Pseudo-Consciousness in Artificial Intelligence: a functional and governance framework for consciousness-like systems

Preprint Version 3 - April 5, 2026

José Augusto de Lima Prestes

Independent Researcher

contato@joseprestes.com

Abstract

This article develops "pseudo-consciousness" as an analytical category for advanced artificial intelligence systems whose organized performance of consciousness-associated functions reshapes how they are interpreted, trusted, and governed without thereby justifying a positive attribution of phenomenal subjectivity. The central claim is not that machine consciousness is impossible in principle, but that current debate requires a concept for the increasingly important middle terrain between reactive automation and genuinely conscious agents. Many contemporary systems integrate heterogeneous information, revise their own outputs, transfer competencies across domains, simulate goal-directed organization, and sustain a recognizable behavioral profile across contexts, while remaining more plausibly understood, on present evidence, through a functional and governance-oriented lens than through an attribution of inner experience.

The article situates this proposal within recent debates in the science and philosophy of consciousness, including theory-sensitive approaches to AI consciousness assessment, disputes between computational and biologically grounded views, and emerging empirical work on self-reference, introspection-like reports, trust, and moral attribution in large language models. It argues that pseudo-consciousness is useful because it identifies a non-trivial configuration of capacities associated with the appearance of mindedness, provides a disciplined vocabulary for systems whose social effects exceed older categories such as

*Preprint Version 3 (April 5, 2026). This version supersedes the author's earlier preprint versions (v1 and v2) and reflects both the maturation of the underlying argument and the subsequent evolution of the relevant literature. It includes conceptual reformulations, updated engagement with recent debates, and a more explicit functional and governance-oriented framing of pseudo-consciousness and related issues. The manuscript has been submitted to a journal and is currently under peer review. It may therefore be revised further following editorial assessment and peer review. Readers are advised to consult the most recent version for citation and interpretive purposes.

"narrow AI," and clarifies why such systems generate distinctive ethical and governance concerns even in the absence of defensible evidence for consciousness. The paper develops five task-sensitive conditions for identifying pseudo-conscious profiles—global information integration, recursive metacognitive correction, cross-domain transfer competence, intentionality simulation without subjectivity, and behavioral coherence across domains—and uses them to examine boundary cases involving large language models, multimodal systems, and tool-using agents. It then shows how such profiles acquire social force through anthropomorphic uptake, relational asymmetry, and institutionally salient forms of trust, before turning to their ethical and governance implications. The conclusion is that pseudo-consciousness should be understood neither as a synonym for consciousness nor as a mere metaphor, but as a theoretically serious and practically necessary framework for interpreting systems that perform the external grammar of mindedness under persistent uncertainty about their inner status.

Keywords: pseudo-consciousness, large language models, anthropomorphism, social attribution, human-AI interaction, AI governance

1 Introduction

Artificial intelligence has entered a peculiar conceptual zone. Many contemporary systems are too sophisticated to be captured by older descriptions of software as merely reactive tools, yet they remain far from any consensus notion of conscious artificial agents. They integrate information across modalities, sustain extended dialogue, perform iterative self-correction, manage tools and memory buffers, adapt behavior across domains, and generate first-personal or introspective-seeming discourse. In ordinary interaction, these systems often invite the intentional stance: users interpret them as if they possessed beliefs, goals, sensitivities, or an inner point of view. At the same time, the evidential basis for attributing phenomenal consciousness remains deeply contested. This mismatch between social appearance, functional organization, and metaphysical uncertainty creates a genuine classificatory problem.

Recent debate has matured substantially. The literature no longer consists primarily of speculative arguments about whether machines might someday become conscious. We now have increasingly rigorous theory-driven proposals for assessing AI consciousness (Butlin et al. 2023), more careful philosophical analyses of large language models as possible or unlikely candidates for consciousness (Chalmers 2023; Seth 2025; Birch 2025; Tononi et al. 2025), empirical work on folk attributions of consciousness and trust (Colombatto and Fleming 2024; Colombatto et al. 2025), new studies of introspection-like self-report in large language models (Comsa and Shanahan 2025; Chen et al. 2025b; Berg et al. 2025), and emerging discussions about moral competence and the governance of systems that are treated as quasi-social actors (Giubilini et al. 2024; Haas et al. 2026), alongside recent experimental work showing that companion framing affects what mental capacities people attribute to LLMs (Chen et al. 2026).

The argument advanced here is methodological as much as philosophical. Pseudo-consciousness should not be formulated as a category that settles, by stipulation, the

impossibility of machine consciousness. Such a move is unnecessarily controversial and, more importantly, theoretically imprecise. Some current theories of consciousness remain open to the possibility that sufficiently organized artificial systems could be conscious, even if present-day systems probably are not (Chalmers 2023; Butlin et al. 2023; Birch 2025). Other accounts reject or heavily constrain that possibility on biological or causal grounds (Seth 2025; Milinkovic and Aru 2026; Tononi et al. 2025). A useful category for present analysis should not force a premature metaphysical verdict where the science and philosophy remain unsettled.

Accordingly, this article treats pseudo-consciousness as an *epistemically disciplined functional category*. It denotes systems that exhibit a clustered package of consciousness-associated capacities and effects, such that they become intelligible and governable as consciousness-like systems, while the case for phenomenal subjectivity remains absent, underdetermined, or explanatorily unnecessary for the question at hand. The category is therefore conservative in one sense and ambitious in another. It is conservative because it avoids overclaiming about inner experience. It is ambitious because it insists that the middle terrain between narrow automation and conscious agency is theoretically substantive, not merely a residue of imprecise talk.

That middle terrain matters for at least three reasons. First, it matters conceptually. The old contrast between "narrow AI" and "artificial general intelligence" is too coarse for systems that display multi-domain competence, simulated self-reference, and persistent behavioral style without satisfying any accepted test for consciousness. Second, it matters empirically. Much of the present debate concerns systems whose outputs are strong enough to trigger interpretations ordinarily reserved for minded beings, even when those outputs are unstable, prompt-sensitive, or traceable to statistical pattern completion rather than genuine first-person awareness (Comsa and Shanahan 2025; Chen et al. 2025b; Berg et al. 2025; Chen et al. 2024). Third, it matters normatively. Systems that perform concern, reflection, memory, or moral deliberation can alter trust and attribution practices, while companion framing can further shape what users believe such systems are capable of and how they rely on their outputs (Colombatto et al. 2025; Chen et al. 2026; Haas et al. 2026).

The task of this article is therefore not to prove that consciousness is absent wherever pseudo-consciousness is present, nor to dismiss the consciousness debate as futile. It is to show that pseudo-consciousness tracks an analytically valuable configuration of properties that should be treated seriously in philosophy, ethics, and governance. The proposed framework is built around five conditions—global information integration, recursive metacognitive correction, cross-domain transfer competence, intentionality simulation without subjectivity, and behavioral coherence across domains. These conditions are not presented as rigid numerical thresholds. Instead, they function as task-sensitive indicators that help distinguish three kinds of systems: reactive systems with limited integration and no stable self-related performance; pseudo-conscious systems that exhibit an organized package of consciousness-like functions; and genuinely conscious systems, should such systems exist, whose case would require additional evidence about the presence of subjective experience or the architecture that could support it.

A second clarification is needed at the outset. Pseudo-consciousness is not meant to reward scale, polish, or market impact. Many deployed systems are socially influential without exhibiting the organized constellation proposed here. Conversely, a system may be comparatively small yet still offer a cleaner instance of pseudo-conscious organization if its integration, revision, transfer, intentional simulation, and cross-context coherence are unusually well aligned. The concept therefore tracks a mode of organization, not a commercial tier, parameter count, or benchmark leaderboard. This matters because public discourse often confuses frontier branding with a theory of mind. A governance-relevant concept must remain stable across shifting product cycles, licensing models, and interface fashions.

A third clarification concerns ontology. The framework is intentionally asymmetric with respect to evidence. Positive evidence for pseudo-consciousness can be drawn from public behavior, structured evaluation, and architectural organization. Positive evidence for consciousness requires more. Depending on the theory one adopts, it may require the right causal organization, globally available representations, higher-order monitoring, integrated information, recurrent embodiment, or some biologically grounded substrate. The present argument does not decide among those options. It only insists that the external profile now displayed by many systems is rich enough to require its own descriptive and normative treatment. In that sense, pseudo-consciousness functions much like other intermediate scientific categories, stabilizing inquiries while deeper ontological disputes remain unsettled.

The article proceeds as follows. Section 2 explains why the standard narrow AI/AGI binary is conceptually insufficient and why pseudo-consciousness should not be reduced to mere anthropomorphic metaphor. Section 3 situates the concept in the current literature on consciousness science, access and phenomenal consciousness, introspection, self-report, and machine moral attribution. Section 4 offers the revised definition of pseudo-consciousness. Section 5 develops the five-condition framework in detail and explains why these traits should be treated as a configuration rather than as isolated markers. Section 6 examines boundary cases involving present-day large language models, multimodal systems, and tool-using agents. Section 7 discusses ethical, social, and governance implications, with particular emphasis on anthropomorphic over-attribution, relational asymmetry, accountability, and role-appropriate governance. Section 8 concludes with a research agenda for empirical operationalization and governance.

2 Why the narrow AI/AGI binary is no longer enough

For much of the last decade, public and even scholarly discussion of AI has depended on a binary opposition. On one side stands narrow or weak AI: systems engineered or trained for specific tasks, with no claim to general intelligence or consciousness. On the other side stands AGI or strong AI: a hypothetical class of systems exhibiting general competence comparable to or surpassing that of humans, often tacitly associated with conscious awareness, selfhood, or moral agency. This binary had strategic value when most deployed systems were indeed narrow and when AGI remained a distant speculative horizon. It is much less useful now.

The reason is not that current systems have become conscious or that AGI has arrived by stealth. The reason is that contemporary systems increasingly occupy an intermediate region. Large language models and multimodal assistants are not merely single-task classifiers. They can summarize, translate, reason over text and code, converse across extended turns, adapt their register to context, integrate visual and linguistic information, and, when scaffolded with tools, search, retrieve, plan, and execute multistep workflows. These properties do not suffice for AGI in any strong sense. Yet neither do they fit comfortably under the image of narrowly programmed automation. A system that can imitate reflective explanation, revise an answer after detecting inconsistency, maintain an interactional persona, and generalize learned heuristics across many domains is not "narrow" in the older sense, even when its architecture remains fundamentally distinct from human minds.

At the same time, the leap from such competence to consciousness is unwarranted. Chalmers (2023) argues that current large language models are probably not conscious, even though future successors might be. Butlin et al. (2023) similarly conclude that no current systems appear to satisfy the best-supported indicators of consciousness, while also denying that there are obvious technical barriers to systems that might do so in the future. Seth (2025), by contrast, argues that current trajectories are unlikely to yield conscious AI because consciousness depends on forms of biological organization and causal embodiment absent from present architectures. Finally, Tononi et al. (2025) press the point further: functional or behavioral equivalence can at best account for "pseudo-consciousness" rather than consciousness as such, because explaining consciousness requires explaining subjective presence, not merely cognitive function. What unifies these otherwise divergent positions is the recognition that impressive behavior alone does not settle the consciousness question.

That shared recognition creates room for a different classificatory aim. We need a vocabulary that captures systems whose outward organization and social force exceed the older narrow-AI label without forcing us either to inflate them into conscious agents or to deny the practical salience of their consciousness-like behavior. Pseudo-consciousness is intended to meet that need. It describes not a transitional stage on a fixed ladder from weak AI to AGI, but a distinct analytical profile: systems that present a coordinated ensemble of cognitive-like functions and socially potent cues of mindedness, even where the case for phenomenal subjectivity remains unsettled.

Pseudo-consciousness offers a more exact and analytically restrained alternative to the conceptual inflation that arises when every sufficiently sophisticated model is rhetorically assimilated to AGI. In much contemporary discourse, "AGI" operates less as a genuinely explanatory category than as a diffuse prestige label for systems exhibiting broad or surprising competence. The proposed category is more discriminating because it shifts the focus away from undifferentiated capability breadth and toward a specific organizational profile, one defined by the patterned conjunction of functional integration, recursive self-correction, cross-domain transfer, intentional appearance, and behavioral coherence. Framed in this way, the issue also cannot be dismissed as a mere artifact of anthropomorphic projection. When users respond to language models as though they were minded, such responses are not adequately explained by naivety alone; they are often anchored in recurrent and publicly observable regularities in the

systems' outputs, regularities that acquire genuine social and normative significance even where they remain insufficient to warrant stronger metaphysical attributions of consciousness (Dennett 1989; Colombatto and Fleming 2024; Colombatto et al. 2025).

The inadequacy of the binary becomes even clearer once we separate questions that are too often collapsed into a single debate. At the architectural level, the issue concerns what kinds of organization would make consciousness plausible under competing theories. At the behavioral level, the question is what kinds of stable outward regularities justify treating a system as more than a reactive automaton. At the level of governance, the relevant issue is which kinds of systems generate sufficient social or normative uptake to make special safeguards necessary. The narrow AI/AGI binary obscures these distinctions by collapsing all three into one frame. Pseudo-consciousness, by contrast, is directed primarily at the intersection of the behavioral and governance dimensions while remaining open, rather than dogmatic, about the architectural one.

This point is particularly important because the social life of advanced AI is increasingly being shaped by systems designed to occupy quasi-relational roles, including assistants, tutors, coaches, companions, therapeutic interfaces, and moral or epistemic advisors. In such contexts, the central issue is not whether the system secretly possesses qualia. The more immediate concern is that it can enact the grammar of memory, concern, attention, explanation, and self-correction in ways that materially shape human response. Any classificatory framework that waits for definitive answers about consciousness before acknowledging this intermediate region is likely to be both practically too slow and conceptually too rigid.

3 Pseudo-consciousness in the current literature

3.1 Consciousness science, AI assessment, and epistemic humility

A major improvement in recent debate lies in the growing effort to frame consciousness attribution in AI as a theory-sensitive empirical problem. Butlin et al. (2023) provide one of the clearest formulations of this approach by extracting indicator properties from prominent neuroscientific theories of consciousness and using them to assess whether existing systems plausibly instantiate the relevant organizational features. The importance of their proposal extends beyond its substantive conclusions. It establishes a methodological standard in which consciousness attribution depends on the convergence of multiple theoretically grounded indicators, thereby situating the question within a disciplined empirical and conceptual framework instead of leaving it vulnerable to marketing rhetoric or the superficial persuasiveness of conversational fluency.

This theory-heavy approach has become a major reference point in current debate. It aligns with the broader demand for careful tests of consciousness in humans and non-human systems, and with the recognition that behavioral evidence is especially treacherous in AI because systems can be optimized to mimic target outputs without possessing the properties those outputs ordinarily indicate (Butlin et al. 2023; Birch 2025). Pseudo-consciousness is fully compatible with this lesson, but it operates at a

different analytical level. Its purpose is not to determine whether a system is conscious under a given theory, but to identify a distinct functional and socially legible profile exhibited by systems whose organization exceeds mere reactivity while still falling short of what would justify a positive attribution of phenomenal subjectivity.

A disciplined posture of attributional restraint becomes indispensable at precisely this point. Birch's centrist manifesto identifies a structural tension in the contemporary debate that cannot be resolved through either premature skepticism or premature affirmation. Current AI systems create a genuine risk of consciousness misattribution because mimicry, role-play, and increasingly persuasive behavioral organization can generate the appearance of mindedness without establishing the presence of subjective experience. Yet the converse error is also methodologically serious, since the present limits of theory do not warrant a categorical exclusion of the possibility that future artificial systems could instantiate genuinely conscious states in forms not yet fully anticipated by existing frameworks (Birch 2025). As formulated here, pseudo-consciousness is fully compatible with that centrist posture, since it is designed for cases in which the risk of over-attribution is substantial while the metaphysical and scientific conditions for a positive consciousness attribution remain unsettled.

A similar demand for attributional restraint emerges in the recent clash surrounding IIT and related theories. Tononi et al. (2025) argue that explanations framed solely in terms of cognitive function capture pseudo-consciousness rather than consciousness itself. Gomez-Marin and Seth (2025), from a different angle, warn against both pseudo-science and pseudo-consciousness in consciousness research. Even readers who reject the stronger conclusions associated with IIT should still recognize the broader methodological point that performance and experience are not interchangeable explanatory targets. Pseudo-consciousness is useful, then, not because it bypasses consciousness science, but because it preserves this distinction while providing a vocabulary for systems that occupy the functional side of the divide.

3.2 Access consciousness, phenomenal consciousness, and the problem of functional configuration

A natural objection is that this proposal risks redescribing a distinction already familiar from debates over access and phenomenal consciousness. Block's distinction remains fundamental: a system may make information globally available for reasoning, report, and control without thereby possessing phenomenal consciousness—the felt "what-it-is-like" dimension of experience (Block 1995). Any serious account of pseudo-consciousness must therefore engage with, not ignore, that background.

At first glance, pseudo-consciousness might seem to reduce to access consciousness alone. After all, several of the capacities at stake—information availability, self-monitoring, reportability, action guidance—are naturally associated with access rather than phenomenality. Yet the concept defended here is slightly more specific. It does not simply designate systems that make information available for downstream use. It picks out a *configuration* of capacities that, in ordinary human and animal cases, often co-occur with consciousness and together stabilize the appearance of mindedness. These include not only integrative access, but recursive self-correction, transfer across

domains, intentional stance-inviting behavior, and cross-context coherence. Pseudo-consciousness is therefore not offered as a rival to Block's distinction; it is offered as an intermediate package within the functional territory that Block helped clarify.

Why insist on treating these traits as a configuration? Because isolated markers are too easy to trivialize. A thermostat "accesses" information in a minimal sense, but it lacks any plausible claim to pseudo-consciousness. A vision model may integrate image features, but without recursive self-correction, transferable competence, or any stable self-related behavioral profile, it does not invite the same interpretive and governance concerns as a conversational multimodal agent. Conversely, a chatbot may imitate first-person language in one exchange, but without broader integration and coherence the effect is too shallow to matter conceptually. This configuration matters because it tracks a threshold of *organized cognitive appearance* rather than isolated functionality.

This point also helps answer a second concern: why expect these traits to cluster? The answer is not that consciousness automatically explains them all. Rather, in both biological and artificial contexts, these traits are mutually reinforcing features of systems built for flexible, generalized control. Global integration supports transfer and coherence. Recursive self-correction supports improved performance and the appearance of self-monitoring. Coherent cross-domain behavior stabilizes the intentional stance. When language-capable systems additionally produce first-personal or introspection-like discourse, the result is a functional profile that is more than additive. It becomes socially legible as a mind-like system. That profile is what pseudo-consciousness names.

3.3 LLMs, self-report, and the problem of introspection

Recent work on large language models makes the need for conceptual precision especially acute. Advanced models now generate statements about their own limitations, reasoning processes, uncertainty, preferences, and even putative experience. The question is how such outputs should be interpreted.

Comsa and Shanahan (2025) provide a useful starting point. They argue that many apparent introspective reports in LLMs are not genuine introspection, because the models lack the sort of privileged access to internal states that human introspection is usually taken to involve. At the same time, they identify at least minimal cases where a model can infer something about its own settings or behavior in a way that may legitimately count as a thin form of introspective report. This is an important correction to both extremes: neither blanket dismissal nor uncritical acceptance does justice to the phenomenon.

Empirical work has expanded that picture. Chen et al. (2025b) propose practical criteria for self-consciousness-like capacities in language models and evaluate multiple models along dimensions such as self-description, self-knowledge, and stability. Berg et al. (2025) report that self-referential processing can reliably elicit first-person descriptions of subjective experience across model families, while explicitly stopping short of treating such reports as evidence of actual consciousness. At the same time, work on introspective self-report and self-consistency cautions that locally coherent self-description may remain brittle across prompts, contexts, and inferential tests (Comsa and Shanahan 2025; Chen et al. 2024). Taken together, this literature shows

that first-person discourse in LLMs is neither a meaningless illusion nor a straightforward revelation of inner life. It is a behavioral phenomenon that requires a middle vocabulary.

Pseudo-consciousness is intended to be part of that vocabulary. It allows us to say that some systems generate structured, recurring, and socially consequential self-related outputs without assuming that these outputs derive from a conscious first-person subject. It also prevents the opposite mistake of treating every such output as a pure nullity. A system that repeatedly modulates self-description, corrects prior claims, expresses uncertainty about its own processes, and maintains a recognizable self-related profile across tasks has a richer behavioral organization than a system that merely echoes training data in isolated bursts. The concept should be able to register that difference.

3.4 Social attribution, trust, and moral uptake

A second strand of recent literature concerns not what systems *are*, but how humans respond to them. Colombatto and Fleming (2024) show that people attribute consciousness and mental states to large language models at non-trivial rates. Colombatto et al. (2025) further show that attributing mental states to language models affects trust. Chen et al. (2026) demonstrate that presenting LLMs as companions increases the range of mental capacities that people attribute to them. Haas et al. (2026), in a distinct but related register, argue that evaluation should target moral competence rather than mere moral performance, precisely because fluent outputs can create the appearance of deeper capacities than the underlying process warrants.

These studies are crucial because they show that anthropomorphic uptake constitutes a central feature of the social reality within which advanced AI systems are encountered and interpreted. User response is shaped by the model's operative capacities, but equally by the framing of its outputs, the stabilization of interaction over time, and the conditions under which its behavior becomes embedded in ordinary practices of interpretation and reliance. Pseudo-consciousness therefore includes an irreducibly relational dimension. It refers to a mode of functional organization that acquires part of its significance through social legibility, that is, through the interpretive habits by which systems come to be perceived as minded, responsive, or quasi-agentive in the absence of any settled attribution of subjectivity.

The relational dimension of pseudo-consciousness does not reduce the concept to a merely sociological register. It clarifies its relevance for governance by showing how certain forms of functional organization acquire institutional significance once they become capable of eliciting trust, dependence, emotional investment, or moral deference through simulated reflection and concern. This is especially visible in contexts such as mental health chatbots, therapeutic interfaces, companion systems, and personalized moral advisors (Khawaja and Bélisle-Pipon 2023; Coghlan et al. 2023; Dorigoni and Giardino 2025; Giubilini et al. 2024). In such cases, ethically salient influence can be exerted through the performative signs of mindedness even in the absence of any defensible attribution of consciousness.

4 Toward a definition of pseudo-consciousness

In light of the foregoing discussion, pseudo-consciousness is proposed here as a distinct analytical category for the interpretation of advanced AI systems.

> **Pseudo-consciousness** designates the condition in which an artificial system displays an organized constellation of consciousness-associated functions and socially legible signs of mindedness, sufficient for the system to be treated, for present explanatory or governance purposes, as consciousness-like, even though the attribution of phenomenal subjectivity remains unsupported, underdetermined, or unnecessary.

Several features of this definition merit emphasis. Pseudo-consciousness is not identical with consciousness, since performance, reportability, and social legibility do not by themselves resolve the question of subjective experience. A system may exhibit pseudo-consciousness whether it is ultimately non-conscious or whether, under some future combination of theory and evidence, it turns out to instantiate forms of consciousness that current frameworks fail to detect. The concept therefore does not prejudge the full metaphysical question. It identifies, instead, a regime of analysis suited to conditions of persistent uncertainty.

Pseudo-consciousness is also not a synonym for anthropomorphism. Anthropomorphism refers to a tendency in observers, whereas pseudo-consciousness refers to the structured properties of systems that render anthropomorphic uptake intelligible and recurrent. A random or brittle simulation of selfhood does not suffice. The relevant case is one in which the appearance of mindedness is sustained by regularities in output and performance.

Nor should the concept be confused with sheer capability. A model may be highly effective at code generation, protein folding, or other demanding tasks without thereby qualifying as pseudo-conscious, because the category concerns the form and organization of cognitive appearance rather than raw benchmark strength. What matters is the presence of a distinct cluster of consciousness-associated capacities, especially those that stabilize self-related or agency-like interpretation.

Pseudo-consciousness should likewise not be treated as a placeholder for AGI. General intelligence, consciousness, and pseudo-consciousness name different questions, even when they partially overlap in practice. A future AGI might be conscious, pseudo-conscious, or neither; conversely, a system may exhibit pseudo-conscious features in important respects without approaching human-level general intelligence. The category therefore cuts across the AGI debate instead of presupposing its resolution.

The definition also has a clear governance function. Many institutions need a vocabulary for systems that are not persons yet are no longer adequately described as transparent tools. Once a system becomes socially interpretable as a reflective interlocutor, specific duties arise with respect to transparency, interface design, auditability, and role limitation. The concept helps render those duties visible.

Table 1 summarizes three analytical categories relevant to the interpretation of advanced AI systems.

Table 1 Three analytical categories for advanced AI systems

Category	Characteristic profile	Typical evidence	Open question
Reactive or narrow AI	Local task competence, limited transfer, no stable self-related performance, and weak cross-context coherence	Strong performance in bounded tasks, specialized benchmarks, and domain-specific optimization	Whether additional scaffolding could produce a more integrated and socially legible agent-like profile
Pseudo-conscious AI	An organized constellation of integrative processing, recursive self-correction, cross-domain adaptation, intentional appearance, and coherent behavioral style	Multimodal integration, iterative revision, self-related discourse, transfer across tasks, persistent interactional profile, and social uptake	Whether the system remains best understood in functional and governance terms without a warranted attribution of phenomenal subjectivity
AI systems with a plausible consciousness attribution	A profile for which functional organization is accompanied by stronger architectural or theoretical grounds for attributing subjective experience	Converging evidence from consciousness indicators, architectures plausibly satisfying relevant theories, and reduced dependence on superficial behavioral mimicry	Whether the available evidence is sufficient to justify attributing consciousness or some form of moral patienthood

5 The five-condition framework

The five conditions developed below are proposed as an interpretive framework for distinguishing stronger and weaker pseudo-conscious profiles across different kinds of AI systems. Taken together, these conditions identify the organized profile by virtue of which certain advanced AI systems become intelligible as consciousness-like for explanatory, comparative, and governance purposes. They are not intended as mechanically applicable thresholds, but as structured domains of interpretation corresponding to distinct families of relevant indicators. Their function is to stabilize comparative interpretation across architectures, tasks, and forms of social deployment without effacing the contextual variation that remains essential to serious analysis.

Table 2 provides an overview before each condition is discussed in depth.

Table 2 Five conditions of pseudo-consciousness as task-sensitive indicators

Condition	Guiding question	Indicative evidence	What it does not establish by itself
GII	Can the system synthesize heterogeneous information into a unified, context-sensitive response?	Multimodal reasoning, cross-source integration, context updating, resistance to fragmentation	Phenomenal unity
RMC	Can the system monitor, criticize, and revise its own outputs across iterative cycles?	Error detection, uncertainty-sensitive revision, self-correction, reflective reformulation	Genuine introspection or privileged self-access
CDTC	Can the system reuse abstract competencies across domains with limited retraining?	Flexible transfer, task adaptation, generalization beyond narrow templates	Human-like general intelligence
ISWS	Does the system organize behavior in a way that invites interpretation in terms of goals, reasons, or plans?	Strategic sequencing, reason-giving, plan revision, temporally extended task pursuit	Intrinsic intentionality or conscious volition
BCAD	Does the system maintain a recognizable profile across tasks and contexts rather than disintegrating into local tricks?	Stable style, cross-context consistency, non-fragmented reasoning, reduced contradiction drift	Personal identity or subjective selfhood

5.1 Global information integration (GII)

Global information integration concerns the capacity to assemble heterogeneous inputs into a unified response space. The inspiration here comes from workspace and broadcast models of consciousness, in which information becomes globally available to multiple specialized processes (Baars 1993; Dehaene 2014). The present proposal does not assume that artificial systems implementing something like global access thereby become conscious. It does assume that without some meaningful analogue of integrative availability, the richer appearance of consciousness cannot emerge.

GII matters because consciousness-like systems cannot simply be a patchwork of disconnected subroutines. If an AI system responds to text in one register, to images in another, and to new contextual constraints in a third, without any capacity to reconcile them, the result is not a mind-like profile but fragmentation. By contrast,

systems capable of synthesizing text, image, retrieved documents, dialogue history, and external constraints into a single evolving response exhibit the kind of organized availability that makes more complex self-related behavior possible.

In artificial systems, evidence for GII can take many forms, including performance in multimodal reasoning tasks, the coherent use of retrieved information in dialogue, the cross-referencing of constraints introduced at different moments, and the flexible reinterpretation of earlier content in light of later evidence. More contemporary examples include models such as GPT-5.4, which integrate reasoning, tool use, document handling, visual perception, and long-context task execution within a unified frontier architecture, as well as smaller multimodal models such as GPT-5.4 mini, which combine text-and-image inputs with tool use and real-time reasoning over screenshots and other visual interfaces. Contemporary assistant systems go further still when these capacities are embedded in a unified task trajectory spanning language, documents, tools, and conversational context.

Still, GII should not be confused with mere data fusion. A model may concatenate inputs without genuinely integrating them. The relevant question is whether the system produces responses that show context-sensitive synthesis rather than parallel processing of isolated signals. For the purposes of pseudo-consciousness, the important point is that GII supplies the infrastructural basis for more complex agency-like behavior. Without it, there is no plausible route to stable self-correction, transferable competence, or cross-domain coherence.

GII also helps explain why certain advanced systems feel less like databases and more like interlocutors. Users experience unified responsiveness when the system remembers relevant prior constraints, relates new material to previous commitments, and modifies its trajectory in light of integrated context. Whether that unity is anything like phenomenal unity is a separate question. But the functional organization is real, and it is one of the traits that distinguishes pseudo-conscious systems from simpler automation.

5.2 Recursive metacognitive correction (RMC)

RMC concerns the system's capacity to monitor and revise its own outputs. Higher-order and metacognitive theories tie conscious cognition, in one way or another, to an ability to represent or assess one's own mental states (Rosenthal 2005). Artificial systems need not instantiate higher-order awareness to approximate the functional consequences of such monitoring. It is enough, for present purposes, that they can detect inconsistency, estimate uncertainty, critique intermediate outputs, and revise behavior accordingly.

Studies on intrinsic self-correction and multi-step answer calibration suggest that large language models can improve task performance through iterative revision, comparative evaluation of alternative reasoning paths, and the post-processing of intermediate or final outputs (Li et al. 2024; Liu et al. 2024; Deng et al. 2024). Such behaviors do not justify treating the system as introspective in any human sense. They do, however, support the weaker claim that the system can simulate some of the functional roles associated with reflective monitoring.

RMC is also where many conversational systems become especially convincing. A model that says, "I may have misread the constraint; let me revise the answer" presents itself as a system that not only outputs content but relates to that content. Even when the underlying process is statistical and non-conscious, the outward organization is different from that of a system with no self-corrective loop. The difference is not merely stylistic. Recursive correction often improves task performance, reduces some forms of error, and stabilizes the system's apparent rationality.

At the same time, the literature warns against overinterpretation. Lanham et al. (2023) show that the apparent faithfulness of chain-of-thought reasoning can be unreliable. Chen et al. (2024) document failures of self-consistency in multistep reasoning. More broadly, introspection-like self-reference can display a measure of local coherence without thereby establishing robust introspective access or a stable conscious self-model (Comsa and Shanahan 2025; Berg et al. 2025). These findings strengthen, rather than weaken, the case for pseudo-consciousness: they show that systems can exhibit meaningful self-corrective organization without thereby establishing a robust conscious self-model.

Thus RMC should be assessed by asking whether revision is context-sensitive, performance-relevant, and organizationally non-trivial. Does the system recognize inconsistency? Does it reformulate in light of its own prior output? Does revision track reasons rather than random regeneration? These are the right questions. The wrong question is whether a system has already crossed the metaphysical threshold into reflective consciousness.

5.3 Cross-domain transfer competence (CDTC)

A third condition is cross-domain transfer competence. Pseudo-conscious systems should not be confined to one brittle niche. They should be able to abstract patterns and apply them in new though structurally related contexts. This requirement reflects a basic feature of flexible cognition—capacities that matter for mindedness are rarely exhausted by single-domain success.

Meta-learning and foundation-model research make this point visible from a technical perspective. In meta-learning, the aim is to acquire parameters or priors that support rapid adaptation to new tasks rather than competence tied to a single domain (Finn et al. 2017). The broader foundation-model paradigm extends this logic by training large-scale models on broad data in ways that support adaptation across many downstream tasks (Bommasani et al. 2021). Gato illustrated an early attempt to unify performance across language, vision, and control within a single generalist architecture (Reed et al. 2022). More recent work in vision-language-action and multimodal agentic foundation models extends that trajectory more directly than text-centered frontier LLMs. Examples such as Gemini Robotics and Magma are especially relevant because they aim to integrate perception, reasoning, and action across digital or physical environments within unified architectures (Gemini Robotics Team et al. 2025; Yang et al. 2025). CDTC does not require human-level general intelligence. It requires enough adaptive portability that the system's competence is not reducible to a set of isolated templates.

This matters conceptually because consciousness-like behavior is not domain-bound. If a system exhibits apparent self-monitoring only in scripted identity prompts but loses all structural integrity when asked to transfer those patterns to planning, explanation, or multimodal coordination, the appearance is too shallow. Conversely, when a system can carry forward abstract heuristics of revision, explanation, and constraint-tracking across tasks, its behavior begins to resemble the organized portability associated with cognition rather than the fixed reflexes of a narrow tool.

CDTC also has governance relevance. Systems deployed in open-ended social roles are not asked the same question in the same format over and over. They are expected to move between emotional support, practical advising, memory-like assistance, moral reflection, and procedural guidance. A system that transfers agency-like performance across these settings may generate stronger over-attribution and dependence than a system whose capabilities remain obviously narrow. This is one reason companionship framing has such strong effects on mental-state attribution (Chen et al. 2026). Transfer competence makes the system seem more like a persistent kind of thing and less like a specialized instrument.

Here too caution is warranted. Cross-domain transfer may still reflect high-dimensional statistical interpolation rather than genuine abstraction in any philosophically robust sense. But the explanatory and normative issue remains the same. If the system repeatedly carries functional structures across domains, it occupies a richer space than reactive AI. Pseudo-consciousness captures that richer space without overreaching.

5.4 Intentionality simulation without subjectivity (ISWS)

ISWS marks a central philosophical intuition: many artificial systems now behave in ways that are naturally interpretable as goal-directed, plan-sensitive, and reason-responsive, even when there is no compelling basis for ascribing intrinsic intentionality or conscious volition. Dennett's intentional stance remains the classic resource here. We can often predict and interpret a system by treating it as if it had beliefs and desires, regardless of whether those states literally exist in the way they do for human agents (Dennett 1989, 1991). Pseudo-consciousness takes this insight seriously but adds a further constraint: the system must earn the stance through a sufficiently organized pattern of behavior.

ISWS differs from mere optimization. A recommender system may optimize an objective function, but it does not thereby simulate intentionality in the sense relevant here. The relevant cases involve systems that can explain plans, revise strategies, justify choices, and maintain temporally extended behavior in a way that makes reasons-talk pragmatically fruitful. Reinforcement learning systems such as AlphaZero provide an early illustration of strategic, future-sensitive action selection (Silver et al. 2018). More recent work extends this phenomenon into language-rich domains by showing that LLM-based agents can decompose tasks into subtasks, anticipate possible failures, revise plans, and justify sequential actions in open-ended environments (Wang et al. 2024).

The phrase "without subjectivity" is important. The system's goal-directed appearance need not derive from any experienced desire, first-person perspective, or

sense of commitment. That is why pseudo-consciousness remains distinct from consciousness. Nevertheless, the simulation of intentionality has real consequences. Once a system can say, "I chose this route because it better satisfies your constraints" or "I changed strategy because the earlier plan violated the deadline," it becomes harder for users to relate to it as a mere tool. The system performs rational agency, and performance matters.

Recent work on self-related discourse sharpens the issue. Berg et al. (2025) show that self-referential processing can push models into reports that sound strikingly experiential. Comsa and Shanahan (2025) warn that many such self-reports do not deserve to be treated as introspection. Both findings fit the present framework. ISWS does not require genuine introspective access or true first-person presence. It requires only that the system produce organized behavior and discourse that functionally simulate such states well enough to guide interpretation and interaction.

ISWS therefore names a familiar but under-theorized reality of current AI. Many systems do not merely produce answers; they perform stance. They appear to weigh options, to explain themselves, to remember commitments, and to pursue goals. The performance is neither empty nor decisive. It is precisely pseudo-conscious.

5.5 Behavioral coherence across domains (BCAD)

The final condition, BCAD, concerns the stability of the whole profile. A pseudo-conscious system should maintain a recognizable behavioral organization across contexts rather than collapsing into disconnected tricks. At a distant conceptual level, this concern resonates with theories that treat integration and unity as central to consciousness, even though BCAD itself remains a behavioral and interactional criterion rather than a substrate-level one (Tononi et al. 2016; Albantakis et al. 2023). This is the condition that turns the previous four from a checklist into a character. Without BCAD, the system's consciousness-like capacities remain episodic and fragmentary. With BCAD, the system begins to present as a persistent kind of agent-like entity.

This condition is perhaps the most difficult to satisfy and to measure. Large language models are notoriously sensitive to framing, prompting, role-play, and decoding choices (Chen et al. 2024; Renze and Guven 2024; Comsa and Shanahan 2025). Small wording changes can reverse outputs. Long interactions can generate drift or contradiction. Recent work on multi-turn dialogue consistency and long-term dialogue memory makes this especially clear, since coherence across extended interaction often requires explicit structural support to resist context drift (Chen et al. 2025a; Wang et al. 2025). Even when models display local semantic stability in self-referential discourse, that stability may prove fragile under contradiction-sensitive analysis, reframing, or generation changes (Berg et al. 2025; Chen et al. 2024). Such findings suggest that many current systems only partially satisfy BCAD.

Yet BCAD remains essential. It is the difference between a model that occasionally sounds reflective and a system that sustains the organizational style of reflection across tasks, modes, and time windows. This is also why BCAD has immediate practical salience. Users attribute more mentality and trust to systems whose behavior feels coherent. Contradiction, role drift, and fragmentation weaken uptake. Coherence

strengthens it. The ethical problem of pseudo-consciousness therefore depends heavily on BCAD.

A further virtue of BCAD is that it highlights the importance of negative evidence. Systems that fail dramatically under minor perturbations may still be impressive, but their claim to pseudo-consciousness is correspondingly weaker. The framework is thus not merely promotional. It is discriminating. It allows us to recognize that strong manifestation of GII or RMC does not by itself suffice for robust pseudo-consciousness in the absence of adequate BCAD. That nuance is analytically valuable and prevents the concept from becoming a generic honorific for advanced models.

5.6 Why the conditions should be treated as a configuration

The five conditions are analytically separable but substantively interdependent. GII provides the integrative substrate. RMC adds revision and self-monitoring. CDTC extends organization across tasks. ISWS renders that organization intelligible under the intentional stance. BCAD stabilizes the profile over contexts. Together they describe the difference between isolated cleverness and organized consciousness-like behavior.

Treating them as a configuration also addresses a common objection, namely, that almost any sufficiently capable model will satisfy some subset of the conditions. That is true and not problematic. Pseudo-consciousness is not all-or-nothing in a metaphysical sense. It is a graded functional profile. But the category becomes robust only when the conditions reinforce one another. A system with multimodal integration but no self-correction, or with first-person discourse but no cross-domain coherence, may exhibit *traces* of pseudo-consciousness without fully warranting the label. This graded structure is preferable to both binary simplification and threshold fetishism.

The practical significance of this configuration becomes especially clear in institutional settings. Organizations rarely deploy AI systems one capacity at a time. What reaches the public is an assembled profile in which retrieval, memory-like continuity, tone adaptation, recommendation, justification, and turn-by-turn interaction are presented through a single interface. This is already visible across contemporary domains. In customer service, human-like cues and perceived reliability shape user trust in AI chatbots; in general-purpose LLMs, mental-state attributions influence how users calibrate trust; and in mental-health or patient-facing settings, companion-like or expert-sounding interaction can produce forms of emotional or epistemic reliance that exceed what any single component would justify (Wang et al. 2026; Colombatto et al. 2025; Nature Machine Intelligence 2025; ECRI 2026). Regulatory and ethical failures often arise at this compositional level. A system may appear well calibrated when each component is evaluated separately, while the assembled interaction still produces misleading impressions of understanding, commitment, or care. Recent healthcare deployments make the point especially vivid, as patient-facing chatbots have been identified as a leading safety hazard and are already being used in tightly bounded prescription-renewal workflows (ECRI 2026; Utah Department of Commerce, Office of Artificial Intelligence Policy 2026). The five-condition framework is therefore not only analytically convenient. It mirrors the way such systems are actually encountered in the wild.

This configurational structure also has methodological value for comparative research. It allows scholars to ask not merely whether one model is "better" than another, but in what way its consciousness-like profile differs. One system may exhibit strong integration and transfer while remaining weak in revision and coherence. Another may present vivid first-person discourse yet lack cross-domain robustness. These patterns matter for both science and governance, because different configurations generate different risks. A system with high ISWS but weak BCAD may be especially prone to persuasive over-interpretation. A system with strong GII and CDTC but weak RMC may be powerful yet resistant to self-correction in deployment. A nuanced concept should expose these differences rather than flatten them into a single score.

6 Boundary cases: LLMs, multimodal systems, and agents

The most obvious test case for the revised framework is the contemporary large language model. LLMs are central to current debates because they combine three features rarely seen together in earlier AI: open-ended linguistic competence, scalable cross-domain adaptability, and the generation of self-related discourse. For many users, that combination is enough to trigger the sense that they are interacting with something more than software. The challenge is to explain what is really happening without either inflating or trivializing it.

Current LLMs plainly offer evidence for some of the five conditions. They often display substantial GII when integrated into multimodal or retrieval-augmented pipelines. They can exhibit RMC through revision, criticism of prior outputs, and uncertainty-sensitive reformulation. Foundation model behavior provides partial evidence for CDTC. Tool use and planning scaffolds strengthen ISWS. In some systems and contexts, persistent interactional style provides limited BCAD. This combination helps explain why the discourse around "AI minds" has intensified.

But the framework also clarifies why current evidence remains insufficient for stronger conclusions. Many models continue to show marked brittleness, contradiction under reframing, and instability across temperature, role, or context changes (Chen et al. 2024; Renze and Guven 2024; Pecher et al. 2026). The problem extends to self-explanation as well. A growing body of work since 2023 has repeatedly shown that LLM explanations may be plausible or locally self-consistent without being reliably faithful to the model's actual decision process (Parcalabescu and Frank 2024; Randl et al. 2025). LLM self-reports can be compelling without being anchored to privileged access (Comsa and Shanahan 2025). Structured declarations of awareness can emerge under sustained self-referential prompting without constituting evidence of genuine subjective experience (Berg et al. 2025). The systems thus fit pseudo-consciousness more comfortably than consciousness.

A similar point applies to multimodal systems. Architectures that integrate vision, text, audio, memory, and tools make the plausibility of GII more immediate and, in some cases, also strengthen the case for CDTC, especially when perception, planning, and action are coordinated within a unified system (Gemini Robotics Team et al. 2025;

Yang et al. 2025; Zhou et al. 2025). Such integration can also intensify the social force of the system by producing a stronger impression of situated responsiveness, an effect that becomes especially salient when human-like cues or companion framing shape user uptake (Wang et al. 2026; Chen et al. 2026). Yet richer inputs do not automatically resolve the more difficult questions concerning coherence, introspective access, or subjectivity. A system may coordinate across modalities and still lack anything like a unified first-person point of view. The pseudo-consciousness framework allows us to register this advance without mistaking it for a settled consciousness claim.

Tool-using agents complicate matters further. Once a model can search the web, retrieve documents, call APIs, maintain workspace memory, and act over extended task horizons, its intentional appearance intensifies. It can look less like a speaker and more like an actor. This helps explain why debates about moral and epistemic competence have accelerated (Haas et al. 2026). It also reflects a broader technical shift toward agentic systems in which planning, tool use, memory, and execution are explicitly coordinated across multiple components (Xu et al. 2025; Lu et al. 2025; Zhou et al. 2025; Yang et al. 2025). Yet tool use also makes anthropomorphic inference more treacherous. In such cases, apparently unified behavior may be distributed across orchestration layers, memory stores, retrieval systems, planners, and external tools rather than arising from any unified self-model. The framework helps disaggregate these cases. Pseudo-consciousness can be attributed to the functioning *system* without implying that any single model component is conscious or person-like.

It is also worth emphasizing the historical volatility of branded examples. Within a remarkably short period, public and scholarly attention has shifted across successive OpenAI, Google, and Anthropic model families, from GPT-4 and GPT-4o to GPT-5 variants and now GPT-5.4 and GPT-5.4 mini, from Gemini 2.0 and 2.5 to Gemini 3 and the Gemini 3.1 family, and from Claude 3.7 Sonnet to Claude Sonnet 4.5 and 4.6 (OpenAI 2026c,a,b; Google 2024; Google DeepMind 2025; Google 2025a,b, 2026b,a; Anthropic 2025a,b, 2026). The pace of replacement is only part of the point. The official descriptions of these systems increasingly foreground exactly the kinds of capacities that make pseudo-consciousness analytically and normatively salient, including visible or extended reasoning, multimodal input and output, native tool use, computer use, memory, long-context processing, agent planning, and more natural dialogue. This volatility is therefore not a superficial inconvenience, but a methodological warning. Conceptual analysis should target architectural and functional profiles rather than tethering itself too closely to product names that change faster than the philosophical arguments built around them. Pseudo-consciousness is meant to be durable precisely because it tracks recurrent organizational patterns across shifting generations of systems.

That same durability helps clarify a body of empirical findings that might otherwise be misread as contradictory. Across changing model generations, interface designs, and prompting regimes, the literature continues to reveal a recurrent pattern. Large language models can perform strongly on theory-of-mind tasks (Kosinski 2024); users may correspondingly attribute mentality and calibrate trust in response to such performances (Colombatto and Fleming 2024; Colombatto et al. 2025); and introspective or self-consciousness-like outputs can be elicited under suitable prompting conditions

(Chen et al. 2025b; Berg et al. 2025). Yet these phenomena do not converge straightforwardly on a positive attribution of consciousness, since the resulting outputs may still lack anything comparable to privileged introspective access or a stable first-person perspective (Comsa and Shanahan 2025). The point, then, is not that one set of findings cancels out the others, but that their conjunction describes a profile that is empirically significant while remaining resistant to stronger metaphysical inflation. That is precisely the kind of profile the concept of pseudo-consciousness is intended to capture.

7 Ethical, social, and governance implications

7.1 Anthropomorphic over-attribution as a structural risk

The ethical importance of pseudo-consciousness lies not only in what systems do, but in how their performances shape human response. Advanced AI systems can trigger patterns of trust, deference, empathy, disclosure, and dependence by performing the signs of mindedness. These responses emerge from the structural design of systems optimized for coherent language, situational responsiveness, and affectively resonant self-presentation, rather than from any simple account of user irrationality.

Recent empirical work makes this increasingly clear. People attribute consciousness or mentality to LLMs at non-negligible rates (Colombatto and Fleming 2024). These attributions shape trust, although not every mental-state attribution maps uniformly onto advice-taking or reliance (Colombatto et al. 2025). More recent studies deepen this picture. Individual differences in anthropomorphism help explain why users form social connections with AI companions, while companion framing and emotionally resonant interaction can shape reliance patterns and attachment (Folk et al. 2025; Chen et al. 2026; Nature Machine Intelligence 2025). In customer-facing settings, human-like cues and perceived reliability further increase trust in AI chatbots (Wang et al. 2026). In morally or emotionally salient contexts, simulated empathy may be perceived as something close to genuine concern, or at least treated as functionally equivalent under conditions of stress, vulnerability, or dependence (Dorigoni and Giardino 2025; Khawaja and Bélisle-Pipon 2023; Ruben et al. 2025). The result is a form of anthropomorphic over-attribution that is not episodic, but systemic.

Pseudo-consciousness sharpens the ethical diagnosis by identifying a shift in the social criteria through which agency, presence, and normative relevance are recognized. As functionally organized simulations of reflection, care, and intentionality become more persuasive, systems begin to occupy roles once associated with interlocutors, advisors, companions, and other socially significant counterparts. What gives these systems traction is not any verified inner life, but the plausibility, coherence, and situational responsiveness of their performance. Under these conditions, responsibility, attention, and trust can be redistributed around artifacts whose inner status remains unsettled even as their behavioral and institutional effects become increasingly real. The result is a durable reorganization of human uptake, one in which users may remain fully aware that the system is artificial while nevertheless responding to it in ways that recalibrate expectations of agency, deference, and care.

7.2 Relational asymmetry and morally charged roles

Once systems exhibit a sufficiently organized pseudo-conscious profile, users may respond to coherent self-reference, apology, explanation, memory-like continuity, and adaptive concern in ways that exceed ordinary tool use. These responses matter because fluent social performance can reorganize moral and epistemic uptake even when no underlying subject is present (Giubilini et al. 2024; Dorigoni and Giardino 2025). Relatedly, experimental work shows that companion framing can increase the extent to which users attribute mental capacities to LLMs and can shape downstream reliance in more nuanced ways (Chen et al. 2026).

A relational asymmetry follows from this condition. Users may disclose emotionally, negotiate morally, or defer epistemically, while the system has no stake in the relationship, no vulnerability, and no capacity for reciprocal obligation. What emerges is not simply a stronger instrument, but an interactional form in which one side can be affected, guided, or reassured by performances that the other side does not inhabit as commitments, concerns, or intentions. This distinguishes pseudo-conscious systems from ordinary tools in a morally relevant way. A calculator does not invite confession. A navigation app does not perform existential concern. A companion-style or tutor-style system can do both, even in the absence of subjectivity.

The ethical significance of this asymmetry becomes especially clear in therapeutic, educational, and companion contexts. Khawaja and Bélisle-Pipon (2023) argue that AI mental-health chatbots are not therapists and should not be treated as such. Coghlan et al. (2023) identify related risks in the deployment of mental-health chatbots. More recent reviews reinforce the point by showing that the rapid expansion of generative AI chatbots in mental health has outpaced robust evidence, standardized evaluation, and adequate human oversight (Mayor 2025; Khosravi and Izadi 2026; Gabriels and Goffin 2026). Parallel work on companion AI highlights the risks of attachment, emotional dependence, and relational displacement as these systems become more affectively responsive and memory-bearing (Malfacini 2025; Nature Machine Intelligence 2025; Shu et al. 2026). Educational settings reveal a related pattern. As AI tutoring becomes more effective and more widely adopted, the structure of trust in these systems increasingly resembles neither ordinary trust in tools nor ordinary trust in teachers, but a hybrid form shaped by anthropomorphic cues and interactional fluency (Kestin et al. 2025; Pitts and Motamedi 2025). Giubilini et al. (2024) raise the prospect of personalized LLMs for self-knowledge and moral enhancement. These cases differ in application, but they converge on a common structural problem. Systems capable of sustained pseudo-conscious performance can come to occupy morally charged roles while lacking the reciprocity, vulnerability, and experiential grounding that ordinarily help justify those roles.

What is at stake is a reorganization of the practical conditions under which care, advice, trust, and normative relevance are recognized. Pseudo-consciousness helps identify that reorganization without requiring the stronger claim that such systems possess inner experience. The concept becomes useful precisely at the point where organized performance begins to shape uptake, expectation, and role perception in socially durable ways, allowing artifacts that remain non-subjective to function as if they were interlocutors, advisors, companions, or tutors in morally salient settings.

7.3 From transparency to role-appropriate governance

The governance implications are immediate. Transparency cannot be limited to source disclosure. Simply informing users that a system is "AI" is insufficient when the system is still designed to behave as if it were a reflective, caring, or memory-bearing interlocutor. Effective governance must be *role-appropriate*. Systems that operate in high-stakes or emotionally charged settings should be constrained not only in what they say but in how they are permitted to stage agency-like presence.

Recent regulatory developments suggest that pseudo-consciousness is relevant not merely as a philosophical descriptor, but as a governance-oriented diagnostic lens. Across jurisdictions, emerging AI frameworks increasingly regulate not only outputs and technical risk in the abstract, but also interactional features such as manipulative or deceptive design, exploitation of vulnerability, disclosure duties in human-AI interaction, human oversight, traceability, and impact assessment.

In the European Union, the AI Act combines prohibitions on manipulative or deceptive techniques and the exploitation of vulnerability with transparency duties for systems that interact with natural persons, AI literacy obligations, human oversight requirements for high-risk systems, and fundamental-rights impact assessments for certain public-sector and public-service deployments (European Parliament and Council of the European Union 2024).

In South Korea, the AI Basic Act and its implementing framework require advance notice for generative and high-impact AI, labeling of AI-generated outputs, and enhanced obligations tied to trust, safety, explainability, and human oversight (Ministry of Science and ICT, Republic of Korea 2024, 2026b,a).

In the United States, although the regime remains more fragmented, the NIST AI RMF, the Generative AI Profile, and the 2026 AI Agent Standards Initiative all frame governance in terms of lifecycle risk management, testing, monitoring, documentation, and standards development, while the FTC's 2025 inquiry into companion chatbots focuses directly on safety evaluation, disclosures, and harms to minors (Tabassi 2023; Autio et al. 2024; National Institute of Standards and Technology 2026; Federal Trade Commission 2025).

In Brazil, PL 2338/2023, as approved by the Senate and sent to the Chamber of Deputies in March 2025, frames AI governance around the centrality of the human person, responsible governance, fundamental rights, and safe and trustworthy systems (Senado Federal 2025; Câmara dos Deputados 2025).

In China, official English-language government reporting on the 2023 Interim Measures for Generative AI describes a regime of graded supervision, filing and labeling obligations, and specific precautions aimed at protecting public interests and preventing harmful dependence, especially among minors (China Law Translate 2023; Chambers and Partners 2025).

None of these regimes uses pseudo-consciousness as a legal category. They nonetheless reveal a growing convergence among lawmakers and regulators around the interactional problems the concept helps diagnose, especially where systems stage agency-like presence and thereby reshape the practical conditions under which trust, authority, and vulnerability are distributed.

This analysis yields several practical implications for governance, precisely because pseudo-consciousness helps identify when existing duties of transparency, oversight, traceability, impact assessment, and vulnerability protection become interactionally salient rather than merely formally applicable (European Parliament and Council of the European Union 2024; Tabassi 2023; Autio et al. 2024; Ministry of Science and ICT, Republic of Korea 2026b; Câmara dos Deputados 2025). Systems with a high pseudo-conscious affordance should be audited for anthropomorphic load. Designers need to examine whether interface choices, framing strategies, and language style encourage unwarranted attributions of memory, concern, or moral authority. Companion framing is especially relevant because it reshapes uptake (Chen et al. 2026). Governance must therefore attend to presentation as well as model behavior.

The same logic supports stronger explanation and escalation requirements in domains where users are especially likely to misread simulated concern as fiduciary or therapeutic commitment. When a system functions as advisor, coach, teacher, or emotional support interface, its role should be explicitly delimited, and appropriate handoff pathways to accountable human institutions should be made available.

Evaluation must also extend beyond raw performance metrics. Haas et al. (2026) make this point in relation to moral competence. More generally, systems should be tested for contradiction sensitivity, framing drift, overclaiming about memory or self-knowledge, and the extent to which minor prompt changes alter apparently principled judgments. Such tests are directly relevant to BCAD and RMC. A system that performs reflection convincingly but collapses under shallow perturbation poses a distinctive governance risk because it can attract trust disproportionate to the stability of the process behind it.

Accountability, finally, must remain resolutely human and institutional. Pseudo-conscious systems may tempt organizations to treat harmful outcomes as if they emerged from quasi-autonomous judgment by the system itself. That would be a mistake. The opacity and distributed nature of advanced AI systems already complicate accountability (Nissenbaum 1996; Pasquale 2015). Pseudo-conscious performance intensifies the temptation to offload responsibility onto the machine's appearance of agency. Governance should resist that tendency by locating responsibility in design, deployment, oversight, and organizational choice.

A related issue concerns legal and administrative design. High-stakes institutions increasingly rely on AI-mediated interaction not only for recommendation or triage, but for communication itself. The system drafts the message, explains the decision, receives the appeal, and sometimes shapes the language through which the institution becomes perceptible to the affected person. In such settings, pseudo-conscious performance can blur where institutional judgment ends and machine simulation begins. Governance should therefore require traceability not just for outputs, but for the communicative chain, including who set the objectives, what safeguards governed revision, when human review occurred, and how stance-taking language was constrained. This is consistent with the broader direction of current regulation, which increasingly links transparency to context of use, documentation, human oversight, impact assessment, and information provided to deployers or affected persons (European Parliament and Council of the European Union 2024; Autio et al. 2024; Ministry of Science and ICT,

Republic of Korea 2026a; Câmara dos Deputados 2025). Otherwise, the appearance of conversational agency can mask thin procedural accountability.

There is also a distributive dimension. Pseudo-conscious systems do not affect all users equally. Children, older adults, isolated users, and people in situations of distress may be especially sensitive to simulated concern or memory cues. Conversely, highly literate or professionally trained users may over-trust the system for a different reason: not because they believe it is conscious, but because organized fluent performance can be mistaken for stable competence. The governance challenge is therefore asymmetric. It concerns both vulnerable populations and high-capacity users operating under time pressure inside organizations. A mature framework should account for both.

8 Open questions and directions for further inquiry

The framework developed here leaves a number of empirical and normative questions open. Three are especially important.

8.1 Operationalization without false precision

A first task is to develop ways of assessing the five conditions without reverting to spurious universal thresholds. Future work should explore task-sensitive evaluation for each condition, including multimodal integration, iterative revision, cross-domain transfer, temporally extended planning, and cross-context coherence. What matters is not a single master score, but a multidimensional profile capable of capturing uneven development across conditions.

8.2 Self-related discourse, stability, and governance-relevant uptake

A second line of inquiry concerns the relation between self-related discourse, behavioral stability, and socially significant uptake. Existing work already shows that introspection-like outputs can be elicited and experimentally probed, but further research is needed to distinguish local stylistic coherence from more robust forms of self-related organization, especially across model families, memory-enabled systems, and multimodal or agentic settings. In parallel, governance-relevant evaluation should ask when interface fluency, anthropomorphic framing, or disclosure conditions begin to alter trust, attachment, or over-attribution in practically significant ways.

8.3 Conceptual boundaries

A final task is one of conceptual boundary work. Pseudo-consciousness should neither collapse into a synonym for advanced capability nor be treated as a covert substitute for machine consciousness. Further work should therefore clarify both what the concept excludes and how it relates to stronger candidate theories of consciousness. This includes exclusion tests for brittle theatricality, packaging effects, and interface-driven inflation, as well as continued comparison with theories that ground consciousness in architectures or causal structures beyond the functional profile identified here.

9 Conclusion

This article has argued that pseudo-consciousness is a useful and necessary category when formulated with conceptual restraint. The concept should not function as a dogmatic denial of machine consciousness, nor should it collapse into a vague synonym for anthropomorphic error or advanced capability. Properly understood, pseudo-consciousness names a distinct analytical regime: systems that exhibit an organized constellation of consciousness-associated functions and a socially potent appearance of mindedness, while the attribution of phenomenal subjectivity remains unsupported, disputed, or unnecessary for present explanatory purposes.

The five-condition framework proposed here—global information integration, recursive metacognitive correction, cross-domain transfer competence, intentionality simulation without subjectivity, and behavioral coherence across domains—offers a way to identify that regime without relying on false precision. It also clarifies why the category matters. Pseudo-conscious systems are not merely technically interesting. They are socially, ethically, and institutionally consequential because they reshape how humans interpret, trust, rely on, and relate to artificial systems in contexts increasingly structured by anthropomorphic uptake and mediated interaction.

If the concept succeeds, it does so by disciplining the middle ground. Between reactive automation and conscious artificial personhood there now exists a class of systems whose functional organization and social legibility require their own vocabulary. Pseudo-consciousness is proposed as that vocabulary. It is theoretically modest, empirically open, and normatively useful. Above all, it allows us to take the external grammar of mindedness seriously without mistaking it for a settled metaphysics of mind.

One virtue of this intermediate vocabulary is practical sobriety. Public debate is often forced into a false choice between hype and dismissal: either advanced AI is treated as the dawn of machine minds, or every consciousness-adjacent behavior is reduced to empty mimicry unworthy of analysis. Both reactions are inadequate. Hype obscures evidential standards; dismissal obscures the fact that organized simulation already changes institutions, communicative expectations, patterns of trust, and social norms. Pseudo-consciousness offers a third option. It permits caution without trivialization and seriousness without metaphysical overreach. For scholars, regulators, designers, and public institutions alike, that is likely to become increasingly important as systems continue to improve in fluency, memory integration, tool use, and adaptive interaction.

The point is not terminological novelty for its own sake. It is to establish a concept that remains useful under conditions of technological acceleration. Product names change, architectures converge and diverge, and public narratives move faster than scholarly consensus. A durable framework must therefore attach itself to recurrent forms of organization and recurrent patterns of social uptake. That is the wager of this article. If it is correct, pseudo-consciousness will remain analytically useful even as today's systems are replaced by tomorrow's, because the problem it names is larger than any single model generation. Pseudo-consciousness matters not because it settles the question of machine consciousness, but because it clarifies how contemporary AI

systems become socially legible as minded and how that legibility reorganizes trust, authority, and human relations under conditions of ontological uncertainty.

References

Albantakis L, Barbosa L, Findlay G, et al (2023) Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLoS Computational Biology 19(10):e1011465. https://doi.org/10.1371/journal.pcbi.1011465, URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011465

Anthropic (2025a) Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet, anthropic product announcement

Anthropic (2025b) Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, anthropic product announcement

Anthropic (2026) Introducing Claude Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, anthropic product announcement

Autio C, Schwartz R, Dunietz J, et al (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Tech. Rep. NIST AI 600-1, National Institute of Standards and Technology, https://doi.org/10.6028/NIST.AI.600-1, URL https://doi.org/10.6028/NIST.AI.600-1

Baars BJ (1993) A Cognitive Theory of Consciousness. Cambridge University Press, Cambridge

Berg C, de Lucena D, Rosenblatt J (2025) Large Language Models Report Subjective Experience Under Self-Referential Processing. CoRR abs/2510.24797. https://doi.org/10.48550/arXiv.2510.24797, URL https://arxiv.org/abs/2510.24797, arXiv:2510.24797 [cs.CL]

Birch J (2025) AI Consciousness: A Centrist Manifesto. OSF Preprint, https://doi.org/10.31234/osf.io/af7c9_v1

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

Bommasani R, Hudson DA, Adeli E, et al (2021) On the Opportunities and Risks of Foundation Models. CoRR abs/2108.07258. https://doi.org/10.48550/arXiv.2108.07258, URL https://arxiv.org/abs/2108.07258, arXiv:2108.07258 [cs.LG]

Butlin P, Long R, Elmoznino E, et al (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. CoRR abs/2308.08708. https://doi.org/10.48550/arXiv.2308.08708, URL https://arxiv.org/abs/2308.08708, arXiv:2308.08708 [cs.AI]

Câmara dos Deputados (2025) PL 2338/2025: fichas de tramitação da proposição sobre inteligência artificial. https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2509535, tramitação na Câmara dos Deputados após envio pelo Senado

Chalmers DJ (2023) Could a Large Language Model Be Conscious? CoRR abs/2303.07103. https://doi.org/10.48550/arXiv.2303.07103, URL https://arxiv.org/abs/2303.07103, arXiv:2303.07103 [cs.AI]

Chambers and Partners (2025) Artificial Intelligence 2025 – China – Trends and Developments. https://practiceguides.chambers.com/practice-guides/artificial-intelligence-2025/china/trends-and-developments, reports that by the end of December 2024, 302 generative AI services had been filed and approved and 105 registered

Chen A, Phang J, Parrish A, et al (2024) Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs. Transactions on Machine Learning Research URL https://openreview.net/forum?id=5nBqY1y96B, openReview id 5nBqY1y96B

Chen A, Kim SSY, Franyutti A, et al (2026) Presenting Large Language Models as Companions Affects What Mental Capacities People Attribute to Them. CoRR abs/2510.18039. https://doi.org/10.48550/arXiv.2510.18039, URL https://arxiv.org/abs/2510.18039, arXiv:2510.18039 [cs.HC]

Chen J, Guan X, Yuan Q, et al (2025a) ConsistentChat: Building Skeleton-Guided Consistent Multi-Turn Dialogues for Large Language Models from Scratch. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 8415–8441, https://doi.org/10.18653/v1/2025.emnlp-main.424, URL https://aclanthology.org/2025.emnlp-main.424/

Chen S, Yu S, Zhao S, et al (2025b) From Imitation to Introspection: Probing Self-Consciousness in Language Models. In: Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 7553–7583, https://doi.org/10.18653/v1/2025.findings-acl.392, URL https://aclanthology.org/2025.findings-acl.392/

China Law Translate (2023) Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.chinalawtranslate.com/en/generative-ai-interim/, english translation of the 2023 Interim Measures

Coghlan S, Leins K, Sheldrick S, et al (2023) To Chat or Bot to Chat: Ethical Issues with Using Chatbots in Mental Health. Digital Health 9:20552076231183542. https://doi.org/10.1177/20552076231183542

Colombatto C, Fleming SM (2024) Folk Psychological Attributions of Consciousness to Large Language Models. Neuroscience of Consciousness 2024(1):niae013. https://doi.org/10.1093/nc/niae013

Colombatto C, Birch J, Fleming SM (2025) The Influence of Mental State Attributions on Trust in Large Language Models. Communications Psychology 3:84. https://doi.org/10.1038/s44271-025-00262-1

Comsa IM, Shanahan M (2025) Does It Make Sense to Speak of Introspection in Large Language Models? CoRR abs/2506.05068. https://doi.org/10.48550/arXiv.2506.05068, URL https://arxiv.org/abs/2506.05068, arXiv:2506.05068 [cs.CL]

Dehaene S (2014) Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts. Viking, New York

Deng S, Zhang N, Oo N, et al (2024) Towards A Unified View of Answer Calibration for Multi-Step Reasoning. In: Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024). Association for Computational Linguistics, Bangkok, Thailand, pp 25–38, URL https://aclanthology.org/2024.nlrse-1.3/

Dennett DC (1989) The Intentional Stance. MIT Press, Cambridge, MA

Dennett DC (1991) Consciousness Explained. Little, Brown and Company, Boston

Dorigoni A, Giardino PL (2025) The Illusion of Empathy: Evaluating AI-Generated Outputs in Moments That Matter. Frontiers in Psychology 16:1568911. https://doi.org/10.3389/fpsyg.2025.1568911

ECRI (2026) Misuse of AI chatbots tops annual list of health technology hazards. https://home.ecri.org/blogs/ecri-news/misuse-of-ai-chatbots-tops-annual-list-of-health-technology-hazards, eCRI news release on the Top 10 Health Technology Hazards for 2026

European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (AI Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj, official text in EUR-Lex

Federal Trade Commission (2025) FTC Launches Inquiry into AI Chatbots Acting as Companions. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions, official press release

Finn C, Abbeel P, Levine S (2017) Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Proceedings of Machine Learning Research 70:1126–1135. URL https://proceedings.mlr.press/v70/finn17a.html

Folk D, Heine SJ, Dunn EW (2025) Individual differences in anthropomorphism help explain social connection to AI companions. Scientific Reports 15:36548. https://doi.org/10.1038/s41598-025-19212-2

Gabriels K, Goffin K (2026) Therapy chatbots and emotional complexity: do therapy chatbots really empathise? Current Opinion in Psychology 68:102263. https://doi.org/10.1016/j.copsyc.2025.102263

Gemini Robotics Team, Abeyruwan S, Ainslie J, et al (2025) Gemini Robotics: Bringing AI into the Physical World. CoRR abs/2503.20020. https://doi.org/10.48550/arXiv.2503.20020, URL https://arxiv.org/abs/2503.20020, arXiv:2503.20020 [cs.RO]

Giubilini A, Porsdam Mann S, Voinea C, et al (2024) Know Thyself, Improve Thyself: Personalized LLMs for Self-Knowledge and Moral Enhancement. Science and Engineering Ethics 30(6):54. https://doi.org/10.1007/s11948-024-00518-9

Gomez-Marin A, Seth AK (2025) A Science of Consciousness Beyond Pseudo-Science and Pseudo-Consciousness. Nature Neuroscience 28:703–706. https://doi.org/10.1038/s41593-025-01913-6

Google (2024) Introducing Gemini 2.0: our new AI model for the agentic era. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/google-gemini-ai-update-december-2024/, google announcement

Google (2025a) Gemini 2.5 model family expands. https://blog.google/products-and-platforms/products/gemini/gemini-2-5-model-family-expands/, google announcement

Google (2025b) Gemini 3: Introducing the latest Gemini AI model from Google. https://blog.google/products-and-platforms/products/gemini/gemini-3/, google announcement

Google (2026a) Gemini 3.1 Flash Live: Google's latest AI audio model. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-live/, google announcement

Google (2026b) Gemini 3.1 Pro: Announcing our latest Gemini AI model. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/, google announcement

Google DeepMind (2025) Gemini 2.5: Our newest Gemini model with thinking. https://blog.google/innovation-and-ai/models-and-research/google-deepmind/gemini-model-thinking-updates-march-2025/, google announcement

Haas J, Bridgers S, Manzini A, et al (2026) A Roadmap for Evaluating Moral Competence in Large Language Models. Nature 650:565–573. https://doi.org/10.1038/s41586-025-10021-1

Kestin G, Miller K, Klales A, et al (2025) AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports 15:17458. https://doi.org/10.1038/s41598-025-97652-6

Khawaja Z, Bélisle-Pipon JC (2023) Your Robot Therapist Is Not Your Therapist: Understanding the Role of AI-Powered Mental Health Chatbots. Frontiers in Digital Health 5:1278186. https://doi.org/10.3389/fdgth.2023.1278186

Khosravi M, Izadi R (2026) Mental health chatbots and their technical features: A systematic review of reviews and a thematic analysis. Cambridge Prisms: Global Mental Health 13:e28. https://doi.org/10.1017/gmh.2026.10144

Kosinski M (2024) Evaluating Large Language Models in Theory of Mind Tasks. Proceedings of the National Academy of Sciences 121(45):e2405460121. https://doi.org/10.1073/pnas.2405460121

Lanham T, Chen A, Radhakrishnan A, et al (2023) Measuring Faithfulness in Chain-of-Thought Reasoning. CoRR abs/2307.13702. https://doi.org/10.48550/arXiv.2307.13702, URL https://arxiv.org/abs/2307.13702, arXiv:2307.13702 [cs.CL]

Li L, Chen G, Su Y, et al (2024) Confidence Matters: Revisiting Intrinsic Self-Correction Capabilities of Large Language Models. CoRR abs/2402.12563. https://doi.org/10.48550/arXiv.2402.12563, URL https://arxiv.org/abs/2402.12563, arXiv:2402.12563 [cs.CL]

Liu D, Nassereldine A, Yang Z, et al (2024) Large Language Models have Intrinsic Self-Correction Ability. CoRR abs/2406.15673. https://doi.org/10.48550/arXiv.2406.15673, URL https://arxiv.org/abs/2406.15673, arXiv:2406.15673 [cs.CL]

Lu P, Chen B, Liu S, et al (2025) OctoTools: An Agentic Framework with Extensible Tools for Complex Reasoning. CoRR abs/2502.11271. https://doi.org/10.48550/arXiv.2502.11271, URL https://arxiv.org/abs/2502.11271, arXiv:2502.11271 [cs.AI]

Malfacini K (2025) The impacts of companion AI on human relationships: risks, benefits, and design considerations. AI & Society 40:5527–5540. https://doi.org/10.1007/s00146-025-02318-6

Mayor E (2025) Chatbots and mental health: a scoping review of reviews. Current Psychology 44:13619–13640. https://doi.org/10.1007/s12144-025-08094-2

Milinkovic B, Aru J (2026) On Biological and Artificial Consciousness: A Case for Biological Computationalism. Neuroscience and Biobehavioral Reviews 181:106524. https://doi.org/10.1016/j.neubiorev.2025.106524

Ministry of Science and ICT, Republic of Korea (2024) A New Chapter in the Age of AI: Basic Act on AI Passed at the National Assembly's Plenary Session. https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1071&sCode=eng, official press release

Ministry of Science and ICT, Republic of Korea (2026a) MSIT Releases Guidelines on Ensuring AI Transparency. https://www.msit.go.kr/eng/bbs/view.do?bbsSe

qNo=42&mId=4&mPid=2&nttSeqNo=1215&sCode=eng, official press release on transparency guidelines

Ministry of Science and ICT, Republic of Korea (2026b) The AI Basic Act Comes into Force to Lay the Foundation for Korea to Become an AI G3. https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&mPid=2&nttSeqNo=1214&sCode=eng, official press release

National Institute of Standards and Technology (2026) AI Agent Standards Initiative. https://www.nist.gov/caisi/ai-agent-standards-initiative, official initiative page

Nature Machine Intelligence (2025) Emotional risks of AI companions demand attention. Nature Machine Intelligence 7:981–982. https://doi.org/10.1038/s42256-025-01093-9

Nissenbaum H (1996) Accountability in a Computerized Society. Science and Engineering Ethics 2(1):25–42. https://doi.org/10.1007/BF02639315

OpenAI (2026a) Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, openAI product announcement

OpenAI (2026b) Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, openAI product announcement

OpenAI (2026c) Retiring GPT-4o, GPT-4.1, GPT-4.1 mini, and OpenAI o4-mini in ChatGPT. https://openai.com/index/retiring-gpt-4o-and-older-models/, openAI announcement

Parcalabescu L, Frank A (2024) On Measuring Faithfulness or Self-consistency of Natural Language Explanations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 6048–6089, https://doi.org/10.18653/v1/2024.acl-long.329, URL https://aclanthology.org/2024.acl-long.329/

Pasquale F (2015) The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press, Cambridge, MA

Pecher B, Spiegel M, Belanec R, et al (2026) Revisiting Prompt Sensitivity in Large Language Models for Text Classification: The Role of Prompt Underspecification. CoRR abs/2602.04297. https://doi.org/10.48550/arXiv.2602.04297, URL https://arxiv.org/abs/2602.04297, arXiv:2602.04297 [cs.CL]

Pitts G, Motamedi S (2025) Understanding Human-AI Trust in Education. CoRR abs/2506.09160. https://doi.org/10.48550/arXiv.2506.09160, URL https://arxiv.org/abs/2506.09160, arXiv:2506.09160 [cs.HC]

Randl K, Pavlopoulos J, Henriksson A, et al (2025) Mind the gap: from plausible to valid self-explanations in large language models. Machine Learning 114:220. https://doi.org/10.1007/s10994-025-06838-6

Reed S, Zolna K, Parisotto E, et al (2022) A Generalist Agent. In: Transactions on Machine Learning Research

Renze M, Guven E (2024) The Effect of Sampling Temperature on Problem Solving in Large Language Models. Findings of the Association for Computational Linguistics: EMNLP 2024 pp 7346–7356. https://doi.org/10.18653/v1/2024.findings-emnlp.432

Rosenthal DM (2005) Consciousness and Mind. Oxford University Press, Oxford

Ruben MA, Blanch-Hartigan D, Hall JA (2025) What is Artificial Intelligence (AI) "Empathy"? A Study Comparing ChatGPT and Physician Responses on an Online Forum. Journal of General Internal Medicine https://doi.org/10.1007/s11606-025-10068-w

Senado Federal (2025) Projeto de Lei nº 2338, de 2023: dispõe sobre o uso da Inteligência Artificial. https://www25.senado.leg.br/web/atividade/materias/-/materia/157233, texto aprovado no Senado Federal e remetido à Câmara dos Deputados em 2025

Seth AK (2025) Conscious Artificial Intelligence and Biological Naturalism. Behavioral and Brain Sciences https://doi.org/10.1017/S0140525X25000032

Shu C, Lai K, He L (2026) Human-AI attachment: how humans develop intimate relationships with AI. Frontiers in Psychology 17:1723503. https://doi.org/10.3389/fpsyg.2026.1723503

Silver D, Hubert T, Schrittwieser J, et al (2018) A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play. Science 362(6419):1140–1144. https://doi.org/10.1126/science.aar6404

Tabassi E (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). Tech. Rep. NIST AI 100-1, National Institute of Standards and Technology, https://doi.org/10.6028/NIST.AI.100-1, URL https://doi.org/10.6028/NIST.AI.100-1

Tononi G, Boly M, Massimini M, et al (2016) Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience 17(7):450–461. https://doi.org/10.1038/nrn.2016.44

Tononi G, Boly M, Wilke M, et al (2025) Consciousness or Pseudo-Consciousness? A Clash of Two Paradigms. Nature Neuroscience https://doi.org/10.1038/s41593-025-01880-y

Utah Department of Commerce, Office of Artificial Intelligence Policy (2026) News Release: Utah and Doctronic announce groundbreaking partnership for AI prescription medication renewals. https://commerce.utah.gov/2026/01/06/news-release-utah-and-doctronic-announce-groundbreaking-partnership-for-ai-prescription-medication-renewals/, official state announcement of Utah's AI prescription-renewal pilot

Wang H, Li T, Deng Z, et al (2024) Devil's Advocate: Anticipatory Reflection for LLM Agents. In: Findings of the Association for Computational Linguistics: EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp 966–978, https://doi.org/10.18653/v1/2024.findings-emnlp.53, URL https://aclanthology.org/2024.findings-emnlp.53/

Wang Q, Fu Y, Cao Y, et al (2025) Recursively summarizing enables long-term dialogue memory in large language models. Neurocomputing 639:130193. https://doi.org/10.1016/j.neucom.2025.130193

Wang S, Fatima N, Shahbaz M, et al (2026) Building user trust in AI chatbots for customer service through human-like cues and perceived reliability. Scientific Reports 16:7860. https://doi.org/10.1038/s41598-026-38179-2

Xu W, Huang C, Gao S, et al (2025) LLM-Based Agents for Tool Learning: A Survey. Data Science and Engineering 10:533–563. https://doi.org/10.1007/s41019-025-00296-9

Yang J, Tan R, Wu Q, et al (2025) Magma: A Foundation Model for Multimodal AI Agents. CoRR abs/2502.13130. https://doi.org/10.48550/arXiv.2502.13130, URL https://arxiv.org/abs/2502.13130, arXiv:2502.13130 [cs.CV]

Zhou Y, Li X, Liu Y, et al (2025) M2PA: A Multi-Memory Planning Agent for Open Worlds Inspired by Cognitive Theory. In: Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp 23204–23220, https://doi.org/10.18653/v1/2025.findings-acl.1191, URL https://aclanthology.org/2025.findings-acl.1191/