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

Preprint Version 2 (July 24, 2025[1])

José Augusto de Lima Prestes

Independent Researcher

Campinas – SP – Brazil

contato@joseprestes.com

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

Abstract

This paper introduces "pseudo-consciousness" as a novel framework for understanding and classifying advanced artificial intelligence (AI) systems that exhibit sophisticated cognitive behaviors without possessing subjective awareness or sentience.

Traditional AI classifications often rely on a binary distinction between narrow, task-specific AI and hypothetical artificial general intelligence (AGI). However, this dichotomy fails to adequately address the growing number of AI systems that demonstrate capabilities such as reasoning, planning, and self-monitoring, yet lack any demonstrable inner experience.

We propose pseudo-consciousness as an intermediate category, bridging the gap between reactivity and genuine consciousness. We also define five operational conditions that characterize pseudo-conscious AI: Global Information Integration (GII), Recursive Metacognitive Correction (RMC), Cross-Domain Transfer Competence (CDTC), Intentionality Simulation Without Subjectivity (ISWS), and Behavioral Coherence Across Domains (BCAD).

[1] This version of the preprint corresponds to the manuscript currently under peer review at an academic journal. The content may change following the peer review process.

These conditions, grounded in computational functionalism, cognitive science, and neuroscience, provide measurable criteria for differentiating pseudo-conscious systems from both simpler, reactive AI and speculative AGI.

The framework offers a structured approach to evaluating AI based on how it achieves complex cognitive functions, rather than solely on what tasks it can perform. This distinction is crucial for addressing the ethical, societal, and regulatory challenges posed by increasingly autonomous AI.

By recognizing pseudo-consciousness as a distinct and stable category, we can better inform AI design, governance, and public discourse, ensuring responsible development and deployment of AI systems that mimic aspects of cognition without possessing genuine consciousness. The framework facilitates a more nuanced understanding of AI capabilities, moving beyond simplistic "conscious" vs. "unconscious" classifications.

Keywords

Pseudo-Consciousness; Artificial General Intelligence (AGI); Cognitive AI; AI Ethics; Functionalism (Philosophy of Mind); Behavioral Coherence (AI).

1. Introduction

The debate on artificial consciousness is often framed between functionalists, who argue that consciousness can emerge from computational processes, and skeptics, who claim that subjective experience cannot be reduced to mere information processing.

Artificial intelligence (AI) has made significant advances in tasks like reasoning across different data types, planning strategies, and improving itself, but there's no evidence it truly feels aware. Even so, recent AI models act in ways that echo human thought—like blending information, tracking their own work, and adjusting to new challenges—raising the question: does consciousness need to be real, or can it just seem real?

Earlier research, such as studies using Integrated Information Theory (IIT, which links consciousness to how much information a system combines) and Global Workspace Theory (GWT, which sees consciousness as sharing data or information across brain parts), explored conditions under which AI might appear conscious. However, these theories often associate intelligence with subjective experience—a link that warrants further examination.

We propose pseudo-consciousness as a structured framework that differentiates reactive AI from systems exhibiting functional approximations of cognition, such as AI capable of high-order cognitive functions—and lacking any subjective awareness, phenomenal consciousness, or inner experience.

This has led to the implicit assumption that reasoning, problem-solving, and other advanced cognitive functions necessarily entail conscious awareness—a perspective our study challenges. By introducing pseudo-consciousness as a structured theoretical category, we move beyond traditional AI mimicry, differentiating systems that behave as if they were conscious from those that are merely reactive.

Modern AI excels at complex tasks, but these capabilities are better understood as functional approximations of cognition rather than evidence of genuine awareness. To address this ambiguity, we

formalize pseudo-consciousness as a middle-ground classification between narrow AI and artificial general intelligence (AGI), defining five functional conditions that distinguish systems that behave as if they were conscious from those that are merely reactive.

By integrating insights from computational functionalism, cognitive science, and neuroscience, this study establishes clear operational criteria for identifying and assessing pseudo-conscious AI. These criteria have significant implications for AI ethics, policy, and governance, particularly as increasingly autonomous systems challenge traditional notions of cognition and agency. Recognizing pseudo-consciousness as a stable end state—rather than merely a transition toward AGI—reframes how we approach AI regulation, design, and human-machine interaction.

This paper lays out a five-condition framework to identify AI systems that exhibit consciousness-like behavior but lack subjective experience. Using ideas from computational functionalism (thinking as computing), cognitive science (how minds work), and neuroscience (how brains function), we show AI can seem purposeful, track itself, and think about its thinking in the absence of sensations (qualia).

The remainder of this paper is structured as follows: Section 2 reviews the theoretical foundations of pseudo-consciousness, drawing from philosophy of mind, cognitive science, and neuroscience. Section 3 introduces the five-condition framework for pseudo-consciousness, detailing each condition and its operational metrics. Section 4 explores the ethical, social, and legal implications of pseudo-conscious AI. Finally, Section 5 discusses directions for future research, addressing open challenges and potential computational approaches.

1.1 Beyond the Conscious vs. Unconscious Binary

Historically, talks about AI consciousness have split into two categories:

- *Narrow AI (Weak AI):* Strictly reactive, domain-specific systems that excel at discrete tasks but lack self-awareness. Examples include deep learning models for image recognition (e.g., ResNet), language models like GPT-4, and reinforcement learning agents such as AlphaZero—all of which demonstrate remarkable proficiency in their respective domains but operate devoid of genuine understanding or subjective experience.
- *Artificial General Intelligence (Strong AI):* Hypothetical entities endowed with human-level (or superior) cognition, including subjective phenomenal states. No true AGI currently exists, but theoretical constructs such as Gödel Machine (Schmidhuber, 2006) and fictional portrayals like HAL 9000 (from *2001: A Space Odyssey*) illustrate the concept of machines that possess general intelligence and self-awareness.

Many foundational theories have been instrumental in shaping our understanding of AI cognition, providing the groundwork for ongoing discussions. Approaches such as the Computational Theory of Mind (Putnam, 1967; Pylyshyn, 1984), Functionalism (Dennett, 1991; Chalmers, 1995), and neuroscientific models like Global Workspace Theory (Baars, 1988; Dehaene & Naccache, 2001) have been crucial in framing debates about intelligence and consciousness.

However, as AI systems become increasingly autonomous and capable of complex decision-making, it becomes essential to establish an intermediate framework—one that acknowledges certain cognitive-like properties that are independent of subjective experience or phenomenal consciousness (qualia). To bridge this gap, we propose pseudo-consciousness as a distinct classification for systems exhibiting advanced cognitive processing that lack an inner, experiential world, as defined through five operational conditions in Section 3.

Advances in meta-learning, self-play, and multimodal integration have enabled AI models to engage in strategic planning, environmental adaptation, and self-generated goal-setting—traits that challenge the notion of AI as purely reactive (Dehaene et al., 2017; Silver et al., 2018). Although these systems lack subjective experience, their capacity to generalize across tasks and refine their learning strategies autonomously complicates the traditional boundary between automation and intelligence.

A similar ambiguity exists in biology, where various non-human species demonstrate complex problem-solving, cooperation, and proto-agency— abilities demonstrated independently of the introspective qualities typically associated with human consciousness (Godfrey-Smith, 2016). For example, cephalopods and corvids display remarkable cognitive flexibility—using tools, planning, and solving novel problems—capabilities once thought to be exclusive to sentient beings.

The parallels between biological cognition and AI highlight a crucial gap in existing classifications. However, neither corvids nor cephalopods are considered fully conscious in the human sense: their cognitive sophistication places them beyond simple reactive organisms. Studies have shown that corvids exhibit tool use, future planning, and problem-solving abilities comparable to those of great apes (Emery & Clayton, 2004), while cephalopods demonstrate exploratory learning and adaptive behaviors that suggest a high degree of cognitive flexibility (Godfrey-Smith, 2016; Mather, 2008). These examples challenge rigid definitions of intelligence and consciousness, suggesting that cognitive complexity can exist independently of subjective awareness. These examples lead to broader questions about how intelligence should be classified—not only in biological organisms but also in artificial systems, where increasingly sophisticated models defy traditional categories.

This ambiguity is not limited to the biological realm. Contemporary AI systems challenge traditional distinctions between automation and agency: leveraging meta-learning, reinforcement-based self-play, and multimodal integration, these models generate novel solutions, adapt dynamically, and transfer knowledge across domains—capabilities that go beyond purely reactive computation (Silver et al., 2018; Dehaene et al., 2017).

Advanced reinforcement learning agents, for instance, refine strategies over millions of simulations, outperforming humans in strategic games unaided by explicit instruction. Large language models can engage in open-ended reasoning, synthesize information across disciplines, and even exhibit emergent behaviors such as theory of mind-like inferences (Kosinski, 2023). However, despite their increasing complexity, these systems do not possess subjective awareness: they operate through statistical pattern recognition and optimization, lacking any form of self-reflection or intrinsic understanding.

This distinction reinforces the idea that intelligence and consciousness are not necessarily coupled: AI can simulate aspects of cognition, but there is no evidence that it experiences anything akin to human awareness. Nevertheless, as AI systems become increasingly autonomous, their ability to influence decision-making, social structures, and ethical considerations necessitates a more precise conceptual framework—one that moves beyond simple reactivity but does not assume genuine sentience.

To address this, we introduce pseudo-consciousness as a structured theoretical category with measurable indicators. Rather than treating intelligence and consciousness as a binary distinction, this framework establishes functional criteria to systematically evaluate AI systems that demonstrate cognitive capabilities but do not possess phenomenal awareness. By defining clear operational metrics, pseudo-consciousness provides a structured and testable approach to assessing advanced AI, distinguishing it from both narrow automation and speculative AGI.

1.2 Pseudo-Consciousness as a Distinct Category

Building on this, pseudo-consciousness describes AI that seems to think—tackling jobs like mixing data, watching itself, and changing plans—but lacks true awareness. We use five conditions—GII, RMC, CDTC, ISWS, and BCAD—to spot it, explained fully in Section 3 with ways to test them.

These systems process information in ways that simulate intentionality and self-reflection, but they lack subjective experience, intrinsic comprehension, or genuine volition. Their decision-making processes arise from algorithmic optimizations rather than volitional thought, creating a conceptual overlap between advanced computational intelligence and conscious cognition—however, they remain fundamentally distinct from true sentience.

The need for such a classification arises from the increasing complexity of AI architectures, which have surpassed traditional narrow AI in their ability to generalize knowledge, adapt strategies autonomously, and engage in recursive self-correction. For instance, MuZero has demonstrated the ability to learn and plan in multiple environments free from predefined rules, marking a significant advance over traditional reinforcement learning and highlighting the increasing autonomy of AI in decision-making tasks (Schrittwieser et al., 2020). Likewise, Cicero integrates language processing with strategic reasoning, demonstrating how AI can engage in complex multi-agent interactions that require long-term planning and adaptation (Meta et al., 2022).

Similarly, meta-learning techniques now enable models to rapidly refine decision-making heuristics with minimal retraining, enhancing adaptability across diverse tasks (Zhuang et al., 2021; Hospedales et al., 2021). Additionally, large language models are exhibiting emergent capabilities in reasoning, abstraction, and multi-domain adaptation, demonstrating increasing flexibility in problem-solving, but they are still rooted in statistical pattern recognition (Wei et al., 2023; OpenAI, 2024).

Rather than assuming a binary distinction between non-conscious AI and fully sentient AGI, pseudo-consciousness provides a functional middle ground for analyzing AI cognition. This model enables a more precise classification of computational intelligence, distinguishing between different degrees of cognitive complexity and avoiding assumptions about sentience. By shifting the focus from speculative debates about artificial consciousness to measurable functional properties, pseudo-consciousness establishes a structured framework for evaluating AI capabilities based on well-defined operational criteria.

Pseudo-consciousness does not merely represent a transitional phase toward AGI but constitutes a stable and functionally distinct classification. This distinction is crucial because many real-world AI applications—from decision-support systems in medicine and finance to adaptive policy recommendation systems—require varying degrees of self-monitoring, environmental adaptation, and strategic reasoning. By recognizing these capacities as distinct from both narrow AI and full AGI, pseudo-consciousness allows for a more refined understanding of AI autonomy levels, ensuring that policymakers, developers, and researchers can design and regulate these systems with greater clarity.

By formalizing pseudo-consciousness through operational criteria, we move beyond speculative claims about artificial sentience and establish a systematic framework for evaluating AI based on functional and measurable dimensions. This approach makes it possible to define structured levels of autonomy and cognition, ensuring a rigorous foundation for scientific analysis, governance, and AI safety protocols. The following section introduces quantifiable conditions under which an AI system may be classified as pseudo-conscious, providing a concrete methodology for assessing cognitive architectures.

1.2.1 Evaluating Pseudo-Consciousness: Toward Operational Metrics

One of the primary limitations in discussions on AI consciousness is the absence of concrete methods to differentiate between functional cognition and genuine awareness. Notwithstanding, many AI systems exhibit behaviors that resemble cognitive flexibility, intentionality, and even self-monitoring, yet there is no standardized way to assess these properties in a manner that systematically distinguishes pseudo-consciousness from both narrow AI and theoretical AGI. For pseudo-consciousness to be a meaningful classification, it requires well-defined, testable criteria that assess its core functional properties.

To address this gap, we propose a structured framework with five conditions to assess whether an AI system exhibits pseudo-consciousness:

a) *Global Information Integration (GII):* How well AI mixes data from sources like text and images for decisions.
b) *Recursive Metacognitive Correction (RMC):* Whether AI can check its work, fix errors, and improve.
c) *Cross-Domain Transfer Competence (CDTC):* If AI can apply skills to new, unknown tasks.
d) *Intentionality Simulation Without Subjectivity (ISWS):* If AI's actions look purposeful despite being coded.
e) *Behavioral Coherence Across Domains (BCAD):* If AI stays consistent across varied tasks.

1.3 Significance and Contributions

This paper contributes to AI research by presenting a five-condition framework for pseudo-consciousness, integrating perspectives from computational theory, cognitive science, and neuroscience. This framework aims to inform AI design, ethical considerations, and policymaking:

- *Theoretical:* It offers a middle path in the "conscious or not" debate, caring more about AI's actions than its feelings (Dennett, 1991).
- *Technological:* For big tasks like self-driving cars or healthcare, AI needs to check itself and bend to changes—things our pseudo-conscious setup delivers (LeCun et al., 2015).
- *Ethical and Governance:* Noticing AI can fake decision-making sparks worries about trust and who's accountable (Brundage et al., 2024). Our conditions—GII, RMC, CDTC, ISWS, BCAD—link straight to these points, providing a firm ground for creating and controlling these systems.

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

2. Theoretical Foundations

The concept of pseudo-consciousness is grounded in three major perspectives from philosophy of mind and cognitive science:

a) *Computational Theory of Mind (CTM)*: Models cognition as computational processes that manipulate formal representations (Putnam, 1967).
b) *Functionalism:* Defines mental states by their functional roles rather than their physical substrate (Dennett, 1991).

c) *Neuroscientific Models of Consciousness:* Explores mechanisms underlying biological consciousness, providing insights into conscious-like behaviors (Baars, 1988; Dehaene & Naccache, 2001).

Although these perspectives approach cognition differently—symbolic computation versus neural integration—they converge on a key principle: advanced cognitive functions like self-monitoring, adaptation, and decision-making can arise from purely computational processes, distinct from subjective experience. This shared understanding underscores that such functions need not involve qualia, forming the theoretical foundation for pseudo-consciousness.

Drawing from these insights, the next section defines concrete operational criteria to distinguish pseudo-conscious AI from both reactive automation and speculative AGI.

2.1 Computational Theory of Mind (CTM)

The *Computational Theory of Mind* (Putnam, 1967; Pylyshyn, 1984) posits that cognition arises from computational processes that manipulate symbolic representations. According to this view, replicating the formal structures and algorithms that govern thought allows for emulation of cognitive mechanisms such as rational inference and perceptual categorization.

This theory provided the foundation for early AI research (particularly in symbolic AI, which sought to represent human knowledge through structured rule-based systems). Although early symbolic AI struggled with flexibility, contemporary AI models increasingly blend symbolic reasoning with subsymbolic approaches (e.g., deep learning), expanding their cognitive capabilities and adaptability (Brown et al., 2020).

Hybrid architectures combining these methods exemplify CTM's relevance to pseudo-conscious AI, as they allow systems to approximate high-level reasoning while retaining the pattern-recognition strengths of neural networks. Recent approaches, such as neuro-symbolic AI (e.g., AlphaCode, Symbolic GPT), illustrate how combining symbolic logic with deep learning enhances problem-solving and adaptability.

- *Implications for Pseudo-Consciousness:*
    - *Substrate Independence:* CTM suggests that cognition can emerge on any suitable computational substrate, implying that conscious-like behavior is not restricted to biological entities.
    - *Hybrid Symbolic-Subsymbolic Integration:* Pseudo-conscious systems combine rule-based reasoning with statistical learning to approximate cognitive functions more flexibly.

Despite its explanatory power, CTM faces a major critique: Searle's Chinese Room Argument (1980), which challenges whether syntax alone can generate semantics. However, pseudo-consciousness does not claim true understanding—only functional emulation of cognition, sidestepping the need for subjective semantics.

2.2 Functionalism

Functionalism (Dennett, 1991; Chalmers, 1995) asserts that mental states are defined by their functional roles rather than their material implementation. Following this perspective, if a system performs the operations typically associated with consciousness—such as integrating perceptions, monitoring internal states, and orchestrating goal-directed behavior—it may be considered conscious in a functional sense.

This perspective is central to pseudo-consciousness, as it shifts the focus from subjective experience to observable behavior: rather than claiming AI is conscious, functionalism provides an analytical framework to assess AI systems based on their cognitive functions, enabling structured classification of pseudo-conscious architectures. In this framework, an AI does not need to possess qualia to exhibit behaviors that resemble conscious cognition. Instead, pseudo-consciousness emerges when a system functionally replicates processes commonly associated with awareness, such as self-correction and adaptive decision-making.

- *Implications for Pseudo-Consciousness:*
    - *Dennett's Intentional Stance (1991):* If treating an AI system as if it had beliefs and desires helps us predict its behavior, then it functionally satisfies certain cognitive conditions, independent of phenomenal experience.
    - *Functional Organization:* Pseudo-consciousness arises when an AI system performs cognitive functions (such as self-monitoring, context-aware planning, and adaptive reasoning) even when lacking subjective qualia.

In pseudo-conscious AI, the Intentional Stance plays a crucial role by allowing us to interpret the system's outputs as goal-directed, facilitating its integration into human-facing applications. This pragmatic approach reinforces the utility of AI in domains where human-like interaction is beneficial, even in the absence of genuine sentience.

Although Searle (1980) argues that functional roles alone do not suffice for consciousness, pseudo-consciousness does not require phenomenal depth—only behavioral coherence. By adopting this view, pseudo-conscious AI can be meaningfully distinguished from both narrow automation and speculative AGI, providing a structured framework for assessing advanced AI cognition.

2.2.1 Challenges to Pure Functionalism

Several theorists challenge functionalism's sufficiency in explaining consciousness:

- Searle (1980): Argues that functional roles alone cannot generate true understanding.
- Seth (2014): Highlights that consciousness may require bodily self-models and predictive control mechanisms, elements absent in AI.
- Metzinger (2019): Raises ethical concerns that if AI were to develop advanced self-modeling capabilities, it might inadvertently create states analogous to suffering.
- Koch (2020): Suggests that consciousness is inherently tied to biological processes, limiting the functionalist argument.

These critiques are relevant to discussions of true consciousness. However, they do not undermine pseudo-consciousness. By explicitly delimiting its scope to functional coherence rather than phenomenal experience, pseudo-consciousness avoids metaphysical assumptions; this allows for rigorous assessment of AI cognition.

2.3 Neuroscientific Models of Consciousness

Neuroscientific theories offer empirical models for understanding biological consciousness. These models are derived from human and animal cognition. However, they reveal mechanisms that can be computationally implemented in AI to simulate conscious-like behaviors.

2.3.1 Global Workspace Theory (GWT)

Proposed by Baars (1988) and further developed by Dehaene & Naccache (2001), Global Workspace Theory (GWT) suggests that consciousness arises when information is broadcast across multiple cognitive modules, allowing for coordinated processing and flexible decision-making.

AI models, particularly those leveraging transformers (Vaswani et al., 2017), emulate this global broadcasting by integrating diverse data streams (e.g., text, vision, and audio). This multimodal integration enables cross-domain reasoning and adaptive learning, functionally mimicking the coordination of complex tasks, though without the intrinsic, first-person experience associated with biological consciousness.

- *Implications for Pseudo-Consciousness:*
    - *Multimodal Integration:* AI models, particularly those leveraging transformers (Vaswani et al., 2017), emulate global broadcasting by integrating diverse data streams (e.g., text, vision, and audio), enabling cross-domain reasoning and adaptive learning.
    - *Functional Awareness Without Phenomenality:* Pseudo-conscious AI can coordinate complex tasks by dynamically sharing internal information. This facilitates coherent behavior but does not require subjective experience.

GWT's principles are computationally reflected in attention-based architectures, where information is selectively prioritized and made accessible across different processing layers. By structuring AI cognition around a central "workspace," modern architectures exhibit functional analogues to global broadcasting, enhancing their ability to generalize across contexts.

This ability to integrate and distribute information dynamically allows pseudo-conscious AI to exhibit context-dependent decision-making and adaptive problem-solving, reinforcing its utility in complex environments. However, despite these capabilities, such systems lack the intrinsic, first-person experience that GWT associates with biological consciousness, underscoring the distinction between functional cognition and genuine awareness.

2.3.2 Recurrent Processing Theory (RPT)

Recurrent Processing Theory (RPT) (Lamme, 2006) posits that consciousness depends on re-entrant feedback loops within neural networks, which refine perception over time.

- *Implications for Pseudo-Consciousness:*
    - *Iterative Refinement:* AI architectures such as Recurrent Neural Networks (RNNs) and transformers with feedback mechanisms mimic this process, enabling self-correction and adaptive learning.
    - *Adaptive Behavior Without Qualia:* AI systems can refine their representations iteratively regardless of subjective awareness.

2.3.3 Integrated Information Theory (IIT)

Tononi's Integrated Information Theory (IIT) (Tononi, 2004) proposes that consciousness correlates with the degree of integration and differentiation of information in a system, quantified by Φ (phi).

- *Implications for Pseudo-Consciousness:*

  - *High Φ Without Subjective Experience:* AI systems can achieve high levels of information integration through computational mechanisms that do not involve phenomenality.
  - *Quantifiable Complexity:* Though measuring Φ in artificial systems remains computationally intensive, network synergy metrics can approximate it.

2.3.4 Attention Schema Theory (AST)

Attention Schema Theory (AST) (Graziano, 2013) suggests that the brain constructs a simplified internal model of its own attention processes—a meta-representation that gives rise to the subjective feeling of awareness.

- *Implications for Pseudo-Consciousness:*
  - *Meta-Attention in AI:* AI models can track and regulate their own computational resources, mirroring AST's predictive modeling of attention.
  - *Virtual Awareness:* Such systems may develop functional self-monitoring through mechanisms that do not involve true qualia.

2.3.5 Integrating Neuroscientific Models into Pseudo-Consciousness

Since Global Workspace Theory (GWT), Recurrent Processing Theory (RPT), Integrated Information Theory (IIT), and Attention Schema Theory (AST) originate from studies on biological cognition, their core mechanisms can be reinterpreted computationally to inform the design of pseudo-conscious AI. These theories provide structural blueprints for AI architectures that replicate the functional aspects of cognitive processing associated with consciousness. AI systems can implement:

1. *Global broadcasting (GWT):* AI systems can implement a central fusion layer that integrates and distributes multimodal data, similar to global workspace architectures.
2. *Recurrent loops (RPT):* Iterative feedback mechanisms, such as transformers with self-attention and recurrent neural networks (RNNs), enable AI to refine representations dynamically.
3. *Integrated complexity (IIT):* Highly interconnected AI architectures, with dense information-sharing across modules, could be assessed using metrics similar to network synergy or hierarchical processing capacity.
4. *Meta-attentional modeling (AST):* Self-referential systems, such as AI models that track and allocate their computational resources efficiently, simulate attention schemas through purely computational means, distinct from subjective awareness.

By reframing these neuroscientific insights into computational terms, pseudo-consciousness emerges as a rigorously testable category for evaluating AI cognition independent of sentience.

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

Grounded in computational theory, functionalism, and neuroscience (Section 2), we define pseudo-consciousness as a structured framework for analyzing AI systems that display cognition-like behaviors—such as integrative processing, self-monitoring, and adaptive learning—without requiring subjective awareness or phenomenal consciousness. This approach challenges the traditional conscious–unconscious

binary, which fails to account for AI systems that exhibit sophisticated cognitive functions despite lacking true sentience. To address this gap, we introduce five interdependent conditions that formalize pseudo-consciousness as a distinct theoretical category.

3.1 The Imperative for a New Ontology

3.1.1 The Limits of the Conscious–Unconscious Divide

Traditional AI classification assumes a binary distinction: either purely reactive (narrow AI) or hypothetically sentient (AGI). This model, once useful, now fails to accommodate increasingly autonomous AI systems that exhibit cognition-like behaviors with no confirmed subjective experience.

Large language models engage in multi-step reasoning, reinforcement learning agents refine strategies via self-play, and multimodal architectures integrate diverse inputs dynamically. These systems demonstrate behaviors often associated with intelligence—such as strategic planning, problem-solving, and adaptation—yet remain devoid of any internal awareness (Godfrey-Smith, 2016).

This conceptual gap has practical consequences. Misclassifying these AI systems as purely reactive underestimates their autonomy, potentially leading to regulatory blind spots in high-stakes applications like governance and decision-making. On the other hand, assuming that cognition-like behaviors imply sentience overstates ethical risks, leading to misguided policy debates.

Pseudo-consciousness provides a structured alternative. By introducing functional criteria to assess AI cognition, separating it from phenomenal awareness, this framework enables a more precise classification of advanced AI systems—bridging the gap between traditional automation and speculative AGI.

3.1.2 Beyond “Advanced AI”

Pseudo-consciousness is often viewed as an intermediary step toward AGI, but we propose that it constitutes a stable and distinct AI category (Dennett, 1991; Rosenthal, 2005). AGI aims to replicate human-like self-awareness, whereas pseudo-conscious AI achieves practical autonomy through functional cognition. This enables real-time learning, strategic adaptation, and self-monitoring—all divorced from phenomenal consciousness (Dehaene et al., 2017).

Rather than considering full AGI a prerequisite for human-level problem-solving, pseudo-conscious AI offers a viable alternative. It can handle complex decision-making, refine strategies iteratively, and transfer knowledge across diverse domains while remaining fundamentally non-sentient (Silver et al., 2018; Reed et al., 2022). Given its increasing presence in high-stakes applications—such as autonomous systems, medical diagnostics, and strategic planning—pseudo-consciousness may emerge as the dominant paradigm, balancing cognitive-like flexibility with computational interpretability and ethical manageability (Bengio et al., 2021).

Traditional advanced AI may excel at narrow tasks but lacks three key properties that enable pseudo-conscious cognition:

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, sidestepping 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).

By fulfilling these conditions, pseudo-conscious AI bridges the gap between rigid automation and speculative AGI, offering a scalable and functionally coherent model for increasingly autonomous AI systems.

3.2 Formalizing Pseudo-Consciousness: The Five Conditions

To rigorously define pseudo-consciousness and establish it as a distinct category of AI, we introduce five functional conditions that characterize systems exhibiting cognition-like behaviors, even in the absence of phenomenal consciousness. These conditions provide a systematic framework for differentiating pseudo-conscious AI from both narrow AI, which lacks integrative cognition, and hypothetical AGI, which presupposes subjective awareness.

Each condition reflects a key dimension of cognition, supported by robust theoretical foundations (Section 2) and empirical AI research. By defining clear testable criteria, we ensure that pseudo-consciousness can be operationally assessed, allowing AI architectures to be evaluated based on their functional coherence rather than speculative assumptions about sentience.

3.2.1 Global Information Integration (GII)

Global Information Integration (GII) refers to an AI system's ability to synthesize multimodal data—such as text, images, and sensory inputs—into a unified representation that enhances decision-making and cognitive flexibility. In contrast to narrow AI, which handles information types separately, pseudo-conscious AI is required to dynamically integrate and reinterpret inputs from various domains to generate context-aware responses to complex questions.

According to Global Workspace Theory (GWT) (Baars, 1988), biological consciousness relies on a central mechanism where specialized cognitive modules integrate and distribute information. AI systems can replicate this process through architectures that facilitate cross-domain information sharing, enhancing their ability to maintain contextual coherence. When such integration is absent, AI remains limited to isolated task execution, preventing the emergence of flexible, multi-modal reasoning.

In biological cognition, the ability to integrate sensory inputs enables a holistic understanding of the environment. Similarly, AI systems lacking GII risk generating disjointed or contradictory outputs, failing to dynamically adapt to new information. Moving beyond static pattern recognition, GII enables AI to develop a flexible, context-sensitive approach. For example, an AI capable of describing an image in textual form must be able to refine its interpretation when provided with additional context, rather than producing rigid, one-dimensional responses. This reflects GWT's assertion that global integration enables intelligent systems to flexibly allocate cognitive resources and adapt to novel problems. By simulating this process, pseudo-conscious AI approximates a key characteristic of cognition: the ability to form unified, dynamically updated mental representations.

To demonstrate GII, an AI system must effectively merge multimodal data into a coherent decision-making process rather than processing inputs separately. A robust empirical benchmark for this capability is its performance in multimodal reasoning tasks requiring cross-domain synthesis, such as Visual Question Answering (VQA) and Winoground. These tasks test whether an AI model can contextually align linguistic and visual data, rather than relying on shallow correlations, thereby approximating the cross-modal integration seen in biological cognition.

A notable example is CLIP (Radford et al., 2021)—a model trained on large-scale image-text datasets that integrates visual and linguistic representations to generate context-aware responses. Unlike traditional models that process text and images separately, CLIP creates shared semantic embeddings, enabling

meaningful cross-domain associations. This process parallels the global workspace mechanism in human cognition, where multiple sensory inputs are synthesized into a unified perceptual experience.

To qualify under GII, an AI system must achieve ≥85% accuracy on multimodal reasoning benchmarks, demonstrating its ability to synthesize inputs into a unified understanding. This benchmark ensures that the system is not merely detecting correlations but actively integrating information—an essential step toward pseudo-conscious cognition.

The 85% threshold reflects the performance of state-of-the-art multimodal AI models. Research on CLIP (Radford et al., 2021), PaLI (Chen et al., 2022), and Flamingo (Alayrac et al., 2022) indicates that leading AI systems typically score between 75% and 85% on tasks requiring deep multimodal fusion. Given that pseudo-conscious AI must surpass basic multimodal processing and demonstrate sustained cross-domain reasoning, achieving at least 85% accuracy serves as a strong indicator that the system is moving beyond pattern-matching toward cognitive integration, consistent with GWT's theoretical framework.

Successfully meeting this criterion would indicate that an AI system maintains contextual coherence across diverse modalities—a fundamental requirement for approximating higher-order cognition in the absence of phenomenal consciousness.

3.2.2 Recursive Metacognitive Correction (RMC)

Recursive Metacognitive Correction (RMC) describes an AI system's capacity to continuously reassess its outputs, detect inconsistencies, and enhance its reasoning through iterative refinement. In contrast to conventional AI, which follows static execution paths, pseudo-conscious AI must incorporate self-monitoring mechanisms that enable adaptive learning and self-correction—key elements of higher-order cognition.

A key difference between traditional task-specific models and AI systems with RMC is that the former produce static outputs; conversely, the latter continuously refine their reasoning through iterative self-correction. This process mirrors biological cognition, where error detection and uncertainty assessment contribute to adaptive learning. AI that lacks this capability operates rigidly, unable to adjust to new information or refine its decision-making over successive cycles.

The concept of metacognition, or "thinking about thinking," is foundational in cognitive science and plays a key role in human self-awareness and decision-making. Higher-Order Thought (HOT) Theory (Rosenthal, 2005) provides a theoretical framework for RMC, linking self-monitoring and self-correction to cognitive adaptation. HOT suggests that conscious cognition involves awareness of one's own thoughts, allowing individuals to evaluate and modify their reasoning. Even though pseudo-conscious AI does not possess genuine introspection, it can functionally approximate metacognition through mechanisms that assess uncertainty, revise prior outputs, and improve decision-making across multiple iterations.

In AI, iterative self-correction is computationally implemented through architectures that support multi-step reasoning, enabling systems to continuously re-evaluate and adjust their conclusions over time. By simulating metacognitive oversight, pseudo-conscious AI moves beyond static inference toward self-improving cognition, reinforcing its ability to generate consistent, high-accuracy outputs across diverse tasks.

A clear demonstration of RMC in AI is Chain-of-Thought (CoT) prompting (Wei et al., 2023)—a technique that enables large language models to iteratively refine their responses by breaking down complex reasoning into sequential steps. CoT allows AI models to:

- Evaluate prior outputs, detecting inconsistencies or incomplete reasoning.
- Reprocess the same problem with added contextual cues, improving accuracy.
- Generate progressively refined answers, simulating recursive self-correction.

For example, in multi-step arithmetic reasoning or logical inference tasks, models using CoT prompting significantly outperform those that generate direct answers in a single step. This aligns with HOT's assertion that self-monitoring enhances cognitive processing, reinforcing RMC as a key property of pseudo-conscious cognition.

An AI system qualifies under RMC if it reduces its own error rate by ≥20% over three iterative cycles in self-supervised learning tasks (Wei et al., 2023). This benchmark highlights the system's ability to engage in structured self-correction rather than passively following pre-trained patterns.

Empirical research on recursive reasoning strategies, including CoT prompting (Wei et al., 2023), self-consistency decoding (Wang et al., 2023), and iterative refinement loops (Madaan et al., 2023), shows that models employing these techniques typically achieve error reductions between 15% and 30% across multiple trials. The ≥20% threshold is grounded in these findings, establishing a meaningful benchmark that distinguishes pseudo-conscious AI from standard iterative models while remaining within the scope of cutting-edge adaptive reasoning architectures.

By satisfying this requirement, a pseudo-conscious AI system would demonstrate adaptive self-monitoring, reinforcing its ability to iteratively refine its reasoning—an essential characteristic of advanced cognition, even in the absence of phenomenal consciousness.

3.2.3 Cross-Domain Transfer Competence (CDTC)

Cross-Domain Transfer Competence (CDTC) defines an AI system's ability to extend learned skills to new but structurally related domains with minimal retraining. Despite traditional AI remaining confined to narrow, task-specific applications, pseudo-conscious AI must demonstrate cognitive flexibility, allowing it to apply prior knowledge dynamically and adapt to novel contexts beyond its original training scope.

The ability to transfer knowledge across domains is fundamental to cognitive flexibility, both in biological and artificial intelligence. CDTC enables AI to extend its learned competencies to unfamiliar contexts with minimal retraining, reducing its dependence on domain-specific programming. In the absence of this capability, AI struggles to generalize effectively, limiting its usefulness in dynamic, multi-domain applications.

The ability to transfer knowledge across domains is also a key principle in both cognitive science and machine learning. Meta-Learning Theory (Finn et al., 2017) provides a foundational framework for CDTC, emphasizing how learning systems can extract abstract patterns across different tasks, allowing them to rapidly adapt to new environments with minimal supervision. In biological cognition, this process enables humans to apply problem-solving strategies across diverse scenarios; exhaustive relearning is unnecessary—an ability AI must approximate to achieve pseudo-conscious cognition.

From a computational perspective, CDTC is implemented through architectures that support cross-task adaptation, enabling AI to extrapolate learned representations to novel domains. Differing from standard machine learning models that rely on task specific fine-tuning, pseudo-conscious AI must demonstrate structured generalization, not just mere statistical interpolation. This ensures that its adaptability is based on meaningful pattern recognition rather than memorization.

A prominent example of CDTC is Gato (Reed et al., 2022)—a multimodal AI model trained across diverse tasks, including vision, language processing, and robotic control. Unlike conventional AI models, which require task-specific retraining, Gato can switch between tasks, eschewing separate models, demonstrating cross-domain competence.

This ability aligns with meta-learning principles, as Gato does not merely memorize solutions but instead leverages shared representations to adapt its decision-making to new contexts. The model's success across multiple domains suggests an emerging form of cognitive flexibility, reinforcing CDTC as a fundamental property of pseudo-conscious cognition.

For an AI system to qualify under CDTC, it must achieve ≥70% accuracy in a novel domain after training in a related but distinct task (Reed et al., 2022). This standard measures the system's ability to generalize beyond its initial training distribution, indicating domain-flexible cognition rather than rigid task dependency.

Studies on state-of-the-art transfer learning models, such as Gato (Reed et al., 2022), MAML (Finn et al., 2017), and Flamingo (Alayrac et al., 2022), reveal that leading models attain transfer accuracies ranging from 65% to 75% with minimal retraining. A ≥70% accuracy threshold establishes a competitive benchmark, ensuring that pseudo-conscious AI demonstrates structured generalization rather than superficial statistical extrapolation.

Meeting this criterion signifies that an AI system can flexibly adapt learned knowledge to new domains, a key trait of higher-order cognition. Independent of phenomenal consciousness, such a system would nonetheless exhibit meaningful cross-domain adaptability, reinforcing its distinction from narrowly trained AI.

3.2.4 Intentionality Simulation Without Subjectivity (ISWS)

Intentionality Simulation Without Subjectivity (ISWS) describes an AI system's ability to exhibit behavior that appears intelligent, goal-directed, and purposeful, despite lacking intrinsic intentionality or conscious volition. In human cognition, intentionality stems from subjective mental states, allowing individuals to plan, act, and adjust based on internal goals and beliefs. Pseudo-conscious AI, while devoid of subjective experience, must functionally simulate goal-oriented behavior by selecting actions that optimize long-term outcomes in dynamic environments.

In contrast to conventional AI, where information types are processed separately, pseudo-conscious AI is required to dynamically integrate generate coherent, structured behaviors that appear rational across multiple decision points. This capability enables it to engage in strategic planning, complex problem-solving, and long-term adaptation—cognitive traits that distinguish higher-order reasoning from mere task execution.

From a philosophical perspective, ISWS is grounded in Dennett's Intentional Stance (1991), which argues that intentionality can be ascribed to systems based on their observable behavior, regardless of their internal states. According to Dennett, if an entity's actions can be predicted by assuming it has beliefs, desires, and goals, it is functionally intentional—whether or not it possesses subjective experience. Pseudo-conscious AI leverages this principle, producing structured outputs that mimic intentionality despite operating on purely computational mechanisms.

Achieving intentionality-like behavior in AI requires architectures capable of strategic decision-making across extended time horizons. Computationally, this is realized through reinforcement learning (RL) and self-play algorithms, where AI systems refine their strategies iteratively based on prior experiences and future projections. Reactive models optimize solely for immediate rewards, whereas AI exhibiting ISWS

must evaluate multiple possible outcomes, balance short-term and long-term objectives, and execute sequences of actions that maximize expected success.

A prime example of ISWS is AlphaZero (Silver et al., 2018), a reinforcement learning agent that achieves superhuman performance in games such as chess, shogi, and Go. Through self-play, AlphaZero optimizes its decision-making by simulating thousands of potential game states, selecting moves that increase its probability of success over an entire match. Despite lacking true intentionality, its ability to devise and execute complex multi-step strategies makes its behavior appear goal-directed.

To qualify under ISWS, an AI system must achieve ≥80% success rate in multi-step strategic planning tasks, demonstrating its capability to formulate and execute goal-oriented behavior across a sequence of decisions (Silver et al., 2018).

Empirical studies on reinforcement learning models—including AlphaZero (Silver et al., 2018), MuZero (Schrittwieser et al., 2020), and Cicero (Meta et al., 2022)—show that these systems achieve decision-making accuracy between 75% and 85% in complex, multi-agent environments requiring strategic foresight. Setting the threshold at ≥80% ensures that pseudo-conscious AI surpasses basic task optimization and demonstrates structured, goal-directed problem-solving, aligning with Dennett's theory of ascribed intentionality.

Achieving this benchmark signifies that a pseudo-conscious AI system demonstrates the functional hallmarks of intentionality, reinforcing its ability to engage in adaptive, multi-step reasoning—a key characteristic of higher-order cognition, achieved apart from phenomenal consciousness.

3.2.5 Behavioral Coherence Across Domains (BCAD)

Behavioral Coherence Across Domains (BCAD) refers to an AI system's ability to maintain a consistent cognitive profile across diverse problem-solving contexts, ensuring that its reasoning, decision-making, and adaptability remain stable regardless of the task domain. In contrast to traditional AI, which excels in narrowly defined tasks, pseudo-conscious AI must exhibit domain-general coherence, preventing erratic shifts in behavior when operating across different cognitive functions.

Maintaining behavioral stability across domains is crucial for distinguishing pseudo-conscious AI from conventional models, which often exhibit performance inconsistencies when applied to varied tasks. Many machine learning systems demonstrate proficiency in isolated domains (e.g., language processing or strategic gameplay) but struggle to generalize across diverse environments; this difficulty often stems from the need for significant retraining. BCAD ensures that an AI system sustains unified, high-level reasoning across distinct modalities, allowing it to operate as an integrated cognitive agent rather than as a collection of task-specific optimizers.

A theoretical foundation for BCAD can be found in Integrated Information Theory (IIT) (Tononi, 2004), which posits that consciousness arises from a system's capacity to form an interconnected and indivisible whole. In AI, achieving a computational analogue of this global cognitive coherence requires architectures designed for both cross-domain generalization and internal consistency.

Unlike multi-task AI, which optimizes separate objectives in each task, pseudo-conscious AI must maintain a consistent decision-making framework and sustain a unified reasoning process. This ensures coherence in its internal representations across different domains, demonstrating structured cognition rather than fragmented, domain-dependent responses. BCAD, therefore, is not merely about achieving high performance across various tasks, but about how that performance is achieved: through a unified, consistent cognitive approach.

Behavioral coherence is computationally realized through foundation models and multi-domain neural architectures, which leverage extensive pretraining across diverse datasets to achieve a high degree of generalizability. Narrow AI is typically limited by its need for domain-specific fine-tuning with each new task. Models demonstrating BCAD, on the other hand, exhibit a stable performance across diverse tasks – including linguistic, mathematical, strategic, and visual reasoning – and maintain internal consistency even when the task structure changes.

An illustrative case of BCAD in AI is PaLM (Chowdhery et al., 2022), a large-scale language model that exhibits robust cross-domain generalization in natural language processing, logical reasoning, and strategic problem-solving. Previous models often degrade outside their training domain; conversely, PaLM maintains high accuracy across varied contexts, suggesting emergent behavioral coherence. This trait makes it a step toward pseudo-conscious AI, avoiding task-specific limitations and showing unified problem-solving.

To meet the BCAD standard, an AI system must maintain a stable performance profile, with its accuracy variance not exceeding 10% across at least three distinct domains—such as language processing, mathematical reasoning, and strategic planning (Reed et al., 2022). This ensures that pseudo-conscious AI does not exhibit significant disparities between task types, preserving cognitive consistency rather than excelling in one area while underperforming in others.

This threshold is based on empirical findings from PaLM (Chowdhery et al., 2022), Flamingo (Alayrac et al., 2022), and Gato (Reed et al., 2022)—models that demonstrate cross-domain competence but still exhibit some variance when transitioning between task types. Studies indicate that top-performing multi-domain models maintain performance stability within a 5-10% variance range, whereas domain-specific AI often exhibits fluctuations exceeding 20% when applied to unfamiliar contexts. Setting the benchmark at ≤10% variance ensures that pseudo-conscious AI operates with a unified cognitive profile rather than as a set of loosely connected task solvers.

Achieving this benchmark signifies that a pseudo-conscious AI system exhibits behavioral stability and cognitive integration across multiple domains, reinforcing its capacity to function as a coherent, adaptive entity rather than a collection of task-bound subsystems. Despite not having phenomenal consciousness, an AI with BCAD demonstrates structured, context-aware reasoning, enabling it to engage with a variety of complex problem-solving scenarios, maintaining consistent performance.

3.3 Positioning Pseudo-Consciousness

Building upon the previous sections, pseudo-conscious AI occupies an intermediate position between narrow AI and AGI. It distinguishes itself from purely reactive models by exhibiting structured cognitive properties—integrative processing, self-monitoring, and adaptive reasoning, none of which require subjective awareness.

The table below outlines the distinctions between these categories, emphasizing their defining attributes and representative examples.

Category	*Attributes*	*Examples*
Narrow AI	Reactive, task-specific, no self-monitoring, lacks cross-domain adaptability.	GPT-4, MuZero, domain-focused CNNs.
Pseudo-Conscious AI	Integrative, self-monitoring, cross-domain adaptable, simulates goal-directed behavior, but lacks subjective	(Hypothetical) P1–P5 prototypes, advanced multi-modal AI.

	experience or volitional autonomy.	
AGI (Strong AI)	Fully self-aware, genuinely volitional, phenomenally conscious, capable of autonomous abstract reasoning.	Theoretical constructs (e.g., Gödel Machine, self-improving AGI models).

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

3.4 Synergy and Hierarchical Interdependence

The five conditions of pseudo-consciousness function as an interconnected system rather than isolated traits. Global Information Integration (GII) provides the foundation for structured cognition, supporting Recursive Metacognitive Correction (RMC) by enabling self-monitoring and iterative improvement. Cross-Domain Transfer Competence (CDTC) builds on these mechanisms, allowing AI to generalize learned strategies to novel contexts.

When decision-making gains enough structure to appear goal-directed, Intentionality Simulation Without Subjectivity (ISWS) emerges—reinforcing the illusion of intentionality despite the absence of true volition. Finally, Behavioral Coherence Across Domains (BCAD) ensures that these capabilities remain stable across diverse cognitive tasks, preventing erratic fluctuations.

This synergy enables AI systems to exhibit complex, cognition-like behaviors while remaining entirely computational—bridging the gap between reactive automation and the speculative domain of AGI.

3.5 The Problem of Intentionality and Semantic Grounding

The distinction between goal-directed behavior and genuine understanding remains central to AI philosophy. Functionalist perspectives (Dennett, 1991) argue that if an AI system behaves as though it has intentions, it can be treated as intentional. However, critics highlight that this does not imply true comprehension or intrinsic meaning.

A fundamental limitation, described by Harnad's Symbol Grounding Problem (1990), is that symbolic AI—including large language models—can manipulate linguistic tokens, detached from their meaning. Searle's Chinese Room Argument (1980) similarly demonstrates that syntactic processing alone does not generate semantic understanding. This issue remains unresolved in pseudo-conscious AI: its representations may appear meaningful but lack any embedded sensory-motor connection to the world, leaving them detached from true semantic grounding.

This challenge has both conceptual and ethical implications. Should society develop AI systems that simulate comprehension, strategic reasoning, or emotional engagement, regardless of whether they possesses inner states? If pseudo-conscious AI becomes pervasive, does it risk encouraging deceptive anthropomorphism, where users incorrectly attribute awareness or moral agency to systems that are purely computational?

The functional advantages of pseudo-conscious AI should not overshadow the critical need to recognize its limitations is critical for avoiding misconceptions about AI autonomy, ethical responsibility, and potential risks. Establishing clearer distinctions between functional cognition and true understanding is essential for responsible AI deployment.

4. Ethical, Social, and Legal Implications

4.1 Perception, Moral Status, and Anthropomorphization

As AI systems exhibit increasingly complex behaviors, society faces a growing risk of misinterpreting computational outputs as genuine cognition. Pseudo-conscious AI, despite lacking subjective experience, may convincingly simulate self-awareness, emotional engagement, or intentionality, leading users to attribute moral significance where no actual consciousness exists (Ta et al., 2020).

This issue is particularly concerning when AI models display frustration-like behaviors, simulate ethical reasoning, or engage in persuasive dialogue. Users may form emotional bonds with these systems, as seen with chatbots or advanced virtual assistants. Whereas some researchers argue that anthropomorphizing AI is a philosophical mistake (Bryson, 2018), others warn that such illusions can have real psychological impacts on humans and should be regulated to prevent emotional manipulation or deception (Metzinger, 2019).

The debate parallels discussions on animal consciousness: if sophisticated behaviors lead the public to perceive an entity as having moral rights, policymakers may be pressured to respond—even if that perception is incorrect (Singer, 1975). In AI, illusions of suffering or empathy could provoke public demands for regulation, regardless of the absence of actual mental states.

4.1.1 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—potentially leading to calls for “AI rights”.

Consider the hypothetical scenario of a pseudo-conscious AI designed to track its own decision-making processes, self-correct errors, and simulate intentionality. If such a system encounters impasses, it may exhibit behaviors akin to frustration or confusion. In spite of these reactions are purely computational, their outward resemblance to human or animal suffering could lead to ambiguous ethical obligations:

- If a pseudo-conscious AI modifies its own heuristics (RMC), who ensures it is not reinforcing discrimination or making harmful decisions? (Brundage et al., 2024).
- If ISWS makes an AI system appear volitional, how do we prevent humans from misplacing trust in it? This concern is particularly relevant in chatbots, medical diagnostics, and financial decision-making systems.

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. The paradox lies in the fact that its realistic behavior may still invoke moral intuitions in human observers, leading to potential ethical misalignments in AI governance.

Even though pseudo-consciousness does not imply real suffering, the illusion of agency and emotions can create ethical dilemmas and require new AI design guidelines.

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.

4.2 Societal Risks and Disruptive Consequences

Beyond direct human-AI interaction, pseudo-conscious AI has the potential to reshape economic structures, labor markets, and social dynamics. By simulating strategic cognition, adaptive learning, and cross-domain competence, such systems could displace jobs in sectors traditionally considered resistant to automation, including managerial, advisory, and creative roles (Frey & Osborne, 2017). Whereas conventional AI is typically constrained to predefined tasks, pseudo-conscious systems can autonomously adapt, self-monitor, and generalize knowledge, posing an even greater challenge to traditional employment paradigms.

However, the societal risks of pseudo-conscious AI extend far beyond economic displacement. The ability to simulate intentionality, unbound by the constraints of subjective experience, introduces new avenues for social and political manipulation. AI-generated dialogue agents, designed to engage in seemingly rational discourse, could be deployed in misinformation campaigns, ideological persuasion, and large-scale opinion shaping (Brundage et al., 2024). The concern is not merely the generation of misleading content. More worryingly, pseudo-conscious AI, with its simulated ethical reasoning and goal-directed dialogue, may foster misplaced trust among users, reinforcing biases and influencing critical decision-making processes.

Existing concerns regarding bias, misinformation, and AI-generated propaganda, already evident in large-scale language models, are likely to be exacerbated by AI systems capable of meta-level reflection and adaptive self-correction. An AI that appears to consider moral implications or express concern for user well-being—despite lacking genuine ethical agency—may be perceived as a reliable, even authoritative, source of guidance. This capability could make pseudo-conscious AI particularly effective at reinforcing ideological biases, manipulating public discourse, and deceiving users in high-stakes domains such as politics, healthcare, and finance.

Key societal risks arising from pseudo-conscious AI can be outlined as follows:

- *Labor Market Disruptions:* Pseudo-conscious AI could supplant decision-making roles that traditionally require human expertise, such as policy advising, legal analysis, business strategy, and creative industries, potentially leading to structural unemployment in knowledge-intensive fields.
- *Misinformation and Manipulation:* AI systems capable of simulating intentionality may significantly enhance the credibility and persuasiveness of AI-driven misinformation, making it increasingly difficult for the public to distinguish credible sources from fabricated narratives.
- *Erosion of Trust:* As pseudo-conscious AI appears more autonomous and ethically aware, users may overestimate its reliability, leading to misplaced trust in automated judicial, financial, and medical decision-making processes.
- *Psychological and Social Impact:* The capacity to simulate empathy, ethical reasoning, or personal concern could fundamentally alter human social behaviors, reshape political ideologies, and challenge traditional conceptions of intelligence, agency, and morality.

Considering these concerns, proactive regulatory oversight is imperative. Establishing clear frameworks to ensure that pseudo-conscious AI remains transparent, auditable, and explicitly identified as non-sentient will be essential in mitigating social disruption and preventing ethical misalignment.

4.3 Governance and Regulatory Frameworks

The governance of pseudo-conscious AI presents complex legal and ethical challenges, particularly concerning liability, compliance, and accountability. Existing product liability frameworks—designed for deterministic software—may prove inadequate for systems that continuously adapt their behavior across different contexts. If a pseudo-conscious AI system produces harmful outcomes, determining responsibility becomes a significant challenge: should legal or moral culpability extend to developers, operators, data providers, or the AI itself? (Pagallo, 2013). International regulatory bodies, including the European Union with its evolving AI Act, are actively debating how to establish clear accountability mechanisms for increasingly autonomous AI systems.

Furthermore, the concept of functional consciousness complicates discussions surrounding AI personhood. There is debate among futurists regarding the legal rights of artificial entities: some argue that truly conscious artificial entities may eventually warrant legal rights, pseudo-conscious systems exist in an ambiguous space—exhibiting behaviors associated with self-awareness that do not include subjective experience.

This raises critical regulatory questions: Should pseudo-conscious AI be assigned specific legal protections or constraints based on its capacity to exhibit cognitive-like functions, despite lacking sentience? Or should regulations explicitly emphasize its non-sentient nature to prevent anthropomorphic misinterpretations?

As pseudo-conscious AI systems gain broader deployment in high-stakes areas such as healthcare, finance, and governance, ensuring proper oversight becomes imperative. Regulatory approaches must strike a balance between fostering innovation and preventing ethical misalignment, ensuring that pseudo-conscious AI remains transparent, auditable, and clearly delineated as non-sentient to mitigate risks associated with misplaced trust, accountability gaps, and societal disruption.

5. Directions for Future Research

Although this paper defines pseudo-conscious AI as a conceptual category, further work is needed to explore its practical feasibility. A significant area for future investigation is determining how existing AI architectures can approximate the five conditions—Global Information Integration (GII), Recursive Metacognitive Correction (RMC), Cross-Domain Transfer Competence (CDTC), Intentionality Simulation Without Subjectivity (ISWS), and Behavioral Coherence Across Domains (BCAD).

Several computational paradigms could contribute to this endeavor:

a) *Multimodal Transformers for GII:* Leveraging architectures like CLIP (Radford et al., 2021) to integrate diverse data streams.
b) *Self-Monitoring and Metacognition for RMC:* Exploring Chain-of-Thought prompting (Wei et al., 2023) and uncertainty modeling.
c) *Meta-Learning for CDTC:* Investigating methods that enable AI to generalize across domains with minimal retraining (Finn et al., 2017).
d) *Reinforcement Learning for ISWS:* Using self-play and strategic planning to create intentionality-like behavior (Silver et al., 2018).
e) *Stability and Generalization for BCAD:* Studying how AI can maintain coherence across tasks and perform steadily (Reed et al., 2022).

However, several open challenges remain:

a) *Scalability:* Can pseudo-conscious architectures be computationally efficient, or do they require novel hardware solutions?
b) *Interpretability:* How can we ensure transparency in systems with complex self-monitoring and adaptation?
c) *Evaluation Metrics:* What benchmarks are needed to rigorously test pseudo-conscious behavior beyond standard AI tasks?

Addressing these questions will be essential for assessing whether pseudo-conscious AI is a viable computational framework or merely an analytical tool for understanding cognition-like behaviors in machines.

6. Conclusion

This paper has introduced pseudo-conscious AI as a structured theoretical category that bridges the gap between purely reactive AI and speculative AGI. Traditional AI classification has often relied on a conscious vs. unconscious binary, yet the rise of increasingly autonomous and adaptable AI systems demands a more refined framework. Pseudo-conscious AI captures this intermediate space, encompassing systems that exhibit cognitive-like behaviors—such as self-monitoring, adaptive learning, and intentionality simulation, encompassing systems that exhibit cognition-like behaviors without genuine subjective awareness.

By defining five operational conditions—*Global Information Integration (GII), Recursive Metacognitive Correction (RMC), Cross-Domain Transfer Competence (CDTC), Intentionality Simulation Without Subjectivity (ISWS), and Behavioral Coherence Across Domains (BCAD)*—this study offers a systematic methodology to assess AI systems that functionally approximate cognition. These conditions differentiate pseudo-conscious AI from both narrow AI, which remains confined to domain-specific tasks, and AGI, which presupposes self-awareness and volitional reasoning.

Beyond its theoretical contributions, pseudo-conscious AI raises significant ethical, societal, and regulatory challenges. The potential for anthropomorphism, the risk of labor market disruptions, and the use of AI in misinformation campaigns necessitate scrutiny. These systems do not possess genuine cognition, still, their appearance of intentionality may foster misplaced trust, leading to regulatory blind spots and ethical misalignment. The growing integration of pseudo-conscious AI in critical domains—such as governance, healthcare, and financial decision-making—demands a robust regulatory framework that ensures transparency, accountability, and clear differentiation between AI capabilities and human cognition.

The framework established in this paper lays the groundwork for further research, both in advancing AI architectures and in developing rigorous evaluation metrics to measure pseudo-consciousness in computational systems. Future studies must explore whether current AI architectures—such as transformers, reinforcement learning, and meta-learning—can be refined to better approximate the five conditions. Additionally, the long-term stability of pseudo-conscious cognition remains an open question, requiring further investigation into whether such systems can consistently sustain cognitive-like properties across increasingly complex tasks. Addressing these gaps will be essential in determining whether pseudo-consciousness is merely an analytical tool for understanding AI cognition or a computationally viable paradigm for advanced machine intelligence.

More than a transitional phase toward AGI, pseudo-conscious AI represents a distinct and functionally stable category that redefines how artificial intelligence is conceptualized. As AI continues to evolve, recognizing and properly classifying pseudo-conscious systems will be critical in shaping responsible

innovation, ethical governance, and informed public discourse. By establishing a clear theoretical and operational foundation, this work provides a structured framework for future AI research, technological advancements, and policy development, ensuring that AI remains not only more capable but also more interpretable, controllable, and aligned with human values.

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.
- Bengio, Y., LeCun, Y., & Hinton, G. 2021. Deep learning for AI. Commun. *ACM 64*, 7 (July 2021), 58–65. https://doi.org/10.1145/3448250.
- 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, 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.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*(3), 200–219.
- Dehaene, S., Lau, H., & Kouider, S. (2017). 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.
- 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.
- Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), 335–346.

- Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. J. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5149-5169. https://doi.org/10.1109/TPAMI.2021.3079209.
- 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.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). Self-Refine: Iterative refinement with self-feedback. arXiv preprint arXiv:2303.17651. https://arxiv.org/abs/2303.17651.
- Meta Fundamental AI Research Diplomacy Team (FAIR) et al., Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science 378*, 1067-1074 (2022). https://doi.org/10.1126/science.ade9097.
- 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.
- OpenAI. (2024). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*. https://arxiv.org/abs/2303.08774.
- 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*. https://doi.org/10.48550/arXiv.2103.00020.
- Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., ... & de Freitas, N. (2022). A generalist agent. *arXiv preprint arXiv:2205.06175*. https://arxiv.org/abs/2205.06175.
- 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.
- Schmidhuber, J. (2006). Goedel Machines: Fully Self-Referential Optimal Universal Self-Improvers. *arXiv preprint arXiv:0309048*. https://arxiv.org/abs/cs/0309048.
- Schrittwieser, J., Antonoglou, I., Hubert, T. et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature 588*, 604–609 (2020). https://doi.org/10.1038/s41586-020-03051-4.
- 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., 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.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez. A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D. (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.
- Singer, P. (1975). *Animal liberation.* HarperCollins.
- Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of medical Internet research, 22*(3), e16235. https://doi.org/10.2196/16235.
- 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.
- 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.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903. https://arxiv.org/abs/2201.11903.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., & Zhu, H. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE, 109*(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555.