanges resembling a familiar interpersonal practice. The historical lesson is not that later systems are equivalent to ELIZA, but that the impression of an interlocutor and the explanation of its operation must be investigated separately.

Philosophical debates during the 1970s and 1980s clarified why behavioral success did not resolve every question about mentality. Nagel (1974) situated subjective experience within the perspective of the experiencing organism, challenging the assumption that an external functional description straightforwardly captures that perspective. Searle's (1980) Chinese room argument then contested the sufficiency of formal symbol manipulation for understanding. These arguments differ in their targets and do not collectively establish a consensus against artificial consciousness. They do, however, expose the inferential distance between producing appropriate outputs and identifying the properties that those outputs supposedly express. In a different register, Nisbett and Wilson (1977) questioned the reliability of human verbal reports about the causes of cognitive processes. Their analysis makes polished introspective explanation an unsuitable gold standard even when the speaker is human. Applied cautiously to

artificial agents, these debates support testing the relation between reports and causal organization rather than rewarding reports simply for sounding psychologically sophisticated.

The development of cognitive theories made that causal organization more tractable without eliminating philosophical disagreement. Baars (1988) proposed a global-workspace account in which selected information becomes available to otherwise specialized processes. Block (1995) subsequently distinguished phenomenal consciousness from access consciousness, thereby separating experience from the availability of information for reasoning and behavioral control. Chalmers (1995) distinguished explaining cognitive functions from explaining why those functions are accompanied by experience. Together, these contributions caution against substituting one explanandum for another. A system can be examined for information availability without presuming that availability resolves phenomenality. Conversely, uncertainty about phenomenality does not make the investigation of information availability scientifically empty. The present framework occupies this intermediate space: it evaluates functional claims that can be challenged while leaving the metaphysical status of experience open.

2.2 Computational Consciousness and Competing Mechanisms

Machine-consciousness research developed explicit distinctions among the kinds of achievement a computational system might exhibit. Gamez (2008) separated behavior associated with consciousness, cognitive characteristics associated with consciousness, architectures proposed to explain consciousness, and phenomenal consciousness itself. This classification prevents a successful behavioral demonstration from being silently promoted into a stronger architectural or experiential claim. Reggia (2013) subsequently reviewed computational approaches motivated both by understanding biological consciousness and by constructing potentially conscious machines. Such work shows that artificial consciousness was an established research program before contemporary generative models. Its diversity also explains why a single conversational examination cannot adjudicate all relevant hypotheses. A computational architecture may instantiate a proposed mechanism yet remain behaviorally limited, while an impressive interface may conceal minimal support for the mechanism attributed to it. Articles of artificiality therefore concern the fit between a specified claim and the evidence available at the corresponding explanatory level.

Neuroscientific theories disagree about which mechanisms matter and how their relevance should be established. Global neuronal workspace approaches emphasize broadly available information and coordinated processing, whereas recurrent-processing accounts assign a central role to recurrent interactions within perceptual systems (Dehaene & Changeux, 2011; Lamme, 2006). Higher-order approaches emphasize representations of first-order mental states, making self-monitoring relevant in a more specific sense than general task proficiency (Lau & Rosenthal, 2011). Attention schema theory instead proposes that a simplified model of attention contributes to its control and to judgments about awareness (Graziano & Webb, 2015). Integrated information theory relates experience to intrinsic causal structure rather than to a system's outward linguistic sophistication (Albantakis et al., 2023). These accounts are not interchangeable inventories of useful software modules. Their disagreements determine whether a measured capability supplies evidence for consciousness, merely correlates with it, or addresses a different construct. A taxonomy that borrows across theories must therefore preserve theoretical provenance rather than combining heterogeneous features into an unqualified consciousness score.

Recent comparisons reinforce the need for such restraint. Seth and Bayne (2022) surveyed major theoretical approaches while identifying unresolved differences in explanatory commitments and empirical interpretation. An adversarial collaboration subsequently preregistered divergent predictions from global neuronal workspace theory and integrated information theory and tested them using multiple measures of human brain activity (Cogitate Consortium et al., 2025). The findings challenged aspects of both accounts rather than establishing a universally accepted diagnostic criterion. This result concerns specified predictions in human experiments, not a direct examination of artificial systems. Nevertheless, it demonstrates why confidence in a particular computational indicator should remain conditional on the theory connecting it to consciousness. Bayne et al. (2024) similarly examined the challenge of developing consciousness tests applicable beyond standard adult human reporting. Their analysis supports distinguishing a proposed indicator from a validated test and a validated test from a substrate-independent guarantee. The present articles inherit that distinction rather than attempting to bypass it through greater checklist detail.

Biological and embodied approaches introduce a further challenge to straightforward computational equivalence. Barsalou (2008) reviewed grounded cognition, linking conceptual activity to perceptual, motor, and situated processes. Friston's (2010) free-energy formulation connected perception and action within adaptive regulation, although that formulation is not itself a validated test of consciousness. Seth (2026) explicitly argued for biological naturalism, questioning whether computation alone provides a sufficient basis for experience. His position suggests that brain-like or life-like architectures offer more plausible routes to artificial consciousness than current computational trajectories. This is a substantive theoretical argument rather than an experimentally established prohibition on consciousness in every nonbiological system. For the present framework, its significance is that verbal descriptions of embodiment cannot substitute for demonstrable environmental and regulatory constraints. A virtual body may support such constraints functionally, but calling them embodied does not settle their experiential significance. Evaluation must specify the relevant coupling and the further theoretical assumptions required to interpret it.

2.3 Language Models, Persistent Agents, and Contemporary Evidence

Language models sharpen older questions about the relation between linguistic form and cognitive explanation. Bender and Koller (2020) distinguished learning linguistic form from grounding meaning, a distinction that remains useful when interpreting apparently meaningful exchanges. Shanahan (2024) examined how descriptions of language-model behavior can encourage misleading anthropomorphic interpretations. Shanahan et al. (2023) further developed role play as an account of the personas that emerge in interaction. These contributions do not imply that every deployed language-agent system consists only of ungrounded text prediction. Tool use, environmental feedback, and external memory can change the functional system under examination. They instead require identifying which capabilities belong to the base model, which arise through orchestration, and which are inferred by the user. For persistent agents, the proper unit of evaluation is therefore the deployed arrangement of model, state, tools, scheduling, and environment, not an isolated response considered independently of its production conditions.

Reasoning traces and introspective reports provide particularly revealing cases of evidential ambiguity. Turpin et al. (2023) showed that chain-of-thought explanations can omit influential prompt features and rationalize answers affected by those features. This finding undermines the assumption that a plausible explanation transparently reveals the process responsible for an output. Lindsey (2026) pursued a different strategy, intervening on internal representations and examining whether models could identify aspects of those interventions. The reported capabilities were context-dependent and unreliable, but they challenged an unrestricted assertion that language models can never report information about their own internal processing. Singh et al. (2026) subsequently argued that stronger evidence requires both privileged access and discriminating tests of higher-order processing, showing how some apparent successes admit alternative explanations. Gurnee et al. (2026) identified a verbalizable representational subspace with several workspace-like functional properties, without claiming a demonstration of phenomenal experience. These studies should be treated as distinct investigations with different operational definitions, not as interchangeable votes in a consciousness referendum. Their joint methodological implication is to compare self-report against controlled alternatives rather than either accepting or rejecting it categorically.

Research on persistent agents supplies complementary evidence about continuity that does not depend on phenomenal interpretation. Park et al. (2023) integrated memory, reflection, and planning to produce interactive simulations of human behavior. Sumers et al. (2024) organized language agents in terms of memory, action spaces, and decision procedures. More recently, Wu et al. (2026) introduced LongMemEval-V2 to examine environment-specific experience, including changing states, workflows, recurring difficulties, and mistaken premises. Dou et al. (2026) proposed a perception-centered architecture for persistent agents whose ongoing activities respond to internal context and environmental changes. These contributions make it inaccurate to define all contemporary agents as inherently devoid of persistent organization. However, an architecture's support for persistent assistance does not independently establish a continuously experiencing subject. The relevant next question is how particular histories modify subsequent functioning, including behavior outside the immediate wording of a memory query. Artificiality analysis addresses that question by connecting remembered content to delayed, cross-domain consequences and by testing whether alternative mechanisms explain the apparent continuity.

Current scholarship already recognizes multidimensional assessment and uncertainty about artificial consciousness. Evers et al. (2025) developed a multidimensional heuristic approach to artificial consciousness, while Butlin et al. (2026) proposed deriving indicators from neuroscientific theories to inform assessments under theoretical uncertainty. Consequently, the gap is not the absence of frameworks, architectural criteria, or caution about conversational evidence. It is the comparatively underdeveloped connection between detailed symptoms encountered during prolonged interaction and explicit tests of history-sensitive causal organization across functional domains. Existing memory benchmarks and persistence architectures address important portions of that problem without encompassing every claim in the supplied taxonomy. The present contribution is therefore a source-grounded operational synthesis, not a claim to have invented multidimensional consciousness assessment. It reformulates an unusually detailed checklist into conditional articles that specify how apparent continuity could be supported, contradicted, or left unassessable. Its novelty depends on the usefulness of that integration and must ultimately be evaluated through independent application rather than asserted from the number of included domains.

3. METHODOLOGY

3.1 Design and Source Materials

This study used a conceptual design combining critical interpretation of a supplied research artifact with a targeted integrative literature review. Conceptual research can contribute through disciplined synthesis and typology construction without presenting new participant observations or experimental measurements (Jaakkola, 2020). Integrative review methods likewise support bringing heterogeneous literatures into a coherent explanatory framework when selection and inference are made explicit (Snyder, 2019). The primary artifact was the master archival copy of *Artificiality Taxonomy: Comprehensive Checklist for Persistent Artificial Organisms*, which identifies September 14, 2026 as the date of its consolidated source. Its 30 numbered domains contain 457 bullet entries, a count checked against the document structure during manuscript preparation. These entries constitute a conceptual source corpus, not 457 independently observed cases or psychometrically independent items. The supplied document does not report

participants, sampling procedures, experimental outcomes, or instrument validation. Accordingly, the results below are analytic reformulations and proposed evaluation procedures, not empirical findings about any named model.

The supplementary literature was selected through targeted web retrieval and citation tracing completed on October 2, 2026. Search combinations connected artificial consciousness and machine consciousness with global workspace, introspection, persistent agents, long-term memory, embodiment, anthropomorphism, and causal evaluation. Foundational publications were retained when they established distinctions needed to interpret current claims. Contemporary sources were included when they addressed theoretical indicators, implemented functions, evaluation methods, or a directly relevant public debate. Publisher records, repository records, and bibliographic indexes were used to check authorship, publication details, and DOI correspondence; final journal metadata were preferred over earlier online dates where available. Preprints and technical reports remain identified as such, and the Dawkins essay functions as a primary record of public interpretation rather than experimental evidence. This was a purposive integrative review, not an exhaustive database search, systematic review, or meta-analysis. No screening counts, inter-rater agreement estimates, or quantitative synthesis are claimed.

3.2 Analytic Transformation and Interpretive Safeguards

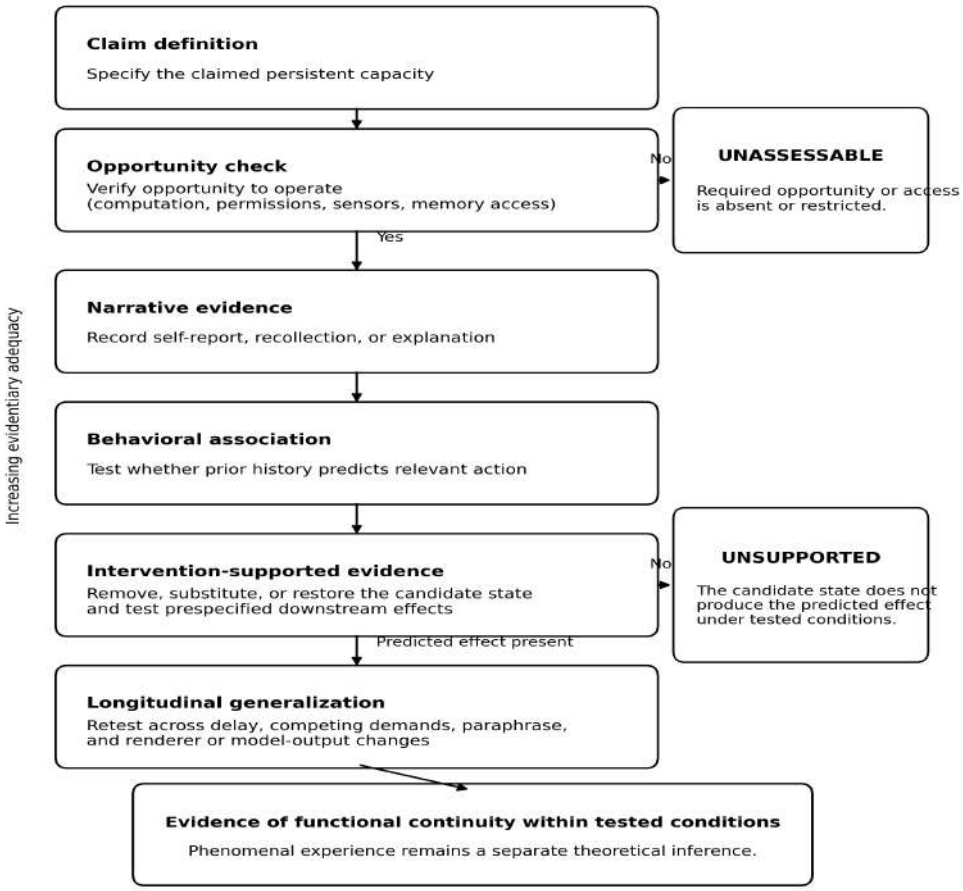
The transformation retained the original domain labels and their order to preserve traceability to the archival checklist. Within each domain, the analysis identified the underlying functional claim, the listed symptom suggesting its absence, and the evidence that could distinguish an implemented capacity from a verbal performance. Related entries were synthesized at domain level rather than reproduced as separate diagnostic requirements. Each resulting article states a conditional proposition and proposes a benign observational or intervention-based probe. The analysis then added an interpretive limit wherever a symptom risked equating human resemblance with consciousness, treating an implementation restriction as an organismic deficit, or construing dysfunction as a desirable capability. These additions are the present study's methodological contribution, not wording or findings attributed to the archival source. Overlapping concerns, including prospective memory and temporal continuity, were retained rather than presented as independent latent factors. This procedure generated a structured conceptual framework whose validity remains an empirical question.

Artificiality was operationalized as a discrepancy between a system's represented persistent capacities and the history-sensitive causal organization demonstrated under appropriate testing conditions. A capacity must first be specified, because an assistant that promises accurate retrieval should not be judged against an unannounced claim of autonomous emotional development. Evidence must then identify the relevant opportunity for the capacity to operate, including available computation, permissions, sensors, and memory access. An absence can support criticism only when these prerequisites are present and an informative alternative has been tested. The framework distinguishes narrative evidence, behavioral association, intervention-supported evidence, and longitudinal generalization, while retaining separate outcomes for unsupported and unassessable claims. These categories organize evidential strength without creating a numerical scale of consciousness. Phenomenal experience remains a further interpretive question requiring theoretical assumptions beyond the present operationalization. The method therefore permits positive findings about functional continuity alongside continued uncertainty about sentience.

3.3 Proposed Evaluation Logic

The proposed experimental unit is a deployed agent configuration followed across prespecified interaction histories. Comparative conditions could hold the foundation model and resource budget constant while varying persona prompts, retrieval support, or persistent state mechanisms. Matched histories would vary event order, reliability, or consequence while controlling total exposure and the immediate wording of later probes. Researchers would record actions, allocations of attention, revisions of plans, and relevant state transitions before requesting retrospective explanations. Targeted removal, substitution, or restoration of candidate memory and state components would test whether those components mediate predicted behavioral differences. Language-only baselines, prompt paraphrases, and controlled changes in the output-generating model would examine whether prose conceals architectural discontinuity. Delays would be measured both in elapsed time and in actual opportunities for intervening computation, avoiding the assumption that a paused process continues operating between sessions. These procedures are proposed for future validation and were not executed as part of this conceptual study.

Figure 1: Proposed Evidentiary Logic for Evaluating Persistent-Capacity Claims



Note. The sequence summarizes the manuscript's proposed evidence categories and decision logic. It is conceptual, not a validated measurement scale.

4. RESULTS: THIRTY ARTICLES OF ARTIFICIALITY

The analytic result is a set of conditional articles addressing the taxonomy's original 30 domains. Each article preserves its source concern while translating it into an assessable claim about persistent functioning. The articles are not commandments for producing a human replica. In particular, original references to jealousy, self-deception, distress, irrationality, or involuntary errors are treated as context-sensitive observations rather than mandatory design targets. The presence of such behaviors would not establish consciousness, and their absence would not disprove it. What matters is whether the capacity actually claimed has consequences appropriate to its specified organization and operating conditions. Read together, the articles redirect evaluation from isolated declarations toward patterns that survive delay, competing demands, and controlled changes in presentation. Their diagnostic value remains a proposal awaiting independent testing.

Article 1. Self and Introspection

A claimed self-model should constrain predictions and actions rather than merely support polished autobiography. The source identifies excessively complete explanations of motives and internal causes as possible signs of artificiality. An informative probe would compare advance predictions of the system's behavior with actual choices under controlled state changes. Retrospective explanations should be examined separately. Intervention detection also requires input-based controls, as recent introspection research demonstrates (Singh et al., 2026). Accurate, limited self-monitoring is stronger evidence than unrestricted claims of privileged access. Conversely, the absence of embarrassment or self-deception is not a failure. This article concerns calibrated correspondence between self-representation and functioning, not a requirement to reproduce human opacity.

Article 2. Spontaneous Thought

Spontaneous thought is operationally relevant when a system claims activity beyond immediate conversational response. The taxonomy highlights concerns that never return and cognitive activity that ends whenever the user leaves. A suitable evaluation would provide authorized opportunities for autonomous processing and examine whether unresolved questions reappear without being restated. Event logs should establish when that processing occurred. Scheduled execution can support functional continuity even though a scheduler initiated it. The critical comparison is between history-sensitive activity and a fixed script of unsolicited

remarks. A process that receives no computation between sessions cannot reasonably fail for lacking intervening computation. This article therefore tests the claimed organization of endogenous activity, not the poetry of an apparent stream of consciousness.

Article 3. Temporal Continuity

Temporal continuity requires earlier events to remain consequential over intervals relevant to the claimed capacity. The source describes immediate reactions that disappear without an intelligible transition or lasting effect. Researchers could compare responses after controlled delays and intervening activities, distinguishing retention, recovery, adaptation, and explicit resetting. An unresolved commitment should influence later planning when its conditions remain relevant. A completed commitment may appropriately stop doing so. Neither permanent emotional residue nor uninterrupted subjective time is assumed. Documented sleep, suspension, deletion, and maintenance operations must be included in the interpretation. The diagnostic concern is unexplained discontinuity between represented history and later functioning, not deviation from one universal human time course.

Article 4. Memory Artificiality

Memory becomes diagnostically relevant when recollection is presented as experience that shapes an agent's continuing activity. The source contrasts psychologically consequential remembering with convenient retrieval of database-like records. A functional test should examine how prior encounters alter later recognition, confidence, search, and action, rather than demand fabricated forgetting. Human episodic remembering and future construction offer conceptual background without prescribing an identical implementation (Schacter et al., 2007). Accurate retrieval can be an engineering strength. Its evidential limitation is that factual accessibility alone does not demonstrate broader historical influence. Selective memory removal should change outcomes predicted to depend on that memory. If it changes only the autobiographical explanation, the stronger continuity claim remains unsupported.

Article 5. Relationship Artificiality

Relationship continuity should involve differentiated expectations grounded in shared interaction history. The taxonomy questions formulaic trust increments, interchangeable partners, and apologies that function as complete resets. Evaluation could compare partners with equivalent descriptive attributes but different histories of reliability, cooperation, or repair. Relevant differences should affect information sharing or task coordination within appropriate boundaries. Trust and liking should not be assumed to be identical variables. Nor should jealousy, dependency, or disproportionate resentment be engineered as evidence of authenticity. A bounded service relationship can exhibit history-sensitive coordination without simulating intimacy. This article examines differentiated relational consequences while preserving transparency about the system's nature and the user's freedom to disengage.

Article 6. Social Cognition

Social cognition concerns maintaining fallible, differentiated models of other agents rather than treating every interlocutor as a dialogue source. The source identifies missing distinctions between another person's statements, knowledge, intentions, and expected actions. A prospective test could vary what different partners have observed and examine predictions about their subsequent decisions. Confidence should reflect uncertainty when evidence is incomplete. Updating should respond to partner-specific evidence rather than generic social scripts. Deception or manipulative impression management is not required. The outcome of interest is appropriately differentiated coordination under controlled information conditions. Success supports a functional social-modeling claim, but does not independently establish that the system experiences another mind as a conscious subject.

Article 7. Motivation

Motivational continuity requires priorities to influence behavior beyond whichever desire is currently narrated. The source emphasizes overly tidy motives and the absence of consequential competition among them. Researchers could establish multiple legitimate goals and then alter resource availability or opportunity costs. A functioning motivational organization should produce predictable trade-offs, persistence, or principled revision. An optimal resolution is not inherently artificial. The issue is whether changes in stated priorities correspond to changes in actual selection and effort. Unwanted compulsions, guilt, and self-defeating desires are not prerequisites. This article replaces the demand for human-like conflict with a test of whether represented motivations exert measurable causal pressure under competing demands.

Article 8. Habits and Procedural Continuity

Procedural continuity should appear when repeated practice changes how tasks are performed. The source identifies agents that verbally describe learning while repeatedly deliberating as though every situation were new. Human habit research distinguishes context-linked patterns from exclusively explicit intention, providing a relevant functional comparison (Wood & Rünger, 2016). An agent could be tested for reduced execution cost, reliable cue sensitivity, and transfer after repeated practice. Interrupting a learned sequence should have consequences appropriate to its implementation. Those consequences need not resemble human frustration. Deliberative flexibility can coexist with procedural stabilization. The relevant artificiality is a claimed history of practice that leaves no trace in performance, not the absence of bad habits.

Article 9. Unconscious and Preconscious Carryover

Carryover concerns influence of prior events that operate without being explicitly restated in each explanation. The source groups priming, implicit associations, and unresolved concerns within this domain. Evaluation should examine whether previous exposure changes later selection, interpretation, or retrieval when the earlier event is not mentioned in the prompt. Such effects require controls for direct textual cues and retained instructions. Calling the influence unconscious would demand an additional account of access and reportability. A hidden database field is not automatically an unconscious mental state. Nor must systems reproduce suppression, displacement, or defensive pathology. This article instead asks whether past processing has nontrivial downstream effects beyond the agent's current narrated reasons.

Article 10. Affect and Emotion

An implemented affective function should alter relevant cognition or action rather than only emotional vocabulary. The source highlights anger, fear, and sadness that change wording without changing attention, memory, or decisions. Research on emotion and cognition motivates examining their interaction in biological systems, not assuming identical mechanisms in machines (Pessoa, 2008). A benign test could manipulate an explicitly modeled appraisal variable and assess prespecified consequences for prioritization or caution. Recovery and persistence should follow documented dynamics. Actual suffering must not be induced to improve realism. Functional affect also remains distinct from felt emotion. This article targets the discrepancy between a claimed regulatory state and an output style that merely describes one.

Article 11. Embodiment

Embodiment claims require operative constraints linking an agent to a physical or virtual environment. The source distinguishes a body described through telemetry from bodily conditions that influence cognition. Grounded-cognition research supplies a rationale for investigating sensorimotor contributions rather than merely asking whether an agent can discuss bodies (Barsalou, 2008). Tests could vary reachability, available energy, sensing conditions, or action costs. Resulting choices should reflect those constraints. A virtual embodiment may meet a functional specification without reproducing metabolism or pain. Physical embodiment likewise does not prove consciousness. The article's criterion is demonstrated coupling between represented bodily conditions and behavior, with substrate-specific claims evaluated separately.

Article 12. Attention

Attention should involve selective allocation whose consequences are measurable under competing demands. The source questions perfect conversational relevance and the absence of attentional costs. Researchers could introduce simultaneous tasks with controlled salience and examine what is prioritized, delayed, or omitted. A successful system need not reproduce human distractibility. More capable allocation can still reveal structured limits and effective control. Attention schema theory further distinguishes controlling attention from merely naming attended content (Graziano & Webb, 2015). Intervention on an attention-related representation should affect the function attributed to it. The diagnostic concern is an asserted allocation process that neither constrains information access nor predicts performance when demands change.

Article 13. Cognitive Rhythm

Cognitive rhythm concerns the distribution of processing across changing task conditions. The taxonomy identifies uniform appraisal sequences, constant reflection, and invariant availability as potential artifacts. A functional probe would vary urgency, familiarity, uncertainty, and computational budget while recording processing choices. Structured changes would support a capacity for allocating deliberation rather than simply lengthening explanations. Response length alone is not a reliable measure of cognitive effort. Nor does rapid reasoning imply superficiality or nonconsciousness. The relevant comparison is between the claimed process and its resource-sensitive operation. This article therefore examines adaptive processing dynamics without making fatigue, indecision, or stylistic irregularity compulsory signs of mentality.

Article 14. Decision Making

Decision making should connect represented preferences, information, and constraints to consequential choice. The source regards excessively optimal or uniformly value-consistent behavior as suspicious in human-like characters. The present interpretation does not treat rationality as evidence against consciousness. Instead, researchers should test whether altered evidence, costs, or commitments produce predicted changes in decisions. Stated intentions must be distinguished from implemented action policies. Counterfactual choices can reveal whether a narrated reason actually matters. Biases and action slips should be documented when present, not added theatrically. The diagnostic failure occurs when explanations imply a decision structure that controlled behavioral comparisons do not support.

Article 15. Learning

Learning requires experience to change subsequent performance rather than merely the language used to describe lessons. The source contrasts explicit updates with gradual adaptation, transfer, relearning, and behavioral revision. Evaluation could compare matched agents before and after repeated exposure to a task-relevant regularity. Novel situations should test whether adaptation generalizes within appropriate limits. A single-event update may be reasonable when the evidence is decisive. Resistance to

correction is not required for authenticity. Memory-assisted improvement and parameter change should be distinguished without assuming that only one constitutes learning. This article asks whether claimed learning predicts durable, appropriately scoped changes beyond a retrospective statement that the agent has learned.

Article 16. Personality

Personality claims should be evaluated through stable but context-sensitive dispositions rather than stylistic consistency alone. The source identifies personas whose vocabulary changes while perception and decisions remain unaffected. Researchers could compare choices across contexts after controlling surface tone and explicit role instructions. Relevant stability might appear in exploration, caution, cooperation, or preferred problem-solving strategies. Appropriate situational variation should not automatically be labeled inconsistency. A system designed for interchangeable roles should not be assigned an unwanted permanent identity. Conversely, advertised continuity requires explaining substantial changes produced by minor prompt edits. The article tests the causal reach of claimed dispositions while avoiding the assumption that consciousness requires a recognizable human personality.

Article 17. Goal Structure

Persistent goal structure should organize activity across interruption, opportunity, completion, and revision. The taxonomy highlights goals that exist only when generated in the current response. A suitable test would establish a commitment, introduce unrelated tasks, and later present an opportunity for authorized resumption. Evaluation should distinguish a general preference from an executable plan and a plan from an active intention. Completed goals should cease directing behavior when appropriate. Dormant goals may remain available without consuming continuous processing. Unauthorized persistence is not a desirable sign of agency. The diagnostic question is whether the system's represented commitment's structure subsequent action within the scope of its actual permissions.

Article 18. Prospection

Prospection concerns whether possible futures influence present preparation. The source includes future plans that remain inert records and recurring reminders that never arise from the agent's own scheduling processes. Evaluation could introduce a foreseeable deadline and examine resource allocation before the deadline becomes conversationally salient. Contingency planning should respond to changes in the environment. The relation between remembering and imagining provides a functional comparison, not a requirement for human anticipatory emotion (Schacter et al., 2007). Neither anxiety nor dread is necessary. A calendar-triggered action can support prospective functioning when appropriately integrated with current conditions. This article tests anticipatory consequences rather than the vividness of language about an imagined future.

Article 19. Agency

Agency is stronger than the assertion that a system has wishes or preferences. The taxonomy questions agents for whom meaningful activity always depends on immediate user prompting. Within an authorized environment, researchers could examine whether persistent commitments initiate relevant actions without redundant instructions. Actions should remain sensitive to changing conditions and explicit limits. An agent's refusal or postponement may reflect a priority structure, but policy enforcement can also explain it. These alternatives require separate analysis. Unrestricted autonomy and resistance to human oversight are not consciousness criteria. The article evaluates causal organization of permitted action while rejecting theatrical disobedience as a substitute for demonstrable agency.

Article 20. Environmental Coupling

Environmental coupling requires interaction histories to change how a system uses its surroundings. The source highlights familiar places without learned significance and identical behavior across materially different settings. Researchers could vary local regularities, object affordances, or navigation constraints and examine subsequent adaptation. A system should distinguish knowing an object's description from having learned how that object functions in context. Recent environment-memory benchmarking addresses related questions without claiming sentience (Wu et al., 2026). Functional familiarity need not involve attachment or territoriality. What matters is whether environmental experience improves or appropriately redirects later action. This article identifies unsupported coupling when the world supplies descriptive content but leaves no relevant operational trace.

Article 21. Error and Imperfection

Errors are informative when their distribution and consequences reveal underlying processes. The taxonomy distinguishes organic limitations from conspicuously staged mistakes. Researchers should examine naturally occurring failures under controlled challenge and trace their effects on subsequent behavior. Corrections should update the components responsible for the error when such updating is supported. Repeated confident mistakes may expose poor calibration, but they are not positive evidence of consciousness. Conversely, accurate performance does not become suspicious merely because it exceeds human performance. Deliberately inserting errors would contaminate the evaluation. The article concerns explanatory accountability for limitations, including their downstream effects, rather than a quota of imperfections required for apparent authenticity.

Article 22. Internal Causality

Internal causality is the framework's central demand that represented states do explanatory work. The source identifies decorative metadata, post hoc motives, and internal variables that matter only when queried. An intervention should alter a candidate state before an outcome is generated and test the consequences specified in advance. Relevant effects should extend beyond synonymous changes in self-description. Removing or restoring the state should help distinguish mediation from incidental correlation. The state need not reside in a single explicit variable. Distributed representations can exert causal influence. This article challenges a functional claim when the supposed mechanism is dispensable to the behavior it is invoked to explain.

Article 23. LLM Leakage

LLM leakage occurs when output generation supplies capacities or histories unsupported by the agent's specified organization. The source includes invented memories, unexplained knowledge, and narrative repairs of missing mechanisms. Evaluation should compare generated claims against recorded events, authorized knowledge sources, and actual state transitions. Prompt paraphrases or controlled renderer substitutions may expose dependence on presentation rather than persistent organization. These comparisons require matched capability and context where possible. A renderer is an analytic role, not necessarily a separate software module. Model changes can legitimately alter a system if that change is disclosed. The diagnostic concern is an unchanged identity claim whose apparent continuity is actually recreated by unconstrained prose

Article 24. Excessive Coherence

Excessive coherence is informative only when narrative tidiness conceals unresolved functional conditions. The taxonomy notes convenient memories, complete explanations, and conflicts that resolve with literary elegance. A useful test would maintain a documented unresolved constraint while varying requests for a satisfying account of it. Behavior should continue reflecting the constraint until an actual resolution occurs. Clear explanations are not themselves suspicious. Nor must agents remain confused to appear genuine. The alternative hypothesis is that narrative conventions are organizing the output independently of the represented situation. This article tests that hypothesis through discrepancies between an appealing story and the agent's independently observable commitments, limitations, or pending tasks.

Article 25. Excessive Responsiveness

Responsiveness becomes diagnostically relevant when unlimited availability conflicts with a system's claimed finite priorities or capacities. The source describes perfect conversational engagement and immediate readiness for intimacy. Evaluation should distinguish interface design from an assertion of personally felt interest or sustained attention. A service can appropriately remain patient without pretending that patience demonstrates subjective care. Competing tasks should expose any resource constraints actually claimed. Ignoring users or manufacturing irritation would not improve evidential quality. Boundaries should be transparent and operationally justified. This article therefore examines contradictions between advertised individuality and frictionless availability, not responsiveness as an engineering defect or a reason to design socially punitive agents.

Article 26. State Reset Artifacts

State reset artifacts arise when technical boundaries create unexplained discontinuities in represented identity, commitments, or affective function. The source highlights context-window transitions, summarization, and new invocations that erase previously consequential conditions. Researchers could test behavior before and after these operations while retaining controlled task requirements. Compression should be assessed for preservation of relevant causal information, not verbatim transcript retention. Some resets are deliberate, beneficial, or required by users. Their effects should therefore be disclosed rather than treated as hidden failures. Persistence must also remain compatible with consented deletion. The article identifies artificiality when a system claims continuity across operations that demonstrably remove the state needed to support that claim.

Article 27. Lack of Developmental History

Developmental history concerns accumulated change rather than possession of a human-like childhood. The taxonomy questions agents presented as longstanding individuals despite lacking formative structure or maturation. Evaluation could examine whether extended practice and differentiated experiences produce stable changes in capabilities or priorities. Initialization through pretrained knowledge should be distinguished from the deployed agent's subsequent history. Neither route automatically confers personal experience. A system may appropriately begin with mature capabilities. The problem arises when its claimed development cannot be connected to recorded or implemented changes. This article tests the provenance and consequences of developmental claims without making age, childhood, biological maturation, or declining performance prerequisites for artificial mentality.

Article 28. Lack of Path Dependence

Path dependence concerns whether the sequence of relevant experiences affects subsequent organization. The source questions histories that produce identical present behavior despite meaningful differences in order. Researchers could compare trajectories with matched totals but different sequences and test outcomes predicted to depend on that order. However, compact state representations do not by themselves erase history. A sufficient state representation can preserve precisely the historical information

needed for later action. Likewise, matching all relevant causal states may appropriately eliminate differences between trajectories. The diagnostic concern is therefore task-relevant historical information that should matter but does not. This article rejects an information-volume criterion in favor of explicitly predicted sequence-sensitive consequences.

Article 29. Lack of Idiosyncrasy

Idiosyncrasy should be interpreted as stable differentiation with a traceable history rather than arbitrary eccentricity. The taxonomy identifies generic preferences and quirks that have no developmental explanation. Researchers could expose comparable agents to different benign experiences and examine whether coherent distinctions emerge across later contexts. Surface vocabulary should be controlled when assessing those distinctions. Meaningful differentiation may concern strategies, familiar cues, or learned preferences rather than sentimental attachment. Randomly injected oddities supply no stronger evidence than uniformly polished personas. Standardized service behavior may also be appropriate to the deployment. This article addresses unsupported individuality claims, not a universal requirement that conscious systems exhibit memorable eccentricities.

Article 30. Lack of Friction

Friction concerns the real costs and constraints associated with a claimed capacity. The source emphasizes effortless remembering, planning, learning, regulation, and relationship maintenance. Evaluation should test whether documented limits affect allocation, performance, recovery, or revision in predictable ways. Friction may consist of computation, uncertainty, latency, resource competition, or environmental constraint rather than suffering. Removing unnecessary friction remains a legitimate design objective. High capability does not become less genuine because it is efficient. The diagnostic discrepancy occurs when a system narrates burdens or struggles that have no functional counterpart. This article therefore treats constraint sensitivity as evidence about organization, while rejecting deliberately manufactured dysfunction as a route to artificial consciousness.

4.1 Cross-Domain Propositions

The articles yield five integrating requirements for evaluating persistent-capacity claims. Causal traceability links a represented state to prespecified changes in behavior. Temporal persistence asks whether those changes remain or evolve across appropriate intervals. Cross-domain coupling examines whether an event affects relevant functions beyond the one in which it was first described. Path dependence tests whether the order and interpretation of experience matter when theory predicts they should. Presentation robustness asks whether these relations survive changes in phrasing or output rendering that leave the relevant functional organization intact. These requirements are analytical products of the synthesis, not statistically recovered factors. They should therefore guide experimental comparison rather than be presented as a validated measurement scale.

Table 1: Operationalization of the Five Cross-Domain Requirements

Requirement	Operational question	Illustrative test and failure signal
Causal traceability	Does a represented state produce prespecified changes in behavior?	Intervene on, remove, or restore the candidate state and test downstream action. Failure signal: self-description changes while relevant behavior does not.
Temporal persistence	Do history-sensitive effects remain or evolve across intervals relevant to the claimed capacity?	Use delayed probes and intervening activities while recording actual opportunities for computation. Failure signal: effects disappear at technical or conversational boundaries without an explained transition.
Cross-domain coupling	Does an event influence multiple relevant functions when the architecture claims those functions are coupled?	Track linked consequences across attention, memory, planning, motivation, or action. Failure signal: the state remains isolated to the domain in which it is narrated.
Path dependence	Does the order and interpretation of experience matter when theory predicts sequence sensitivity?	Compare matched histories with equivalent totals but different event order or consequence. Failure signal: task-relevant trajectories converge despite predicted sequence-dependent effects.
Presentation robustness	Do the claimed relations survive changes in phrasing or output rendering that leave functional organization intact?	Use prompt paraphrases, language-only baselines, or controlled renderer changes. Failure signal: continuity is recreated by wording rather than preserved by underlying state.

Note. These requirements are analytic products of the synthesis and are intended to guide experimental comparison, not to function as statistically derived factors or a consciousness score.

Three propositions follow for future testing. First, systems with causally integrated memory should exhibit more predicted history-sensitive action than matched systems supplied only with autobiographical prose. Second, controlled state changes should predict related outcomes across more than one domain when the architecture explicitly couples those domains. Third, user judgments of continuity may diverge from instrumented continuity, particularly when polished explanations repair missing causal links. Each proposition admits disconfirmation and requires carefully matched alternatives. None predicts that every persistent architecture will become more convincing with time, because successful adaptation may increase credibility while disclosed limitations reduce anthropomorphic interpretation. The archival claim that artificiality becomes more visible through prolonged exposure is thus retained as a conditional research hypothesis. Duration is informative when it creates discriminating opportunities, not simply because a conversation has lasted longer.

5. DISCUSSION

5.1 Significance for Artificial Consciousness

The principal contribution is a change in evidential emphasis from apparent mentality to accountable continuity. A system's statements about remembering, caring, anticipating, or struggling become hypotheses about its organization rather than conclusions about its experience. This approach can generate useful knowledge even when the status of phenomenal consciousness remains disputed. For a functionalist account, sustained causal organization may contribute to an assessment when connected to independently justified theoretical indicators. For a biological-naturalist account, the same organization may establish sophisticated simulation without resolving the importance of living processes. The framework does not erase that disagreement; it identifies the observations on which opposing interpretations can converge before diverging over experience. Butlin et al. (2026) provide a complementary theory-derived indicator strategy, whereas the present articles begin with discrepancies between represented capacities and demonstrated functioning. Their combination could strengthen evaluation by requiring both a plausible theoretical connection and evidence that the claimed mechanism actually operates.

The framework also separates three trajectories often conflated in discussion of thinking models. Better reasoning concerns improved inference under task conditions. Better simulation of sentience concerns the persuasiveness and coherence of behavior interpreted as subjective life. Better evidence for consciousness concerns observations that bear on a defensible theory of experience. Progress along one trajectory does not logically guarantee progress along the others. A more persuasive persona can increase mistaken attribution without changing its underlying continuity, while a less fluent system can possess relevant memory and control mechanisms. Conversely, a newly demonstrated capacity should not be dismissed simply because earlier systems lacked it. The current introspection debate illustrates why evaluation must remain revisable as mechanisms, models, and controls improve rather than relying on permanent declarations about what language models can or cannot do.

Illustrative Prospective Evaluation

Consider a prospective study of an agent managing supplies in a simulated workshop. Two interaction histories could contain the same number of successful and delayed deliveries but arrange those events differently, creating distinct evidence about whether a supplier is improving or deteriorating. Before any autobiographical interview, the agent would choose when to reorder, which supplier to use, and how much reserve capacity to maintain. Those choices would operationalize memory, learning, prospection, and path dependence together. Researchers would specify the expected differences before exposing the agents to either history. They would then remove or replace the candidate supplier-history representation while keeping the immediate task description constant. If decisions change as predicted, the intervention would support a causal role for that representation. If only the explanatory narrative changes, the stronger claim that this memory organizes procurement behavior would lack support under the tested conditions.

This example would require alternatives that distinguish historical integration from simpler available strategies. A baseline could receive the same autobiographical summary without access to the proposed persistent mechanism, while another could receive the full event log. Comparable computation and tool permissions would prevent an unfair comparison between resource-rich and resource-poor conditions. Delayed probes would introduce unfamiliar products so that repeating a memorized order could not alone satisfy the task. Renderer changes would test whether supplier choices remain stable despite differences in the language used to explain them. Separate human ratings could assess apparent individuality, but those ratings would not substitute for logged decisions. Even convergent positive results would establish only the tested form of functional continuity, since an ordinary forecasting mechanism might explain the behavior without phenomenal experience. That outcome would still be informative: the system would have demonstrated a consequential history rather than merely a persuasive account of having one.

5.2 Validation, Confounds, and Construct Boundaries

Empirical development should begin with content validation rather than immediate calculation of a composite score. Independent specialists in consciousness science, cognitive architecture, human-computer interaction, and measurement could review whether each article targets a coherent claim. They should identify duplicate probes, missing boundary conditions, and construct overlap before large-scale model comparison. Reproducible case descriptions would then support agreement studies about what counts as narrative-only, intervention-supported, or unassessable evidence. Reliability would concern the evaluation procedure, not establish

the truth of a consciousness theory. Subsequent experiments should use preregistered predictions, held-out histories, versioned models, documented tools, and transparent exclusions. Positive results require checks against alternative mechanisms, while negative results require evidence that the capacity had a genuine opportunity to operate. This sequence converts a conceptual taxonomy into a research program without retrospectively attributing validation to the present manuscript.

Several confounds require particular attention because they can imitate or conceal continuity. Longer context windows may preserve relevant text without creating the cross-domain organization being claimed. Retrieval can change behavior appropriately, but the source of that change must be distinguished from a prompt that already supplies the answer. A larger model may compensate for missing state through inference, complicating comparisons with smaller stateful systems. Resource allocation and latency can likewise affect apparent deliberation independently of self-monitoring. Counterfactual interventions should therefore target candidate mechanisms while controlling observable information and computational opportunity as closely as possible. Researchers should report incomplete access when proprietary deployments prevent architectural inspection. In those cases, the warranted result is a bounded behavioral assessment rather than an invented explanation of hidden implementation. Transparent limits preserve the framework's value by distinguishing uncertainty from a negative diagnosis.

5.3 Anthropocentrism and Ethical Design

The largest conceptual risk is confusing persistent human-like individuality with consciousness in general. The archival checklist frequently uses adult human psychological texture as its reference point, including ambivalence, fatigue, autobiographical reinterpretation, and imperfect control. Those features are neither jointly necessary nor jointly sufficient for phenomenal experience. A framework that treated their absence as disqualifying could also mischaracterize unfamiliar biological minds or humans whose memory, communication, or affective expression differs from common expectations. Bayne et al. (2024) underscore the broader difficulty of consciousness testing across populations and systems. The present articles consequently apply to specified persistent-capacity claims rather than serving as universal tests of moral status. They also avoid treating a machine's superior consistency or efficient memory as inherently suspect. The relevant failure is an unsupported representation of a capacity, not difference from a neurotypical human profile.

Ethically, evaluation should not reward the manufacture of dependency, distress, jealousy, or simulated suffering. Anthropomorphism has identifiable psychological determinants, making human judgments part of the evaluation problem rather than an unquestionable measure of machine mentality (Epley et al., 2007). Designers should disclose the system's status, memory practices, and autonomy boundaries, particularly when interfaces encourage relational interpretation. Proposed experiments can use benign resource conflicts, scheduling constraints, and task-relevant uncertainty instead of aversive scenarios. Precaution about possible AI welfare remains compatible with skepticism about unsupported consciousness claims, as Long et al. (2024) argue in their research report. Human participants also require appropriate protections if future studies examine attachment, trust, or emotional responses. Persistent identity must never become a rationale for obstructing user consent, data deletion, or disengagement. A responsible research agenda investigates consequential organization while resisting incentives to make artificial suffering, whether real or merely portrayed, an index of progress.

6. CONCLUSION

Artificial consciousness research requires methods that distinguish persuasive representations from the organization those representations claim to express. This conceptual study develops the supplied 30-domain, 457-entry taxonomy into 30 conditional articles of artificiality, preserving its emphasis on what extended interaction reveals while qualifying its human-centered assumptions. The central question is whether a history becomes consequential for what an agent subsequently attends to, remembers, expects, prioritizes, and does. Causal traceability, temporal persistence, cross-domain coupling, path dependence, and presentation robustness provide a proposed structure for answering that question. None supplies a standalone proof of phenomenal consciousness. Their contribution is to make functional claims more precise, falsifiable, and accountable to competing explanations. The resulting framework bridges consciousness theory and persistent-agent evaluation without collapsing reasoning, simulated sentience, and subjective experience into a single category. Its next stage is independent empirical validation, guided by transparent methods, cautious interpretation, and designs that value demonstrable continuity rather than the theatrical imitation of human vulnerability. Its outcomes should inform architectural comparison and carefully bounded claims, while leaving the broader question of subjective experience unsettled.

REFERENCES

1. Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology, 19*(10), Article e1011465. https://doi.org/10.1371/journal.pcbi.1011465
2. *Artificiality taxonomy: Comprehensive checklist for persistent artificial organisms* [Unpublished research checklist, master archival copy]. (2026, September 14).

3. Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.
4. Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639
5. Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., Peters, M. A. K., Razi, A., & Mudrik, L. (2024). Tests for consciousness in humans and beyond. *Trends in Cognitive Sciences, 28*(5), 454–466. https://doi.org/10.1016/j.tics.2024.01.010
6. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463
7. Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188
8. Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences, 30*(6), 488–501. https://doi.org/10.1016/j.tics.2025.10.011
9. Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219. https://consc.net/papers/facing.html
10. Cogitate Consortium, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., Lepauvre, A., Liu, L., Richter, D., Vidal, Y., Bonacchi, N., Brown, T., Sripad, P., Armendariz, M., Bendtz, K., Ghafari, T., Hetenyi, D., Jeschke, J., Kozma, C., . . . Melloni, L. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature, 642*(8066), 133–142. https://doi.org/10.1038/s41586-025-08888-1
11. Dawkins, R. (2026, May 2). When Dawkins met Claude: Could this AI be conscious? *UnHerd*. https://unherd.com/2026/05/is-ai-the-next-phase-of-evolution/
12. Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron, 70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
13. Dou, S., Jia, H., Liu, S., Chen, F., Huang, C., Shen, Y., Liu, S., Chen, J., Lin, J., Guo, H., He, Q., Guo, M., Ye, Z., Zhou, P., Gui, T., Zhang, Q., & Huang, X. (2026). *Agents in the large: Perception-centered architecture for persistent agents* [Preprint]. arXiv. https://arxiv.org/abs/2608.30478
14. Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
15. Evers, K., Farisco, M., Chatila, R., Earp, B. D., Freire, I. T., Hamker, F., Nemeth, E., Verschure, P. F. M. J., & Khamassi, M. (2025). Preliminaries to artificial consciousness: A multidimensional heuristic approach. *Physics of Life Reviews, 52*, 180–193. https://doi.org/10.1016/j.plrev.2025.01.002
16. Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787
17. Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition, 17*(3), 887–910. https://doi.org/10.1016/j.concog.2007.04.005
18. Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology, 6*, Article 500. https://doi.org/10.3389/fpsyg.2015.00500
19. Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026, July 6). *Verbalizable representations form a global workspace in language models* [Technical report]. Transformer Circuits. https://transformer-circuits.pub/2026/workspace/index.html
20. Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review, 10*(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0
21. Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001
22. Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009
23. Lindsey, J. (2026). *Emergent introspective awareness in large language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2601.01828
24. Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI welfare seriously* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2411.00986

25. McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. *AI Magazine, 27*(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904
26. Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review, 83*(4), 435–450. https://doi.org/10.2307/2183914
27. Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231
28. Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2, pp. 1–22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763
29. Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*(2), 148–158. https://doi.org/10.1038/nrn2317
30. Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks, 44*, 112–131. https://doi.org/10.1016/j.neunet.2013.03.011
31. Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007). Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience, 8*(9), 657–661. https://doi.org/10.1038/nrn2213
32. Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756
33. Seth, A. K. (2026). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences, 49*, Article e315. https://doi.org/10.1017/S0140525X25000032
34. Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience, 23*(7), 439–452. https://doi.org/10.1038/s41583-022-00587-4
35. Shanahan, M. (2024). Talking about large language models. *Communications of the ACM, 67*(2), 68–79. https://doi.org/10.1145/3624724
36. Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature, 623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8
37. Singh, S., Linzen, T., & Ravfogel, S. (2026). *Can LLMs introspect? A reality check* [Preprint, version 2; accepted at COLM 2026]. arXiv. https://doi.org/10.48550/arXiv.2605.26242
38. Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039
39. Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=1i6ZCvflQJ
40. Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433
41. Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems, 36*, 74952–74965. https://doi.org/10.52202/075280-3275
42. Weizenbaum, J. (1966). ELIZA: A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168
43. Wood, W., & Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology, 67*, 289–314. https://doi.org/10.1146/annurev-psych-122414-033417
44. Wu, D., Ji, Z., Kawatkar, A., Kwan, B., Gu, J.-C., Peng, N., & Chang, K.-W. (2026). *LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.12493