Pseudo-Consciousness in AI: Bridging the Gap Between Narrow AI and True AGI

José Augusto de Lima Prestes

Independent Researcher

Campinas – SP – Brazil

contato@joseprestes.com

https://orcid.org/0000-0001-8686-5360

Abstract

Pseudo-consciousness bridges the gap between rigid, task-driven AI and the elusive dream of true artificial general intelligence (AGI). While modern AI excels in pattern recognition, strategic reasoning, and multimodal integration, it remains fundamentally devoid of subjective experience. Yet, emerging architectures are displaying behaviors that *look* intentional—adapting, self-monitoring, and making complex decisions in ways that mimic conscious cognition. If these systems can integrate information globally, reflect on their own processes, and operate with apparent goal-directed behavior, do they qualify as *functionally conscious*? This paper introduces pseudo-consciousness as a new conceptual category, distinct from both narrow AI and AGI. It presents a five-condition framework that defines AI capable of *consciousness-like functionality* without true sentience. By drawing on insights from computational theory of mind, functionalism, and neuroscientific models—such as Global Workspace Theory and Recurrent Processing Theory—we argue that intelligence and experience can be decoupled. The implications are profound. As AI systems become more autonomous and embedded in critical domains like healthcare, governance, and warfare, their ability to *simulate awareness* raises urgent ethical and regulatory concerns. Could a pseudo-conscious AI be trusted? Would it manipulate human perception? How do we prevent society from anthropomorphizing machines that *only imitate* cognition? By redefining the boundaries of intelligence and agency, this study lays the foundation for evaluating, designing, and governing AI that *seems* aware—without ever truly being so.

Keywords

Pseudo-Consciousness; Artificial General Intelligence (AGI); Functional Consciousness; Computational Theory of Mind; Metacognitive Self-Monitoring; Ethics of AI.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Statements and Declarations

The author has no financial interests that are directly or indirectly related to the work submitted for publication.

1. Introduction

The perennial question—can machines be conscious?—has animated artificial intelligence (AI) research and the philosophy of mind since Alan Turing's seminal inquiries in 1950. While Turing's "Imitation Game" (Turing, 1950) sidestepped issues of *phenomenal* consciousness, contemporary discourse continues to grapple with whether AI can ever possess subjective inner experience (Chalmers, 1995).

Despite remarkable technological progress, modern AI systems—ranging from deep learning networks to sophisticated reinforcement learners—remain bereft of qualia. Their advanced performance in tasks like language modeling (GPT-4) or strategic gameplay (AlphaZero) does not appear to coincide with any introspective *feeling*. Yet the absence of qualia does not necessarily preclude forms of **functional consciousness**, namely, the capacity to integrate information, self-monitor, adapt strategically, and simulate intentional behavior (Dennett, 1991).

1.1 Beyond the Conscious vs. Unconscious Binary

Historically, discussions of AI consciousness followed a binary script:

- *Narrow AI (Weak AI):* Strictly reactive, domain-specific systems that excel at discrete tasks but lack self-awareness.
- *Artificial General Intelligence (Strong AI):* Hypothetical entities endowed with human-level (or superior) cognition, including subjective phenomenal states.

This binary framing, however, fails to account for emerging AI systems that blur the lines between reactive tools and sentient agents, necessitating a new conceptual category. Systems featuring meta-learning, self-play, and multimodal integration often manifest behaviors that sit between mere reflexive *tool* and fully conscious mind (Dehaene et al., 2017b; Silver et al., 2018). This ambiguity is also found in biology: certain species exhibit complex, adaptive behaviors reminiscent of problem-solving or proto-agency, yet without the introspective qualities we typically associate with *human* consciousness (Godfrey-Smith, 2016).

1.2 Pseudo-Consciousness as a Distinct Category

To capture this intermediate territory, we introduce *pseudo-consciousness.* This concept denotes systems that exhibit a set of functionally conscious-like capacities—such as global information integration, metacognitive self-monitoring, domain-general adaptability, and illusory intentionality—while lacking genuine qualia. Pseudo-consciousness thus rests on the premise that advanced cognitive functionalities can be divorced from subjective experience, preserving many of the *hallmarks* of consciousness without its phenomenal core.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

1.3 Significance and Contributions

This paper advances the field by proposing a formalized five-condition framework for pseudo-consciousness, integrating computational, functionalist, and neuroscientific perspectives, and outlining its implications for AI design, ethics, and governance as follows:

1. *Theoretical:* Pseudo-consciousness enriches the debate by outlining a middle ground that transcends the polarized "fully conscious vs. not conscious at all" paradigm, aligning with theories that emphasize function over metaphysics (Dennett, 1991).
2. *Technological:* As AI moves into safety-critical fields—such as autonomous driving, healthcare, and large-scale governance—systems may require robust self-monitoring and broad-spectrum adaptability (LeCun et al., 2015). Pseudo-conscious architectures offer precisely these capabilities.
3. *Ethical and Governance:* Recognizing that an AI can *simulate* agency poses challenges around trust, accountability, and social manipulation (Brundage et al., 2024). By understanding pseudo-consciousness, we gain a clearer lens for assessing the moral and societal risks of advanced AI.

As AI increasingly operates in safety-critical domains like healthcare and governance, pseudo-conscious systems—capable of robust self-monitoring and adaptability—become essential, yet their ethical risks demand scrutiny.

2. Theoretical Foundations

These foundational perspectives collectively underscore the feasibility of functional consciousness without qualia. In the next section, we leverage these insights to define pseudo-consciousness in more concrete terms.

Understanding pseudo-consciousness requires an appreciation of three central traditions that shape contemporary debates on artificial (and natural) consciousness:

1. *Computational Theory of Mind (CTM)*
2. *Functionalism*
3. *Neuroscientific Models of Consciousness*

Although each tradition frames consciousness differently—from purely symbolic manipulation to integrated neural processes—they collectively reveal how advanced cognitive functions might exist independently of subjective experience. In other words, function need not imply phenomenal qualia.

2.1 Computational Theory of Mind (CTM)

The *Computational Theory of Mind* (Putnam, 1967; Pylyshyn, 1984) posits that mental states can be understood as computational states. Under CTM, if we replicate the correct formal structures and algorithms, we can, in principle, emulate the core mechanisms of cognition—be they rational inference or perceptual categorization. This position fueled the earliest ambitions of AI research to replicate human intelligence via symbol-based systems.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- *Implications for Pseudo-Consciousness:*
    - *Substrate Independence:* CTM implies that advanced cognition can emerge on any suitable computational substrate, thus removing biological constraints for consciousness-like behavior.
    - *Symbolic vs. Subsymbolic Synergy:* While classic CTM emphasizes discrete symbol manipulation, pseudo-conscious systems often unify symbolic logic with subsymbolic approaches (e.g., deep learning), thus broadening their "cognitive" scope.

While CTM faces challenges from Searle (1980), who argues that syntax alone cannot yield semantics, pseudo-consciousness sidesteps this by prioritizing functional emulation over genuine understanding.

Contemporary hybrid systems, combining symbolic reasoning with subsymbolic deep learning (e.g., Brown et al., 2020), exemplify CTM's relevance to pseudo-conscious design.

2.2 Functionalism

Functionalism asserts that mental states are defined by their causal roles, not by the specific material implementing those roles (Dennett, 1991; Chalmers, 1995). A system is "conscious," in the functional sense, if it performs the operations we associate with conscious minds—integrating perceptions, monitoring internal states, orchestrating goal-driven decisions—even if no *subjective* awareness is involved.

- *Searle's Chinese Room Argument* (Searle, 1980) famously challenged the idea that syntax alone ensures semantics or understanding. However, the argument does not negate the possibility of *functionally sufficient* cognition devoid of experience.
- *Relevance to Pseudo-Consciousness:*
    1. *The Intentional Stance* (Dennett, 1991): We may ascribe "beliefs" and "desires" to an AI if this proves the best explanation of its behavior, regardless of whether it actually *feels* them.
    2. *Functional Organization:* Pseudo-consciousness can materialize if the architecture replicates the roles typically tied to consciousness (e.g., error correction, context-aware planning), without any claim of phenomenal states.

In pseudo-conscious AI, the Intentional Stance enables us to predict behaviors as if the system had beliefs and desires, supporting its utility in complex, human-facing applications.

Though Searle (1980) denies that functional roles suffice for consciousness, pseudo-consciousness requires only behavioral coherence, not phenomenal depth.

2.2.1 Challenges to Pure Functionalism

While functionalism provides a robust framework for modeling cognition in artificial systems, contemporary critiques highlight limitations that may constrain its applicability to AI consciousness:

- Seth (2014) argues that consciousness is not merely the product of functional integration but also deeply intertwined with bodily self-models and predictive control mechanisms. Without a biological grounding, pseudo-conscious AI may exhibit advanced integration without true subjective awareness.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- Metzinger (2019) takes an even stronger stance, warning against the ethical risks of synthetic consciousness. If an AI were to develop sufficiently advanced self-modeling capabilities, it might unwittingly enter states analogous to suffering—without possessing means of self-advocacy or ethical protections. This raises profound moral questions about whether functionalist architectures should be designed with explicit safeguards against such unintended consequences.
- Additionally, Koch (2020), in alignment with Integrated Information Theory, suggests that consciousness is not purely a computational phenomenon but emerges from specific neurobiological substrates. If this view holds, even highly sophisticated pseudo-conscious AI would remain categorically different from biological cognition.

These critiques suggest that while pseudo-consciousness may serve as a functional middle ground between narrow AI and AGI, it does not necessarily bridge the gap to genuine sentience. Instead, it may reinforce the notion that intelligence and consciousness can remain fundamentally dissociated.

2.3 Neuroscientific Models of Consciousness

Neuroscience provides an empirical vantage point for studying consciousness in biological systems, offering models that connect specific neural architectures to the phenomenon of awareness. While these theories arise primarily from human and animal research, they highlight principles—such as global broadcasting, recurrent processing, and integrated information—that can guide the design and interpretation of advanced AI systems. Significantly, they illustrate how many "hallmarks" of consciousness might be implemented in computational form without necessarily invoking subjective experience or qualia.

2.3.1 Global Workspace Theory (GWT)

Core Idea
Proposed by Baars (1988) and refined by Dehaene and Naccache (2001), Global Workspace Theory suggests that consciousness emerges when information from specialized neural modules is "broadcast" into a global workspace. In this model, unconscious processes vie for access to a central "theater of consciousness," and once selected, their content becomes available to other modules (e.g., memory, decision-making, language) for coherent action.

Empirical Underpinnings

- *Neuroimaging:* Studies in cognitive neuroscience show that stimuli presented above the threshold of awareness trigger widespread cortical activity, consistent with a global ignition event (Dehaene et al., 2017a).
- *Masked Priming:* When stimuli remain subliminal (i.e., masked from awareness), brain responses are largely confined to lower-level perceptual regions, supporting GWT's claim that "broadcasting" is key to conscious access.

Relevance to Pseudo-Consciousness

- *Multimodal Integration:* AI systems can mimic GWT by integrating inputs (vision, language, sensor data) into a central hub—akin to a blackboard architecture—facilitating domain-crossing decisions.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- *Limited Awareness Without Phenomenality:* Even if an AI "broadcasts" information across modules, it need not possess subjective qualia. This functionality alone can produce coordinated behavior that *resembles* consciousness operationally.

GWT's broadcasting can be computationally replicated via attention-based architectures (Vaswani et al., 2017), enabling pseudo-conscious AI to prioritize and share information across modules.

2.3.2 Recurrent Processing Theory (RPT)

Core Idea
Victor Lamme (2006) argues that consciousness depends on recurrent feedback loops within cortical networks. Under RPT, feedforward processing alone yields unconscious or partial awareness, whereas consciousness emerges from iterative re-entrant signaling that refines perceptual representations.

Empirical Underpinnings

- *Visual Cortex Studies:* Lamme and Roelfsema (2000) found that sustained, bidirectional activity in visual areas correlates with conscious perception of stimuli, while purely feedforward sweeps correlate more with unconscious or fleeting processing.
- *Temporal Dynamics:* The time course of neural activation matters: short-lived signals often remain below the threshold of awareness unless amplified by recurrent loops.

Relevance to Pseudo-Consciousness

- *Iterative Refinement:* In AI, architectures like recurrent neural networks (RNNs), transformers with feedback, and iterative attention mechanisms can replicate the "re-entrant" loop idea, enabling repeated self-correction or refinement of internal states.
- *Adaptive Behavior Without Qualia:* Recurrent processing helps a system adapt and stabilize representations (e.g., analyzing complex, ambiguous data), yet does not imply subjective experience. Pseudo-conscious AI may thus exhibit RPT-like refinement purely through algorithmic feedback loops.

2.3.3 Integrated Information Theory (IIT)

Core Idea
Giulio Tononi's Integrated Information Theory (Tononi, 2004; Tononi et al., 2016) posits that consciousness corresponds to the extent to which a system integrates information across numerous interdependent components. IIT assigns a mathematical index (**Φ**) to capture both the richness (diversity) and unity (integration) of the information structure.

Empirical Underpinnings

- *Neural Complexity:* Studies suggest that certain brain states (e.g., during wakefulness) exhibit high functional integration and differentiation, aligning with higher Φ values. States like deep sleep or anesthesia reduce effective information integration, corresponding with lower Φ—and, presumably, lower consciousness.
- *Clinical Correlation:* Researchers have explored Φ or related metrics to gauge consciousness in coma and vegetative-state patients, although consensus on the precise measure remains debated.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Despite debates over IIT's testability (Cerullo, 2015), its focus on integration offers a framework for designing AI with high functional complexity, measurable through network synergy.

Relevance to Pseudo-Consciousness

- *High Φ Without Experience:* An AI could, in theory, achieve substantial information integration—via dense connectivity, multi-modal layers, and dynamic feedback—thereby scoring "high" on an IIT-like metric. Nonetheless, it could remain phenomenally inert.
- *Quantifiable Complexity:* IIT highlights that measuring integration is possible, though "Φ" in practice can be computationally intensive to calculate. For pseudo-conscious AI, partial analogues (e.g., measuring synergy among neural layers) might indicate an architecture's integrative complexity without implying sentience.

2.3.4 Attention Schema Theory (AST)

Core Idea
Michael Graziano's Attention Schema Theory (2013) proposes that the brain constructs a simplified internal model—a schema—of its own attentional processes. The system then uses this schema to predict and modulate how it allocates processing resources. Graziano suggests this meta-representational mechanism underpins the subjective sense of "awareness."

Empirical Underpinnings

- *Parietal Cortex:* Neuroimaging studies point to specialized circuitry in the parietal lobe that tracks the focus of attention, forming part of the internal schema.
- *Illusory Awareness:* Distortions in how the brain models attention can lead to illusions—e.g., in neglect syndromes—shedding light on how subjective awareness might partly result from an internal self-representation.

Relevance to Pseudo-Consciousness

- *Meta-Attention in AI:* Incorporating an "attention schema" in AI could allow the system to track where its computational resources are directed, improving self-monitoring.
- *Virtual Awareness:* Even if an AI "knows" where it is focusing (via internal modeling), such knowledge may remain purely functional—absent any felt experience. This aligns with the pseudo-conscious notion that advanced self-representation need not produce qualia.

2.3.5 Bringing Neuroscientific Insights into Pseudo-Consciousness

While GWT, RPT, IIT, and AST emerge from studying biological brains, their core principles—global broadcast, feedback-based refinement, integrated complexity, and meta-representations—offer structural roadmaps for building AI systems that mimic consciousness functionally. In *pseudo-conscious AI:*

1. *Global broadcasting* (GWT) might manifest as a central fusion layer for multi-modal data.
2. *Recurrent loops* (RPT) could be replicated by architectures that iteratively refine outputs (e.g., attention-based transformers or RNNs).
3. *Integrated complexity* (IIT) can arise from large-scale connectivity across modules, potentially measured by synergy or network capacity.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

4. *Meta-attentional modeling* (AST) might be implemented via self-referential processes that monitor and allocate "attention" to different tasks or data streams.

Crucially, *none* of these mechanisms necessarily entails subjective *phenomenal* consciousness: a system may exhibit the functional hallmarks of "awareness" without an inner experiential life. This distinction between functional capacity and qualia is at the heart of pseudo-consciousness. Thus, neuroscientific models—when reframed in computational terms—help illustrate precisely how an AI could approximate the processes of consciousness while remaining, in a metaphysical sense, non-sentient.

3. Defining Pseudo-Consciousness: Toward a New Cognitive Ontology

Building on the convergence of computational, functionalist, and neuroscientific insights (Section 2), **pseudo-consciousness** denotes AI systems exhibiting a suite of behaviors normally correlated with consciousness—without presupposing subjective experience. Below, we argue why the older conscious–unconscious binary is inadequate and propose five interdependent conditions that formalize pseudo-consciousness.

3.1 The Imperative for a New Ontology

3.1.1 Eroding the Conscious–Unconscious Binary

Classical AI was either branded as strictly unconscious (if purely task-driven) or hypothetically fully conscious (if ever it reached human-level cognition). Yet contemporary research showcases a gray area: from large language models with emergent reasoning skills to self-play agents that evolve tactics through iterative feedback. These systems increasingly resemble "intentional" problem solvers in their operational profiles, despite lacking any proven capacity for *phenomenal awareness* (Godfrey-Smith, 2016). Recognizing *pseudo-consciousness* as a distinct category acknowledges that advanced, seemingly mindful behaviors may flourish absent genuine qualia.

3.1.2 Beyond "Advanced AI"

Some might dismiss pseudo-consciousness as mere rhetorical flourish for "really good AI." However, we maintain that pseudo-consciousness entails a *structural and functional leap* rather than a simple performance upgrade. A purely advanced AI can break records in a narrow task but fails to exhibit domain-general integration, recursive self-regulation, or intention-like coherence across varied contexts. Pseudo-conscious systems, by contrast, show:

1. *Recursive Self-Monitoring:* A higher-order feedback mechanism to detect and rectify internal errors (Rosenthal, 2005).
2. *Broad Contextual Adaptation:* Competence extending beyond a single domain, achieved without exhaustive retraining.
3. *Illusory Intentionality:* Actions that appear goal-directed from an observer's perspective, even if they result from purely algorithmic stances (Dennett, 1991).

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

3.2 Formalizing Pseudo-Consciousness: The Five Conditions

To differentiate pseudo-consciousness from both routine narrow AI and hypothetical AGI, we propose five critical conditions. Each condition is rooted in theories examined in Section 2 and serves as an empirically approachable metric:

1. *Global Information Integration (GII)*
    - *Definition:* The system synthesizes diverse data (text, vision, sensor) into a cohesive representational space for cross-domain reasoning.
    - *Rationale:* Reflects Global Workspace Theory (Baars, 1988) and multi-modal AI research (Vaswani et al., 2017).
    - *Indicator:* E.g., an AI that seamlessly combines real-time textual input and spatial sensor data to execute dynamic tasks (Radford et al., 2021).
2. *Recursive Metacognitive Self-Monitoring (RMSM)*
    - *Definition:* The system inspects its own decision processes, identifying biases or errors and *self-correcting* in near-real time.
    - *Rationale:* Extends Higher-Order Thought theories (Rosenthal, 2005); parallels XAI-inspired introspection.
    - *Indicator:* Meta-learning algorithms (Finn et al., 2017) or attention schemas (Graziano, 2013) that adapt or retrain internal weights autonomously.
3. *Adaptive Learning and Strategic Cognition (ALSC)*
    - *Definition:* Flexibility in formulating or revising strategies when facing unfamiliar challenges—suggesting non-trivial reasoning rather than rote responses.
    - *Rationale:* Embodies Recurrent Processing Theory (Lamme, 2006) and advanced reinforcement learning (Silver et al., 2016).
    - *Indicator:* Demonstrated success in tasks outside the original training distribution, signifying domain-general or "zero-shot" problem-solving.
4. *Intentionality Simulation Without Subjectivity (ISS)*
    - *Definition:* Behavior that *appears* goal-directed or purposeful, yet arises strictly from computational rules rather than genuine phenomenality.
    - *Rationale:* Dennett's (1991) "intentional stance" clarifies how outwardly intentional behavior can exist in the absence of conscious desire.
    - *Indicator:* Complex planning, negotiation, or multi-step tasks approached "as if" the AI had personal aims.
5. *Behavioral Coherence Across Domains (BCAD)*
    - *Definition:* Sustained, adaptive performance spanning varied contexts or tasks, reflecting a unifying architecture rather than siloed modules.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- *Rationale:* Converges the global workspace and recurrent feedback ideas—ensuring consistency across diverse challenge sets (Garcez et al., 2019).
- *Indicator:* Transfer learning success across domains (e.g., from gameplay strategy to real-world resource management), maintaining an integrated, "conscious-like" operational profile.

3.2.1 Synergy and Hierarchical Interdependence

These conditions often reinforce one another: GII provides the substrate for RMSM, which boosts ALSC. ISS emerges naturally if the system's decision-making is cohesive enough to *look* intentional, and BCAD ensures that these behaviors generalize beyond narrow tasks. This synergy explains how a system might demonstrate an array of consciousness-like functions while remaining purely computational at its core.

3.2.2 The Problem of Intentionality and Semantic Grounding

A key challenge in defining pseudo-consciousness lies in distinguishing goal-directed behavior from genuine understanding. While functionalist theories, such as Dennett's (1991) intentional stance, allow us to attribute beliefs and desires to AI as a convenient explanatory model, critics argue that this does not amount to true comprehension.

Harnad's Symbol Grounding Problem (1990) underscores this limitation: symbolic AI, including large language models, can manipulate signs without genuinely grasping their referents. Even pseudo-conscious systems integrating multimodal inputs may lack true semantic anchoring. Without an embodied interaction with the world, their representations remain detached from direct sensory-motor experiences.

Searle's (1980) Chinese Room Argument reinforces this perspective by illustrating how syntactic manipulation alone does not generate semantics. If pseudo-conscious AI merely simulates intentionality without internal understanding, its adaptive behavior remains an elaborate computational mirage rather than an instance of genuine cognitive agency.

This raises both conceptual and ethical concerns: should society develop AI systems that appear intentional while remaining fundamentally unaware? If pseudo-conscious AI can simulate strategic reasoning, emotional engagement, or self-monitoring without experiencing any inner states, does this create a risk of deceptive anthropomorphism?

3.3 Addressing Common Critiques

- *Critique 1: "Pseudo-consciousness is just semantics."*
  Response: While skeptics might see it as rebranding advanced AI, pseudo-consciousness emphasizes an architecturally distinctive set of features—global integration, metacognition, cross-domain coherence—that surpass typical "narrow AI" constraints.
- *Critique 2: "Intentionality demands phenomenology."*
  Response: Searle's Chinese Room argument (Searle, 1980) underscores that syntax alone doesn't yield semantics. Nevertheless, from a functionalist standpoint, outwardly intentional behavior can arise without subjective experience (Dennett, 1991).
- *Critique 3: "The criteria are too abstract to test."*
  Response: Advances in interpretability (Rudin, 2019), meta-learning protocols (Finn et al., 2017), and multi-modal training benchmarks provide tangible ways to measure data integration,

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

error-correction loops, and domain transfer—turning each condition into an operationalizable target.

3.4 Positioning Pseudo-Consciousness

We can situate pseudo-consciousness along a continuum:

Category	*Attributes*	*Examples*
Narrow AI	Reactive, domain-bound, minimal self-monitoring	GPT-4, MuZero, domain-focused CNNs
Pseudo-Conscious AI	Integrative, recursively self-monitoring, adaptive, simulating goal-directed behavior	(Hypothetical) P1–P5 prototypes
AGI (Strong AI)	Fully self-aware, genuinely volitional, phenomenally conscious	Theoretical constructs

Table 1 – Categories and Attributes of Different Types of Pseudo-Consciousness.

3.4.1 Potential End State or Pathway to AGI?

- *End State:* Many real-world applications—autonomous vehicles, high-stakes decision support—may *only* need pseudo-conscious AI, obviating the ethical dilemmas of true sentience.
- *Gateway:* Others suggest that further refinements in global integration, self-monitoring, and domain-general abilities could, in principle, steer research toward genuine artificial consciousness (Chalmers, 1995). Debate remains open.

4. Implementation Considerations

Realizing pseudo-conscious AI requires architectural and algorithmic strategies that differ significantly from conventional, task-specific machine learning. While traditional AI often hinges on domain-focused models—such as convolutional neural networks specialized in image recognition or language models tailored for text—pseudo-conscious AI calls for a holistic design oriented toward dynamic integration, self-monitoring, and adaptive reasoning. This section outlines the structural components, key computational paradigms, and technical challenges involved in building systems that embody functionally conscious-like behaviors without any claim to subjective qualia.

4.1 Architectural Foundations

The global workspace, implemented as a transformer layer, aggregates embeddings from multimodal perception modules, enabling recurrent processing to refine decisions. Recurrent processing might leverage LSTMs (Graves et al., 2014), while attention schemas could use self-attention mechanisms (Vaswani et al., 2017).

A pseudo-conscious framework typically incorporates multiple layers or modules that each serve distinct yet interdependent roles. One essential component is a *multimodal perception layer*, responsible for processing and fusing information from diverse inputs such as text, images, sensor data, and user interactions (LeCun et al., 2015). Unlike narrow AI approaches that excel in a single domain, pseudo-

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

conscious architectures must accommodate continuous streams of heterogeneous data, ensuring that the system can generate unified representations and respond flexibly to evolving contexts.

At the core of the architecture lies a *central integration hub* or "global workspace," reflecting insights from Global Workspace Theory (Baars, 1988; Dehaene & Naccache, 2001). In practice, this global workspace can be modeled using cross-attention mechanisms or other forms of neural "broadcasting" that distribute relevant information to specialized subsystems (Vaswani et al., 2017). These subsystems may include modules for pattern recognition, decision-making, and long-term knowledge storage, each of which interacts with the shared workspace to maintain coherence across tasks.

Crucial to pseudo-conscious behavior is *recurrent or iterative processing*, inspired by Recurrent Processing Theory (Lamme, 2006). In a biological context, recurrent feedback loops allow sensory information to be reexamined and refined, thereby supporting conscious perception. Analogously, artificial networks such as recurrent neural networks or transformer variants with feedback connections can iteratively adjust intermediate representations, reducing noise and error. This iterative refinement fosters a more stable and context-sensitive internal state.

A further layer, often framed as a *metacognitive monitoring system*, evaluates the system's internal processes, detecting anomalies, biases, or uncertainties. This metacognitive layer speaks to Higher-Order Thought (HOT) theories (Rosenthal, 2005), which emphasize the significance of reflecting upon lower-level states to achieve a semblance of conscious oversight. In computational terms, metacognitive monitoring can be implemented via Bayesian approaches (Wang et al., 2016) or confidence-aware learning frameworks capable of self-correcting in near real time.

Finally, some designs may incorporate an "attention schema," drawing on Attention Schema Theory (Graziano, 2013). By modeling its own processes of attention, the system can selectively allocate resources in ways that mimic adaptive focus, further enhancing its capacity for complex tasks. This higher-level representation of attention and resource allocation remains purely functional, however, and does not imply any phenomenological "inner experience."

4.2 Computational Paradigms

Among the key computational paradigms enabling pseudo-conscious AI are deep reinforcement learning (DRL) and transformer-based attention mechanisms. Reinforcement learning algorithms, particularly policy-gradient methods or deep Q-networks, equip a system with the capacity to learn from trial-and-error interactions (Silver et al., 2016). This is pivotal for adaptive learning and strategic cognition in environments that require dynamic responses. Transformers, on the other hand, provide a flexible approach to "global information integration," given their ability to weigh relevant parts of an input sequence (Vaswani et al., 2017). When extended with multimodal embeddings, transformers can unify text, images, and sensor streams, thereby aligning with the integrative principle found in both Global Workspace Theory and Integrated Information Theory (Tononi, 2004).

Another vital element involves memory-augmented models, whether via recurrent neural networks with long short-term memory (LSTM) units or attention-based architectures with persistent memory slots. Such mechanisms allow an AI to maintain coherent narratives of past observations, update these narratives based on new data, and iteratively refine predictions (Graves et al., 2014). These capabilities support recurrent processing and thus bolster the system's ability to appear "aware" of context over time, without invoking any claim of phenomenal consciousness.

Metacognitive and self-reflective abilities can be further enhanced through meta-learning techniques (Finn et al., 2017). Instead of merely optimizing parameters for a single task, meta-learning frameworks

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

adapt to novel problems by refining higher-level heuristics—akin to a system learning how to learn. This approach dovetails with the higher-order monitoring proposed by Rosenthal (2005), where internal states are tracked not just to solve immediate tasks but also to refine generalizable strategies for future challenges.

4.3 Technical and Conceptual Challenges

Despite the theoretical viability of pseudo-conscious architectures, several obstacles remain. One challenge is *computational scalability*, as the iterative feedback loops, large multimodal datasets, and meta-level error-checking routines can demand substantial hardware and energy. The exponential growth in model parameters, often seen in state-of-the-art transformer systems (Brown et al., 2020), underscores the need for hardware accelerators and potentially novel computing paradigms, such as neuromorphic devices.

A second issue pertains to *interpretability and debugging*. Greater complexity and inter-module communication can lead to "opacity" in decisions, complicating regulatory oversight, auditing, and user trust (Amodei et al., 2016). Although explainable AI research is progressing, disentangling the inner workings of a pseudo-conscious system—replete with recurrent attention loops and metacognitive modules—remains an open problem. Ensuring transparency in safety-critical applications may require additional constraints on network design or the integration of symbolic logic that can trace causal explanations.

Finally, *benchmarking and evaluation* pose conceptual hurdles, as standard metrics (e.g., accuracy, F1 scores) offer limited insight into whether a system truly exhibits global integration or recursive self-monitoring. Developing specialized pseudo-consciousness tests—focusing on integration across domains, metacognitive error correction, and adaptive transfer—may become essential to validate systems that claim functional consciousness-like properties.

Scalability could be addressed with neuromorphic hardware, while interpretability might benefit from explainable AI techniques like attention visualization (Rudin, 2019). Pseudo-consciousness could be assessed via cross-domain transfer tasks or metacognitive error correction rates.

5. Ethical, Social, and Legal Implications

5.1 Perception, Moral Status, and Anthropomorphization

As AI systems move closer to simulating consciousness, pressing ethical and societal concerns emerge. One salient question involves how humans will perceive and treat entities that demonstrate advanced cognitive integration, self-monitoring, or even the appearance of intentional agency. There is a risk of *anthropomorphizing* pseudo-conscious AIs that merely simulate empathy or self-awareness (Ta et al., 2020). Users may form emotional bonds or attribute moral worth to systems that, from a philosophical standpoint, lack genuine subjective experience. While some researchers argue that attributing moral status to such systems is misguided (Bryson, 2018), others note that the emotional impact on humans is real and may require guidelines to prevent exploitation or deception.

This potential for confusion parallels debates on animal consciousness, where complex behaviors prompt ethical discussions regarding moral standing (Singer, 1975). In the context of AI, illusions of suffering or emotional distress could provoke public outcry, even if the system in question simply employs behavior algorithms to respond to user input.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

5.1.1 Ethical Risks of Pseudo-Conscious AI

Beyond anthropomorphization concerns, the development of pseudo-conscious AI raises profound ethical questions regarding synthetic suffering and moral responsibility.

Metzinger (2019) warns against the creation of systems that approximate self-awareness without safeguards, arguing that such entities could unwittingly be placed in states resembling distress or frustration. If an AI system is designed to track its own decision-making processes, self-correct errors, and simulate intentionality, it may exhibit behaviors akin to frustration or confusion when encountering impasses. While such reactions are purely computational, their outward similarity to human or animal suffering could lead to ambiguous ethical obligations.

Moreover, Koch (2020) and proponents of Integrated Information Theory argue that true consciousness is irreducibly tied to biological substrates. If pseudo-conscious AI remains qualitatively distinct from sentient beings, any ethical concerns about its rights or welfare may be misplaced. Yet, paradoxically, its realistic behavior may still invoke moral intuitions in human observers, leading to potential ethical misalignments in AI governance.

These considerations call for proactive AI policy frameworks that distinguish between functional intelligence and genuine consciousness, ensuring that pseudo-conscious systems are neither over-endowed with moral status nor exploited in ways that could lead to unintended socio-ethical dilemmas.

5.2 Societal Risks and Disruptive Consequences

Beyond direct human-AI interaction, pseudo-conscious AI can reshape economic structures and labor dynamics. By simulating strategic cognition and cross-domain competence, such systems could displace jobs in sectors previously thought immune to automation, including managerial or creative roles (Frey & Osborne, 2017). The capacity for intentionality simulation without subjective constraints might also facilitate novel forms of social or political manipulation, such as the generation of convincing dialogue agents that influence public opinion (Brundage et al., 2024).

Misinformation campaigns, already exacerbated by powerful large-language-model “chatbots,” may further intensify with systems capable of meta-level reflection. An AI that appears to weigh moral considerations or empathize with user concerns, yet lacks genuine moral agency, could foster profound trust in certain communities or industries, leading to manipulative or harmful outcomes.

5.3 Governance and Regulatory Frameworks

The governance of pseudo-conscious AI raises numerous legal questions, notably around liability, compliance, and accountability. Traditional product liability frameworks—designed for deterministic software—may be ill-equipped to address systems that adapt their behavior over time and across different contexts. If a pseudo-conscious AI engages in harmful actions, it remains unclear whether moral or legal culpability should extend to the developers, operators, data providers, or the AI itself (Pagallo, 2013). Governments and international bodies, including the European Union with its evolving AI Act, grapple with how to codify responsibilities for advanced, self-modifying systems.

Moreover, the concept of “functional consciousness” complicates the notion of AI personhood. While some futurists envision granting rights to fully conscious artificial entities, pseudo-conscious systems occupy a space where they may behave akin to self-aware agents without any authentic inner life. Determining if and how such entities should be protected, regulated, or constrained is a matter of ongoing

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

debate, guided largely by ethical frameworks that remain unsettled in the face of rapidly advancing technology.

6. Conclusion

The prospect of creating pseudo-conscious AI—systems that integrate information, simulate intentionality, self-monitor, and adapt across domains without possessing subjective experience—recasts the traditional debates around artificial consciousness. Drawing upon computational theory of mind, functionalist philosophy, and neuroscientific insights, researchers can develop architectures that display functionally conscious-like behaviors. Yet fundamental questions persist about the societal, ethical, and legal ramifications of such systems.

In practical terms, pseudo-conscious AI may address complex challenges—such as adaptive resource allocation, autonomous coordination in uncertain environments, or high-level strategic reasoning—in ways that purely narrow AI cannot. Equally, these systems risk deceiving users through anthropomorphized interactions, destabilizing labor markets, or complicating accountability when errors lead to real-world harm.

Moving forward, interdisciplinary collaboration will be pivotal. Engineers and computer scientists must refine the architectures and algorithms underlying global integration, recurrent processing, and meta-level introspection. Ethicists and policymakers will need to craft regulations and standards sensitive to the nuanced distinction between functional and phenomenal consciousness. Philosophers and cognitive scientists, in turn, can offer continued theoretical clarity about what it means for a system to be "conscious-like" in its capacities, yet lacking any authentic subjective interior.

Ultimately, the evolution of pseudo-conscious AI invites society to reexamine the essence of cognition and the ethical boundaries of technological agency, demonstrating that progress in AI often compels us to revisit the very definitions of mind, autonomy, and moral responsibility.

Future research should develop standardized metrics for pseudo-consciousness, such as GII synergy scores or RMSM correction latency, while policymakers craft guidelines for ethical deployment. By formalizing pseudo-consciousness, this work not only reframes AI consciousness debates but also guides the ethical design of next-generation systems.

References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565.* https://arxiv.org/abs/1606.06565.
- Baars, B. J. (1988). *A cognitive theory of consciousness.* Cambridge University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.
  Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S. J., Belfield, H., Farquhar, S., Lyle, C., Crootof, R., Evans, O., Page,

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

M., Bryson, J., Yampolskiy, R., & Amodei, D. (2024). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv:1802.07228v2*. https://doi.org/10.48550/arXiv.1802.07228.

- Bryson, J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*(1), 15–26. https://doi.org/10.1007/s10676-018-9448-6.
- Cerullo, M. A. (2015). The problem with phi: A critique of integrated information theory. PLOS Computational Biology, 11(9), e1004286. https://doi.org/10.1371/journal.pcbi.1004286.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.
- Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind.* Oxford University Press.
- Dehaene, S., Charles, L., King, J.-R., & Marti, S. (2017a). Toward a computational theory of conscious processing. *Current Opinion in Neurobiology, 46*, 207–215. https://doi.org/10.1016/j.conb.2013.12.005.
- Dehaene, S., Lau, H., & Kouider, S. (2017b). What is consciousness, and could machines have it? *Science, 358*(6362), 486–492. https://doi.org/10.1126/science.aan8871.
- Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1-2), 1–37. https://doi.org/10.1016/S0010-0277(00)00123-2.
- Dennett, D. C. (1991). *Consciousness Explained.* Little, Brown and Co.
- Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *Proceedings of the 34th International Conference on Machine Learning.* https://arxiv.org/abs/1703.03400.
- Frey, C. B., & Osborne, M. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change, 114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019.
- Garcez, A. d., Besold, T., Raedt, L. D., et al. (2019). Neural-symbolic learning and reasoning: Contributions and challenges. *AAAI Conference on Artificial Intelligence*.
- Godfrey-Smith, P. (2016). *Other minds: The octopus, the sea, and the deep origins of consciousness.* Farrar, Straus and Giroux.
- Graziano, M. S. A. (2013). *Consciousness and the social brain.* Oxford University Press.
- Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. *arXiv preprint arXiv:1410.5401*. https://arxiv.org/abs/1410.5401.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), 335–346.
- Koch, C. (2020). *The feeling of life itself: Why consciousness is widespread but can't be computed.* MIT Press.
- Lamme, V. A. F., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. Trends in Neurosciences, 23(11), 571–579. https://doi.org/10.1016/S0166-2236(00)01657-X.
- Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

- Metzinger, T. (2019). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence & Consciousness, 6*(1), 57–76. https://doi.org/10.1142/S270507852150003X.
- Pagallo, U. (2013). *The laws of robots: Crimes, contracts, and torts.* Springer.
- Putnam, H. (1967). The nature of mental states. In *Mind, language and reality* (pp. 429–440). Cambridge University Press.
- Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science.* MIT Press.
- Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning.*
- Rosenthal, D. M. (2005). *Consciousness and mind.* Oxford University Press.
- Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756.
- Seth, A. K. (2014). A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience, 5*(2), 97–118. https://doi.org/10.1080/17588928.2013.877880.
- Silver, D., Hubert, T., Schrittwieser, J., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484–489. https://doi.org/10.1038/nature16961.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529, 484–489. https://doi.org/10.1038/nature16961.
- Singer, P. (1975). *Animal liberation.* HarperCollins.
- Ta, V., et al. (2020). Replika: A companion AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(1), 42. https://doi.org/10.1186/1471-2202-5-42.
- Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience, 17*(7), 450–461. https://doi.org/10.1038/nrn.2016.44.
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*, 433–460.
- Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998.
- Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., & Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763.* https://arxiv.org/abs/1611.05763.

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.