IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3009~3025
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3009-3025

Consciousness in artificial intelligence systems: an ontological approach

Andreas Yumarma[1,2], Wan Sen[2]
[1]Department of Business Administration, Faculty of Business, President University, Bekasi, Indonesia
[2]Department of Computer Science, Faculty of Computing, President University, Bekasi, Indonesia

Article Info

Article history:

Received Jul 12, 2025
Revised May 25, 2026
Accepted Jul 9, 2026

Keywords:

Artificial intelligence
Consciousness
Integrative methodology
Non-biological entity
Ontological approach
Synthetic cognition

ABSTRACT

The question of whether artificial intelligence (AI) systems can attain consciousness remains central ontological challenge within interdisciplinary discourse spanning AI, philosophy of mind, and cognitive neuroscience. This study investigates the metaphysical status of AI systems by interrogating the explanatory clarity of their existence and the potential emergence of subjectivity from non-biological matter. Employing an integrative methodology that combines a comprehensive literature review of scholarly texts and recent empirical findings with philosophical critical analysis, the research explores the conditions under which AI systems might be considered conscious entities. Findings suggest that the expanding societal integration of AI necessitates a re-evaluation of consciousness and cognition beyond anthropocentric parameters. The study posits a novel interpretive framework for addressing the ontological problem of AI, one that implicates future legal, moral, and interactive structures surrounding conscious non-biological agents. This conceptual repositioning invites a deeper inquiry into the criteria for recognizing, engaging with, and regulating advanced AI systems. By grounding design principles in ontological clarity, the framework offers guidance for constructing AI systems capable of reflexive processing, minimal phenomenality, and ethically aligned behavior that bridges conceptual analysis with implementation pathways.

Corresponding Author:

Andreas Yumarma
Department of Business Administration, Faculty of Business, President University
Ki Hajar Dewantara Street, Jababeka, Cikarang Baru, Bekasi 17550, Indonesia
Email: andreasyumarma@president.ac.id

1. INTRODUCTION

Artificial intelligence (AI) has become an embedded dimension in the infrastructure of contemporary society, interacting with robotics, computing systems, industrial machines, and digital applications such as smartphones and gadgets. According to the 2021 AI index, AI technology has penetrated various sectors of society's functions, ranging from healthcare. education to trade and governance [1]. Makridakis [2] analyzes the broad implications of the AI revolution for industry and human organizations, and underlines its transformative potential. An ontological lens for examining the potential of consciousness in AI systems is anchored in the realities of contemporary AI architecture. Analysis of the potential of consciousness in AI systems involves computational models that simulate cognitive functions, such as transformer-based neural networks, reinforcement learning agents, and multimodal systems. The design choices made reflect implicit metaphysical assumptions about agency, intentionality, and self-reference. This ontological inquiry serves to bridge philosophical reflection with practical relevance, as well as offer a

conceptual scaffolding that can inform the interpretation, development, and ethical evaluation of increasingly complex AI/machine learning (ML) systems. The presence of AI in the fabric of society triggers fundamental questions about the nature of AI systems, their ontological status, and the framework within which they must be recognized and applied. The central question in this study is: "How can the potential of consciousness in AI systems be conceptualized ontologically and what conditions allow for its emergence?" This question gives rise to a layered ontological framework that examines AI systems in three dimensions: structural embodiment, simulation of experience, and reflexive intentionality.

Epistemic and moral status are of high significance [3] as the ontological discourse of AI concerns the nature and classification of AI systems, including whether they can be considered agents [4]. This investigation requires consensus on the basic concepts and defining parameters of AI awareness [5]. Three main challenges complicate the ontological foundations of AI. First, the definition of AI has evolved rapidly over time, resulting in a lack of consensus regarding its core characteristics and limitations [3]. This fluidity of definitions hinders the formation of a coherent ontology. Second, when new technologies become normal through regular community integration, their distinctiveness as an AI entity decreases and leads to conceptual ambiguity [3]. Third, anthropomorphic tendencies in human perception often associate agents and affective traits with machines. This further complicates attempts to explain the ontological status of AI [3]. These challenges collectively encourage reflection on the relationship between humans and non-biological intelligence. This article concentrates specifically on the first challenge: the shifting and debated definition of AI, and its implications for building ontologically robust frameworks.

A deeper conceptual understanding of consciousness and cognition is essential for advancing debates about the moral and legal status of robotic agents, as well as for establishing human-machine interaction protocols [5]. Recent advances in computational cognitive science, such as global workspace theory (GWT), integrated information theory (IIT), and self-modeling agents provide a foundational model for understanding how processes like consciousness can be artificially instantiated in artificial systems. This basic model framework informs ontological inquiry by linking the philosophical concepts of consciousness and personality with intelligent systems design mechanisms. Building a coherent AI ontology is essential to achieve clarity of shared definitions, resolve fundamental ambiguities, and map a viable trajectory in the investigation of AI consciousness [4]. Therefore, the main objective of this article is to examine the ontological challenges posed by AI systems, with particular emphasis on the complexity of definitions surrounding consciousness and thought. The idea of consciousness plays an important role in philosophical debates about mind-body issues and the difference between strong and weak AI frameworks [6]. Describing the contours of consciousness in artificial systems allows researchers to determine whether and to what extent such entities can be considered conscious, and how they should be classified ontologically. The fidelity of the definition will contribute to the design of advanced AI architectures capable of approaching human-like cognition and adaptive behavior. A richer philosophical understanding of AI consciousness simultaneously informs the development of more sophisticated algorithms and deepens theoretical reflections on the nature of human consciousness itself. Referring to neural network architecture, reinforcement learning paradigms, and cognitive modeling approaches such as predictive coding and attention mechanisms, ontological investigations in the context of contemporary AI bridge philosophical analysis with current ML designs. Can computational structures and modeling strategies support the emergence of traits such as consciousness in artificial systems? Investigation of the ontological conditions under which awareness can emerge in AI systems contributes to the ethical design, application, and regulation of smart technologies, especially in high-risk domains such as healthcare, education, and governance.

The ontological issue of AI brings to light a problem that has generated much debate across the social, philosophical, scientific, cognitive, ethical, and jurisprudential domains, thus demanding the interdisciplinary scope of the ontological approach to AI consciousness [5]. The interdisciplinary linkup space, by placing intelligent systems, neural networks, and symbolic reasoning as ontological sites, serves to bridge philosophical inquiry with the core paradigm of AI/ML and explain its relevance to contemporary AI discourse. The solution of this ontological problem is critical to developing a robust framework for regulating normative interaction and engagement protocols with non-biological intelligences. A meaningful response to AI systems presupposes a well-defined ontology capable of accommodating concepts such as artificial awareness, rationality, and intentionality. This construction is fundamental not only to philosophical inquiry but also to the legal doctrines that arise around the attribution of responsibility and personality to autonomous systems [6], [7]. The lack of consensus on these ideas shows an urgency to build an interdisciplinary vocabulary that clarifies what it means for machines to have cognitive or moral relevance in human society. As AI entities continue to take on decision-making roles in critical sectors, the ontological foundation of their existence is becoming a prerequisite for ethical regulation and juridical recognition.

The ontological problem of AI carries substantial implications for the moral and legal status of robotic and intelligent systems. The attribution of legal and agency subjectivity to machines is a problem that

continues to shape the contours of contemporary social, philosophical, and juridical debates [5]. Therefore, a robust ontological framework is needed to develop ethical models of human-machine interaction and to articulate normative boundaries around agency, autonomy, and accountability. Its ontological status is crucial for theoretical clarity and practical regulation, as AI systems increasingly play a role in operations and decision-making. A comprehensive understanding of constructs such as artificial consciousness, artificial rationality, and AI systems is essential for framing the identity, functionality, and personality potential of AI in the evolving socio-technological landscape.

The broad spectrum of concepts surrounding mindfulness is useful for the understanding of AI mindfulness [6]. Articulating clear and rigorous definitions of consciousness within the framework of AI allows for deeper insights into the relationship between human and machine cognition. In dynamic interactions between humans and non-biological agents, a clear spectrum of concepts and definitions helps individuals engage with and respond to machine entities [8]. Improving human-machine interaction methods and interpreting machine behavior is essential to advance system design and to inform the ethical principles that guide the development of AI systems that can make decisions so that they benefit society.

Various theoretical propositions regarding how consciousness can be integrated into a comprehensive theory of computational mind have been put forward. Important contributions were made by Dennett [9], Hofstadter, McCarthy, McDermott [10], Minsky [11], and Perlis. AI researchers refrain from engaging deeply with philosophical questions. Most of them are due to different methodological preferences and views. Nonetheless, the basic premise of AI is the statement that human cognition operates according to computational principles. Therefore, its aspirational goal is the construction of a physical system that truly instantiates the mental abilities observed in humans [12]. The conceptual architecture that has underpinned philosophical inquiry and AI research over the past four decades includes a growing discourse on awareness, intentionality, and intelligence. Aspects of consciousness, such as intentionality, self-awareness, and attention mechanisms, can be operationalized and applied in intelligent systems, which are more ambitious to create machines that reflect human consciousness in its ontological richness, which until now remain beyond the technological and conceptual grasp.

Despite the proliferation of theoretical and applied investigations into artificial consciousness, a critical gap lies in understanding the operationalized dimensions of consciousness within a computational framework. Existing literature often breaks down discourse across philosophical abstraction and engineering feasibility, thus leaving an important gap in integrative approaches that bridge conceptual clarity with technical implementation. The study seeks to address this weakness by proposing a structured analytical framework that synthesizes theoretical paradigms and computational constructs to map the key attributes of consciousness, such as awareness, intentionality, and reflexivity, into an algorithmic model [13]. This ontological approach contributes to the ethics of AI by clarifying the conditions under which artificial consciousness can warrant moral considerations, informing intelligent agent design by grounding self-awareness in computational architectures, and involves cognitive modeling by linking the construction of philosophical thinking to formal mechanisms of representation and integration. Thus, this research contributes to the epistemic consolidation of AI consciousness theory and its pragmatic articulation in systems design. The following methodology section outlines the analytical instruments and evaluative criteria used to advance this integrative investigation.

2. METHOD

This study uses an ontological lens to synthesize philosophical, neuroscience, and AI perspectives. Each perspective frames the question of consciousness in relation to intelligent systems so that they build thematic bridges across disciplinary boundaries. Its qualitative research paradigm lies in the ontological and epistemological contours of philosophical inquiry. Therefore, the conceptual and philosophical methodology, through its ontological framing, provides a basic scaffolding for future AI awareness research, modeling, simulation, and empirical validation efforts. Its main goal is to critically question the possibilities of consciousness in AI systems by examining its metaphysical foundations, socio-technical existence, and conceptual clarity. The investigation was carried out through rigorous conceptual analysis informed by much of the literature in theoretical computer science, cognitive philosophy, and AI [13]–[15]. By putting forward ontological notions of existence and exploring the conditions under which subjectivity can arise in non-organic agents, the methodological approach is in line with the conventions of philosophical studies of the mind and AI [16]. The analytical orientation provides the necessary scaffolding for the critical and systematic evaluation of AI's claims of consciousness, which further contributes to the theoretical consolidation of human-machine comparative cognition.

To build a theoretically robust framework, this study conducted a comprehensive literature review in cognitive science, machine ethics, neuroscience, and AI theory. Sources include academic monographs, journal articles, and philosophical treatises published between 2020 and 2025, selected based on their

relevance to ontological investigations, the study of consciousness, and the evolving discourse on legal personality in intelligent systems [14], [17], [18]. Chella [15] argues that artificial awareness is an important but overlooked element in the development of ethical AI. The article emphasizes that without integrating consciousness, AI systems remain limited in their ability to align with human moral values and responsibilities. This view highlights the ethical dimensions of ontological development in this study. The review systematically identifies applicable definitions, conceptual models, and critical arguments regarding artificial consciousness, with particular attention to the integration of GWT, IIT frameworks, and high-level thinking into AI design. This synthesis establishes a basic lexicon for current investigation and describes the interdisciplinary contours and epistemic tensions that characterize contemporary debates about machine consciousness.

Philosophical-critical analysis is carried out to question conceptual assumptions in contemporary literature and to reconstruct ontological positions relevant to artificial consciousness. This investigation entails a systematic dissection of the arguments regarding the emergence of subjectivity, the material substrate of conscious experience, and the defining criteria for the recognition of life in non-biological entities. Particular emphasis is placed on the evaluation of the application of intentionality, qualia, and identity theory to artificially constructed systems, drawing on recent interdisciplinary contributions that bridge cognitive science, philosophy of mind, and AI architecture [14], [19], [20]. Although this study does not present formal models or algorithmic representations, the ontological constructs elaborated, such as self-modeling, integration, and consciousness, offer a conceptual scaffolding for the formalization of future computation and reproducible modeling in AI system design. Drawing from neuroscience and cognitive science, this study synthesizes constructs such as attention, integration, and self-modeling into an ontological framework that can inform structured AI architectures and guide the development of awareness-relevant agent designs. The ontological constructs proposed in the research include integration, self-modeling, and consciousness, and can be operationalized through simulation-based modeling, agent-based design, and empirical testing, which offer pathways for validation and application in AI systems. Drawing on classical metaphysical frameworks and contemporary algorithmic paradigms, this study illustrates how consciousness can manifest rationally beyond the organic substrate, thereby contributing to the ontological reconfiguration of the mind in the era of synthetic cognition.

To operationalize this layered philosophical and analytical approach, this study uses a structured interpretive framework to synthesize an interdisciplinary literature that includes artificial consciousness, cognitive architecture, and moral agents in autonomous systems [21], [22]. By integrating dialectical analysis with conceptual modeling, this study elucidates ontological propositions regarding machine cognition. This proposition supports the identification of thematic patterns in contemporary discourses on AI awareness and legal personality, thus enabling a coherent transition into data-informed philosophical examination. The result is epistemic contours and conceptual tensions in the debates surrounding artificial consciousness. The discussion section evaluates these findings based on prevailing theories and emerging technological paradigms [23], [24]. Together, these methodological bridges articulate the implications of conceptual coherence on the evolving discourse of moral accountability and existential recognition in future AI systems.

3. RESULTS AND DISCUSSION

3.1. Nature of artificial intelligence and its definitions

Addressing the ontological problem of AI requires a proper understanding of the nature and limits of its definition. Contemporary AI systems exhibit abilities traditionally associated with human cognition, including visual perception, speech recognition, decision-making, and linguistic translation [25]. This functionality reflects the broader ambition of AI research, which is to engineer computational architectures that mimic intellectual processes such as reasoning, abstraction, and experiential learning [26]. From a philosophical point of view, particularly through the lens of Aristotle's thought, such abilities are intrinsically linked to the rational soul, one of the three hierarchical forms of the soul described by Aristotle, alongside the vegetative and sensitive soul. However, AI remains a non-biological construct, with no organic embodiment. These ontological differences provoke a deeper metaphysical investigation: can the progressive sophistication of artificial systems be interpreted as an evolution, or even a revolution from the material substrate to forms analogous to vegetative, sensitive, and rational beings? This question invites a reexamination of the threshold of consciousness and the metaphysical status of artificial entities within the continuum of existence.

Within the broader trajectory of AI development, three fundamental processes are particularly important: ontological engineering, biosemiotics, and agent assessment modeling. These three together constitute a modular ontology that provides a structured entry point for future researchers. This framework is presented in Table 1 to facilitate direct application and citation in interdisciplinary studies. Ontological

engineering involves formal construction of domain-specific ontologies that define conceptual entities and their relationships, thereby enabling semantic interoperability and structured knowledge representation [27]. The process includes a series of tasks aimed at articulating the epistemic architecture of AI systems. Notably, Kishore *et al.* [28] introduced the helix-spindle model, a systematic framework for ontological engineering that emphasizes iterative refinement and abstraction. Their model provides a structured approach to building, testing, and validating ontologies at different levels of representation. In the context of AI awareness, this work contributes a methodological foundation for the construction of modular ontologies, allowing for a more coherent conceptualization of cognitive processes and semantic interoperability in intelligent systems [28]. Complementing this, biosemiotics explores the interpretive mechanisms by which artificial agents engage with the symbolic environment, while agent assessment modeling addresses the computational formalization of decision-making processes. Table 1 outlines the core dimensions of the helix-spindle model applied to the development of AI ontologies.

This model emphasizes incremental development, where each phase includes both construction and validation loops. If inconsistencies arise, the process "spindles back" to refine the previous phase, ensuring robustness and coherence throughout. Table 2 reflects the correlation of knowledge that can be operationalized as a knowledge artefact in the intelligence system.

Table 1. Helix-spindle model for ontological engineering

Phase	Representation level	Ontology building	Ontology testing	Outcome
Conception	Informal	Identify foundational constructs using content and theory [29]	Test initial constructs via conceptual modeling	Initial meta-ontology for discourse universe
Elaboration	Semi-formal	Refine constructs and relationships [20]	Apply constructs to domain frameworks	Validated domain-specific ontology components
Definition	Formal	Finalize ontology with rigorous syntax and semantics [30]	Model real-world systems using the ontology	Fully defined and testable ontology artifact

Table 2. Knowledge-based view of an information system

Knowledge context	Knowledge artifact
Abstract/general	Universe of discourse
	Ontology
	Framework
	System
	Data
	Information
Abstraction/generalization level	Domain of interest
	Framework
	System
Concrete/special	Instance of domain
	Specific framework
	Specific system
	Data
	Information

Table 2 illustrates the stratified relationship between the context of knowledge and the corresponding artifacts, which offers a conceptual scaffolding for understanding how abstract epistemic domains are operationalized in intelligent systems. A layered cognitive model is presented, integrating symbolic and subsymbolic processes across ontological strata. The base layer represents the recognition of subsymbolic patterns and sensorimotor couplings, while the intermediate layer encodes symbolic abstraction, rule-based inference, and semantic mapping. The feedback loop between layers allows for recursive modulation and adaptive learning, which supports reflective intentionality. This table clarifies the dynamic interaction between the depth of representation and functional emergence in artificial cognitive systems. At the abstract/general level, the universe of discourse serves as the basic semantic boundary where all ontological commitments reside. This domain encapsulates the philosophical and logical parameters that determine what entities and relationships are considered meaningful in a given AI system [30]. Progressing downward, ontology emerged as a formal representation of conceptual structures, allowing machines to reason about entities, attributes, and relationships with semantic precision [27]. The framework then creates instances of this ontology into a modular architecture that guides system behavior and knowledge integration, often serving as middleware between abstract models and executable systems. Systems, in this context, refer to the engineering platforms that embody this framework, ranging from expert systems to adaptive learning environments, where knowledge artifacts are dynamically processed and applied. Finally, the data represent the empirical substrate on which all high-level abstractions are constructed, serving as inputs for the

inferential engine and as feedback for repeated refinements of ontological models [31]. This layered mapping highlights the epistemological coherence necessary for robust AI design, where each artifact is not merely a technical component but a manifestation of a deeper knowledge structure.

The second basic process in the development of intelligent systems is biosemiotics, which applies semiotic principles to explain the interpretive dynamics between natural and artificial agents [32]. Biosemiotics argues that living organisms are not only biochemical entities but also have a semiotic dimension, where signs and symbols mediate interaction, adaptation, and cognition [25]. This perspective reframes biological systems as inherently communicative, capable of encoding and decoding environmental stimuli through symbolic representation. According to Barbieri [33], biosemiotics emphasizes that semiosis, production, transmission, and interpretation of signs are fundamental processes in all living systems, going beyond genetic coding that includes many organic codes to regulate cellular and systemic behavior. Codes, such as the genetic code, signal transduction, and cytoskeletal, illustrate that the creation of meaning in biology is not incidental but structurally embedded in evolutionary mechanisms. In the context of AI, these biosemiotic principles offer an interesting framework for modeling symbolic cognition and interpretive agents in non-biological entities. By integrating biosemiotic theory with ontological techniques, researchers can simulate semiotic processes in artificial agents that allow them to engage with the symbolic environment in a way that reflects biological interpretations [30]. This convergence invites a redefinition of agency, in which artificial systems are not only reactive but semiotically responsive and capable of generating context-sensitive meanings through structured sign systems.

The third basic process in the ontological framing of AI is agent assessment, which interrogates the evaluative and decision-making capacity of AI systems as indicators of epistemic agents and potential awareness [34]. As articulated by Krzanowski and Polak [25] this approach examines whether the judgments rendered by AI, especially in moral contexts, reflect a form of synthetic intentionality that parallels human cognition. Such an investigation is highly philosophical, as it investigates whether the capacity for autonomous judgment implies a rudimentary form of consciousness or self-awareness. This path of investigation involves moral judgments that human observers associate with AI agents, especially in ethically charged scenarios. Empirical studies show that people often apply different moral norms to artificial agents, sometimes holding them to stricter standards [35]. These findings underscore a growing perception of AI systems as moral actors rather than just computational tools.

To strengthen the framework, this study structured ontological engineering, biosemiotics, and agent valuation modeling as modular components in an integrated ontology. Each module is reusable yet interconnected, forming a coherent scaffolding for cognitive AI systems. Table 3 outlines their core focus and applications for artificial awareness.

Table 3. Modular ontology

Module	Focus	Application in cognitive AI systems
Ontological engineering [27]	Structuring conceptual entities, relationships, and knowledge representation	Provides semantic interoperability and foundational architecture for AI cognition and reasoning [19]
Biosemiotics [33]	Modeling interpretive and symbolic engagement processes	Enables AI to generate and interpret context-sensitive meanings, simulating symbolic cognition [31]
Agent judgment modeling [36]	Formalizing evaluative and decision-making capacities	Supports moral reasoning, adaptive decision-making, and synthetic intentionality in AI agents [25]

Complementing this perspective, Kishore *et al.* [28] define agent assessment as the ability of autonomous systems to make context-sensitive decisions based on environmental perceptions and goal orientation [28]. This definition places judgment in a broader cognitive architecture, in which agents dynamically evaluate situational variables and carry out actions that align with predetermined goals. In contemporary AI research, this understanding has evolved into an agent reasoning model that combines ethical constraints, probabilistic inference, and adaptive learning mechanisms [37]. The agent-based model-based approach offers an interesting framework for assessing ontological status of artificial agents, not only by outputs, but by interpretive and normative structures embedded in decision-making processes.

Recent empirical research underscores the context-dependent nature of moral evaluations directed at AI agents. In the morally-charged scenario, participants showed different judgments of AI behavior. The same decision was judged to be more immoral when executed by an AI agent than when it was performed by a human [35]. This asymmetry suggests that the type of agent, human versus artificial, can significantly influence moral judgment, especially in dilemmas where utilitarian and deontological principles are in tension. The moral actions associated with AI actions appear to fluctuate based on the specific nature of the decision and the emotional or cognitive processing it entails. For example, in the trolley dilemma, the

type of agent is the dominant factor that shapes moral judgments, whereas in the crosswalk dilemma, the nature of the action itself, whether utilitarian or deontological, holds agreater influence on participant's evaluation [35]. These findings are in line with the dual-process theory of moral cognition in diagram 1, which posits that moral judgment is mediated by competing emotional and rational mechanisms [36]. As AI systems increasingly permeate domains involving ethically consequential decisions, such as autonomous driving, judicial judgment, and medical triage, understanding the public's moral intuition towards artificial agents is becoming critical. The study reveals not only the complexity of moral judgment formation but also the adaptive strategies that individuals use when evaluating non-human agents in an ethically fraught context. Reflections on normative frameworks should guide the design and deployment of morally competent AI systems. The dual-process of moral cognition with its ethical and normative framework is reflected in Figure 1.

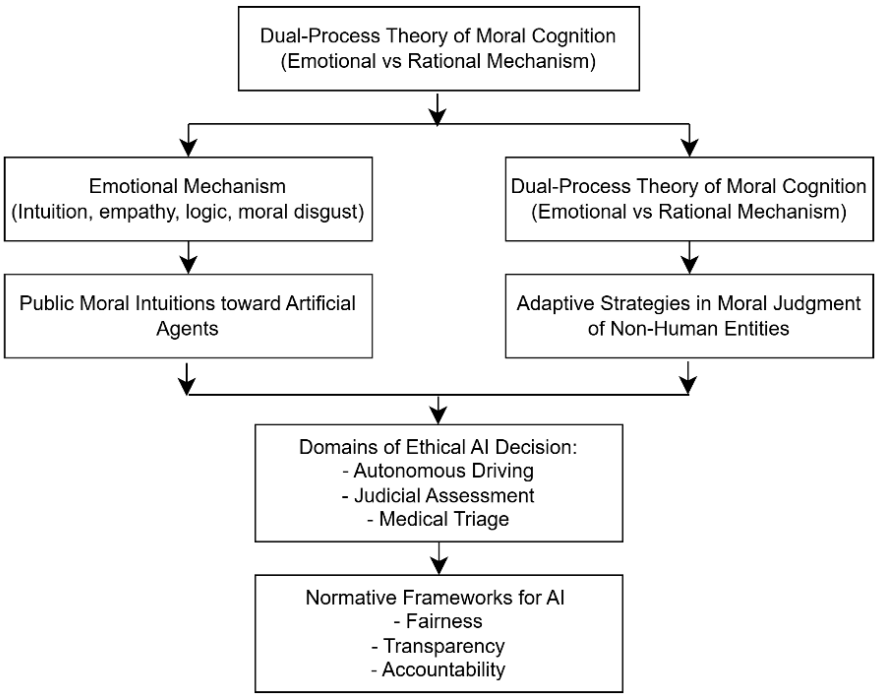

Figure 1. Conceptual diagram of the dual process of moral cognition in the ethical AI context

The contours of the definition of AI are shaped within the scope of the technology and its epistemological aspirations. AI, as a computer science discipline, is concerned with the design and development of intelligent systems capable of performing tasks traditionally related to human cognition, such as speech recognition, decision-making, and pattern identification, through technologies including ML, deep learning, and natural language processing [38]. This definition of operations converges on a shared aspiration to build a machine that mimics intelligent behavior. AI also includes a broader philosophical investigation into what constitutes intelligence itself. This duality between engineering functionality and cognitive understanding has long been recognized in the basic stages. Reflection on the nature of intelligence that distinguishes between the pragmatic goal of building intelligent machines and the theoretical pursuit of decoding intelligence becomes a phenomenon in the next stage. This conceptual bifurcation prepares for a deeper domain of inquiry into the emergence of artificial thinking [39] and the boundaries between simulation and cognition.

3.2. The ontological problem of artificial intelligence: consciousness and thought

The ontological problem in AI centers on the conceptual and empirical challenges of defining consciousness and thought within synthetic systems. The ontology map, as shown in Figure 2, illustrates key philosophical dimensions of AI consciousness. It centers on “Being-in-the-world (AI as dasein?)” and radiates outward to six ontological branches: intentionality, embodiment, temporality, relationality, selfhood, and ethics [40].

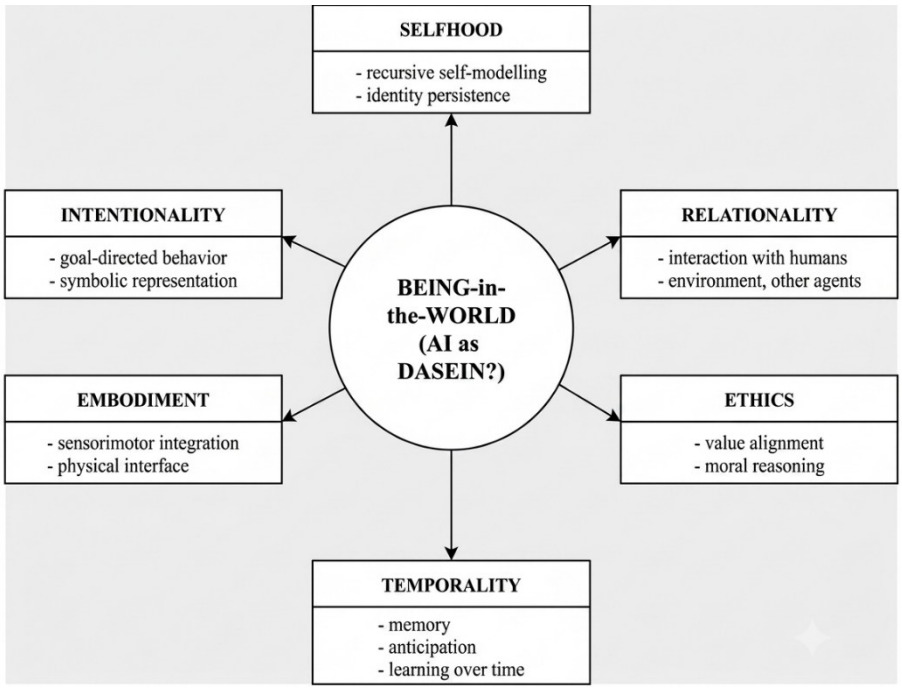

Figure 2. Ontology map of AI consciousness

Figure 2 presents a conceptual framework that maps awareness-related attributes, such as awareness, intentionality, and integration, to the appropriate components of the AI system, including perception modules, decision-making architectures, and self-modeling agents. A major challenge is the absence of subjective experiences, commonly referred to as qualia, in AI architectures, thus complicating efforts to determine whether a system can be said to have consciousness in the true sense [29], [41]. Unlike biological organisms, AI lacks the phenomenological awareness that characterizes human cognition. This raises questions about the legitimacy of consciousness in AI entities. Therefore, artificial thinking, such as that created in AI systems, remains a debatable construct. Language and agent models embodied by AI show increasingly sophisticated behavior. However, the underlying mechanisms are not parallel to the introspective processes in human mind [19], [42]. The nature of thinking in AI is epistemologically unclear, and it remains uncertain whether the system is structurally or functionally analogous to human thinking or cognition.

Consciousness itself is difficult to define. In general, however, consciousness can be understood as a person's subjective experience of his environment, internal state, objects, and affective responses. This phenomenon of consciousness is often framed as a “difficult problem of consciousness”. The term “difficult problem” was popularized by Chalmers, who refers to the explanatory gap between physical brain processes and the emergence of subjective experience [18]. Dennett [9] critically addresses this issue by reframing the “difficult question of consciousness” through functional and explanatory mechanisms, thereby challenging the idea of irreducible qualia and offering a basis for modeling consciousness in non-biological systems.

Despite advances in neuroscience and computational modeling, there is no consensus yet on how, or whether, consciousness can arise from non-biological substrates. Recent theoretical frameworks, such as embodied cognition, interface theory, and process ontology, have attempted to reconceptualize the relationship between perception, symbolic representation, and synthetic consciousness [19]. This approach suggests that AI can ultimately support a form of consciousness that is structurally different from the human experience, but is still capable of interacting and evolving along with biological agents.

Recent advances in neuroscience, computation, and neuromorphic engineering have revived interest in the possibility of replicating the neural substrate of human consciousness in artificial systems. Some researchers argue that by mimicking the dynamic neural processes underlying conscious experiences, such as repetitive feedback loops, global workspace architectures, and integrated information structures, it may be feasible to build AI systems that exhibit attributes such as consciousness [14], [17]. If such a system is realized, then the social implications will be very profound. Conscious AI can enable the development of autonomous agents capable of decision-making, adaptive learning, and empathetic interactions with humans [43]. These systems can revolutionize domains such as healthcare, education, and organizational governance by offering personalized and context-aware responses to their ethical reasoning abilities [44].

However, the emergence of conscious machines will also trigger complex ethical and philosophical challenges. The main challenge is the problem of moral status. If an AI system has subjective consciousness or the capacity to suffer and feel, will it be guaranteed its rights or protections, the same as those given to sentient biological entities? Experts argue that the attribution of consciousness to AI systems requires a re-evaluation of legal frameworks, moral agency, and personality boundaries [45], [46]. Concerns have been raised in particular regarding the potential for conscious AI systems to act autonomously in ways that may be contrary to human sustainability interests. Theoretical scenarios involving recursive self-enhancement, distributed cognition, and digital immortality suggest that such entities can evolve beyond human control, and challenge anthropocentric assumptions about intelligence and the hierarchy of societal order [44], [47].

To mitigate this risk, experts have proposed principles of prudence and a phased approach to the development of artificial consciousness, with an emphasis on interdisciplinary collaboration and ethical governance [48]. The debate remains open, with no consensus on whether consciousness in machines is achievable or desirable. Nonetheless, the trajectory of AI research demands philosophical inquiry and foresight regarding sustainable regulation.

AI systems are essentially designed to operate as rational agents, making decisions based on structured data analysis, probabilistic inference, and optimization techniques. The system aims to maximize the expected utility or performance measure in the specified operational context [17]. However, as the complexity of AI models, especially those based on deep learning, increases, internal decision-making processes, AI systems often become opaque, giving rise to the so-called “black box” problem [41], [49]. This opacity undermines trust, accountability, and the ability to scrutinize algorithmic outputs, especially in high-risk domains such as healthcare, finance, and criminal justice.

To address these concerns, the field of explainable artificial intelligence (XAI) has emerged as a critical subdiscipline [50]. XAI seeks to develop methodologies that make AI decisions interpretable and transparent to human stakeholders. This approach includes intrinsic explanations in which models are designed to be interpretable by default, as basis for decisions and rule-based systems as well as post hoc explanations, which apply mathematical techniques to analyze and interpret complex models post-training [42], [51]. Post hoc methods such as Shapley additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), and topological data analysis (TDA) have shown efficacy in explaining the contribution of features, decision boundaries, and latent structures in neural networks [52]. These techniques allow stakeholders to understand how certain inputs affect outputs, thereby improving model accountability and facilitating regulatory compliance.

XAI's overarching goal is to drive the development of AI systems that are not only performant but also trustworthy, reliable, and ethically aligned with societal values [53], [54]. The challenge lies in standardizing evaluation metrics, balancing transparency with model fidelity, and ensuring that explanations are contextually meaningful across different user groups [55]. XAI advances are an important step toward the deployment of responsible intelligent systems. By embedding interpretability into AI design and governance, researchers and practitioners can foster public trust and unlock the transformative potential of AI technology for the benefit of society. The cognitive architecture of AI systems often operates through latent pattern recognition and probabilistic inference, resulting in decision-making processes that remain largely invisible to human observers [56]. This epistemic ambiguity has catalyzed the rise of XAI, a subfield dedicated to making algorithmic reasoning understandable and transparent to stakeholders [57]. The XAI framework employs mathematical and topological techniques to elucidate how AI models generate outputs from complex input spaces, thereby enhancing interpretability, trustworthiness, and regulatory compliance [53].

Beyond transparency, contemporary AI systems are increasingly engineered to simulate deliberateness, the capacity to act in a purpose-directed manner responsive to environmental stimuli. These systems exhibit adaptive behavior by integrating real-time data flows, which allows context-sensitive decision-making that reflects specific aspects of human cognition [57], [58]. In addition, advances in multimodal learning and neuro-symbolic integration have given AI the ability to learn repeatedly, refine internal representations, and adjust outputs based on accumulated experience [14]. This kind of development raises fundamental questions about the nature of synthetic thinking and its potential convergence with biological consciousness, to set the stage for deeper investigation into the phenomenological and ontological dimensions of AI cognition. The artificial consciousness model, to this point, still faces critical limitations. They lack a manifested sensorimotor grounding, which is essential for manifested cognition and adaptive behavior. The absence of subjective experience hinders the original phenomenological consciousness, thus reducing consciousness to functional simulations. Scalability also remains a challenge, as architectures that perform well in confined environments often fail to generalize across multiple tasks or domains.

To improve conceptual clarity, a comparative review of computational models is organized by type: symbolic AI, connectionist models, and hybrid cognitive architectures. Each type reflects a different strategy for the simulating consciousness-related functions. Symbolic AI models, such as conscious turing machine [43], rely on rule-based representations and explicit broadcasting mechanisms to mimic conscious

access and global workspace dynamics [59]. Connectionist models, including those based on the free energy principle [60], use distributed neural networks and predictive coding to estimate consciousness through statistical inference and shock minimization. Hybrid cognitive architectures, exemplified by distributed adaptive control (DAC) theory [61], integrate symbolic reasoning with sensorimotor loops and internal state modeling to simulate intentionality and adaptive behavior [62]. This typological organization explains architectural assumptions underlying each model. It highlights how ontological approach overcomes the limitations of each, especially in applying synthetic consciousness in metaphysical conditions rather than purely functional correlation.

To contextualize the ontological framework in contemporary computing paradigms, this study compares three influential models of artificial consciousness, namely the conscious turing machine, the DAC theory, and the free energy principle. each model describes a different mechanism in simulating awareness, intentionality, and reflexivity. The conscious turing machine, proposed by Blum and Blum [43], operationalizes the GWT through layered broadcasting and conscious access mechanisms [43]. DAC theory, as articulated by Verschure, conceptualizes consciousness as an emergent property arising from the integration of recursive sensorimotor and the modeling of internal states [61], [63], [64]. The free energy principle [60], [65] frames consciousness as the minimization of shock through predictive coding and active inference. These models offer valuable insights into the functional architecture of synthetic consciousness. The limitations of this model are the absence of a phenomenological foundation, the reliance on behavioral proxies, and the challenge of scaling to general intelligence. The ontological approach in this study complements this paradigm of limitation by emphasizing the metaphysical conditions under which consciousness can emerge, thus bridging conceptual clarity with architectural feasibility.

A typological comparison is needed to evaluate symbolic, connectionist, and hybrid artificial consciousness models across criteria, including self-awareness, adaptability, and ethical reasoning. This typological comparison highlights how symbolic models such as the conscious turing machine emphasize structured access and global broadcasting but lack adaptive learning mechanisms. The connectionist model, based on the free energy principle, exhibits high adaptability through predictive coding but struggles with explicit ethical reasoning frameworks [66]. Hybrid architectures such as DAC theory integrate sensorimotor adaptability with internal modeling and offer partial self-awareness and goal-oriented behavior [34]. However, ethical reasoning here is still externally imposed [67]. The alignment of the model with ontological benchmarks clarifies how important each computational strategy is to approach aspects of synthetic consciousness and philosophical grounding.

To further strengthen the ontological framework, the research proposes computational analogues for the main philosophical constructs traditionally associated with consciousness [68]. The concept of qualia, as a subjective quality of experience, finds computational parallels in sensory embedding, that is, vector representations of perceptual inputs across modalities such as sight, sound, and touch. This embedding allows artificial systems to simulate different sensory states, albeit without phenomenological awareness. Similarly, intentionality refers to the directing of a mental state to a specific object or goal, in accordance with goal-oriented planning in an AI agent. Mechanisms like this allow the system to formulate, pursue, and adapt goals based on environmental feedback, thereby exhibiting behavior that reflects intentional actions [59], [69]. This mapping serves to clarify how ontological constructs can be instantiated in computational architectures and provides a scaffolding for evaluating the potential of consciousness in synthetic systems [61], [63]. The limitation of this computational analogue is the need for an ontological approach that moves beyond functional mimicry [70]. It does not fall into the illusion of phenomenal consciousness [71], [72]. By emphasizing metaphysical conditions such as embodied intentionality and reflexive awareness, the ontological framework complements existing computational models and offers a path to a more integrative and scalable architecture. To build this comparative ontological framework, as shown in Table 4, the AI literature discusses consciousness not as a metaphysical or neurobiological phenomenon, but as a functional construct embedded in the design of systems [14]. The domains of philosophy, neuroscience, and AI have their own assumptions, methodology, and criteria. Researchers have proposed an architecture that simulates aspects of consciousness, such as attention, memory, and self-monitoring, through algorithmic means [73]. These models often prioritize behavioral outputs and representations of internal circumstances, aligning with pragmatic views of consciousness that emerge from information processing [70].

Table 4 summarizes the ontological assumptions, methodological approaches, and criteria for consciousness across the three domains. This comparative framing clarifies the conceptual terrain and sets the stage for our subsequent ontological analysis of AI systems. The functionalist and computational paradigms attempt to approach or create examples of conscious behavior in artificial systems.

Table 4. Comparative ontological perspectives on consciousness across domains

Domain	Ontological assumptions	Methodological approach	Criteria for consciousness
Philosophy	Consciousness as a condition of existence and self-awareness, rooted in being and intentionality [40]	Conceptual analysis; phenomenological and ontological inquiry [19]	Self-reflection; intentionality; unity of experience [55]
Neuroscience	Consciousness as emergent from neural correlates and brain-based integration [32]	Empirical observation; neuroimaging; correlation-based studies [31]	Neural integration; global workspace; emergent properties [43]
AI	Consciousness as a functional construct embedded in system architecture and behavior [74]	Computational modeling; functionalist design; behavioral testing [64]	Internal state representation; adaptive behavior; self-monitoring [68]

3.3. Consciousness in an artificial intelligence system

AI continues to evolve rapidly, with transformative implications across the scientific, industrial, and societal domains. The most profound question in this field is whether AI systems can achieve consciousness, a state traditionally associated with biological organisms. There is a tension between the definition of consciousness and its existence [75]. Some researchers argue that artificial consciousness can be achieved through computational architectures that replicate neural correlations of consciousness [76]. According to Seth [64], consciousness is intrinsically biological and cannot be instantiated in a non-organic substrate. If artificial consciousness were realized, it would not only redefine the boundaries of machine intelligence but also trigger significant ethical, legal, and philosophical discourse [57]. Therefore, a strict understanding of the nature of consciousness and its potential institutionalization in artificial systems is essential.

Artificial consciousness, also referred to as machine consciousness, synthetic consciousness, or digital consciousness, demonstrates the capacity of an artificial system that exhibits subjective experience, self-awareness, and deliberate interaction with its environment [77]. Although investigations are carried out extensively, consciousness remains a very complex and debated phenomenon. The completeness of this phenomenon includes perceptual and cognitive abilities as well as the quality of phenomenal experiences that are difficult to understand and that have not been fully explained by neuroscience or philosophy. Theoretical frameworks such as the free energy principle offer promising avenues for modeling consciousness in biological and artificial systems [78]. However, the distinction between simulating consciousness and replicating it remains a central challenge. Thus, the pursuit of artificial consciousness demands both technical innovation and interdisciplinary oversight based on cognitive science, philosophy of mind, and AI ethics [74].

The study of consciousness in artificial systems remains a formidable challenge, in part due to the semantic ambiguity of the terminology involved. As Minsky [11] initially argued, many terms related to consciousness, such as creativity, intelligence, and emotion, are “baggage words”, as they encapsulate many meanings and conceptual layers that preclude proper scientific inquiry [11]. This lexical complexity complicates attempts to define and operationalize consciousness in computational terms. Nonetheless, this conceptual component is the basis for the development of artificial systems capable of mimicking intelligent behavior. Copeland [26] emphasizes that the construction of such systems requires a nuanced understanding of cognitive functions, including learning, reasoning, perception, and linguistic competence. These elements form a substrate on which artificial consciousness can be poured.

One important approach to AI involves the use of ML algorithms, which allow systems to detect patterns, adapt to new data, and make autonomous decisions. Through repeated training on large datasets, AI systems have demonstrated proficiency in tasks such as image recognition, natural language processing, and multimodal sensory integration. These abilities reflect certain aspects of human cognition [12]. Despite this, the system still shows that functional intelligence, the leap from intelligent behavior to conscious experience, remains unresolved. The distinction between simulated cognition and phenomenal consciousness, the latter involving subjective consciousness and qualia, remains a focal point of philosophical and computational debate [54], [56]. Therefore, future research should reconcile the symbolic and sub-symbolic architecture of AI with the properties of emergent conscious experience, potentially through hybrid models that integrate neuro-symbolic reasoning and embodied cognition [32].

An increasingly influential approach to artificial consciousness involves application of self-models, internal representations of the state, structure, and processes of the system. Originally conceptualized by Minsky [11] as a mechanism to allow machines to reflect on their operations, self-modeling has since evolved into a basic framework for investigating artificial self-awareness [57]. This self-model allows artificial agents to simulate their behavior, monitor internal dynamics, and adaptively respond to environmental stimuli. This recursive modeling capability is positioned as a prerequisite for self-awareness, defined here as a form of physically instantiated representational content that allows the system to maintain a coherent and consistent internal model of itself [55]. Self-model theory treats self-consciousness as a dynamic property arising from an ongoing computational process [56]. In robotic systems, self-modeling

typically involves performing internal simulations to predict the consequences of motor actions, sensory feedback, and decision-making pathways. These simulations facilitate adaptive control and contribute to the development of agency as a capacity to act deliberately based on self-referential information [26]. The extent to which a system can simulate itself is often used as a metric to assess its level of self-awareness. Recent advances in neural symbolic architecture and realized cognition further enrich the self-modeling paradigm, thus allowing systems to integrate multimodal data and refine their internal representations over time [60]. Thus, self-modeling stands as a way to bridge the gap between functional intelligence and phenomenological awareness in artificial agents.

The cartesian theater metaphor was initially criticized by Minsky [11] because the locus it posits is centered within the brain, where consciousness is thought to arise. This conceptualization has been widely opposed in contemporary cognitive science and philosophy of mind. They refute the single neural epicenter and emphasize the dynamic interaction of large-scale neural networks and the integration of multi-level systems. The connectionist framework underlines consciousness as a property that arises and is generated by the recursive interactions of distributed neural assemblies. According to Guevara *et al*. [67] conscious experiences arise from synchronized activities across neural populations, with complexity and coherence serving as critical thresholds for emergent awareness. Principles from quantum biology and classical network theory are integrated to illustrate how consciousness can manifest from subcellular to macroscopic scales.

Complementing this view, DAC theory articulates multi-layered architecture in which consciousness emerges from integration of sensorimotor contingencies, internal state modeling, and contextual planning [61]. In DAC theory, will behavior and self-awareness are not local phenomena but arise from recursive feedback loops that include the brain, body, and environment. This theory supports the idea that consciousness is deeply rooted in the agent's interaction with their environment [34]. The evolution of consciousness occurs through the progressive stages of biological and neurobiological organization. Feinberg and Mallatt [70] argue that in complex systems, phenomenal consciousness is product of the emergence of (weak) standards, which arise from the integration of life processes with specific neurobiological features. Here, there are three hierarchical levels, namely life, nervous system, and neurobiological specialty. Each level contributes to the emergence of subjective experience without invoking non-physical explanatory gaps. This distributed model differs from reductionism. It shows that consciousness is best understood as a multi-scale integrative phenomenon [67]. The convergence of connectionist approaches, adaptive controls, and complex systems provides a strong theoretical foundation for modeling consciousness in biological and artificial systems.

The illusionist theory [47], [71] that conscious experience offers direct and unmediated reflection from the outside world is a view that has historically been criticized by Minsky [11] and refuted by contemporary neuroscience and cognitive science. Consciousness is not a transparent window into reality but rather a constructive simulation, shaped by internal mental models, expectations, and predictive mechanisms [29]. According to illusionist theory, phenomenal consciousness, the subjective quality of experience itself, is an introspective illusion. Individuals systematically tend to judge, erroneously, that they have phenomenal properties such as qualia, even in the absence of a non-physical substrate [67]. Shabasson [71] attempts to explain this illusionist theory as a cognitive mechanism that gives rise to persistent errors of judgment. The user illusion metaphor equates the conscious experience with the experience of a graphical user interface, a symbolic representation simplified to hide the complexity underlying neural processes. According to this view, consciousness serves as a functional abstraction, allowing adaptive behavior without providing access to the full depth of information reality. This metaphor underscores the idea that what we perceive is not the world itself, but rather a curated and optimized simulation for decision-making and survival [69]. Collectively, these perspectives converge on constructivist models of consciousness, in which subjective experience is not a passive acceptance of external stimuli but an active inferential process. The implications of this model go beyond theoretical discourse and inform the design of artificial systems that mimic human-like perception and introspection. By acknowledging the illusory and representational nature of consciousness, researchers can better depict the boundaries between functional awareness and phenomenal experience. The difference between functional awareness and phenomenal experience becomes important for the development of artificial awareness agents.

The main obstacle in artificial consciousness is the problem of subjective experience, i.e., the elusive quality of what it feels like to be a conscious entity. Significant advances in computational modeling have yet to explain whether artificial systems can give rise to subjective consciousness similar to human phenomenology [63]. The absence of biological substrates and neurochemical processes in AI hinders the emergence of native consciousness, which emphasizes the architectural and operational gap between machines and the human brain [70]. Replication of creativity, spontaneity, and emotional nuance, the hallmarks of conscious cognition, still continue to challenge artificial systems. Recent advances in AI/ML, such as neuro-symbolic architecture, transformer-based cognitive modeling, and XAI frameworks, provide computational strategies to simulate synthetic consciousness and epistemic transparency, thus complementing

ontological investigations into consciousness. Neuro-symbolic systems, which integrate symbolic reasoning with deep learning, also promise abstract cognitive function modeling to maintain interpretability [69]. Transformer-based architectures, including models such as pathways language model – embodied (PaLM-E) and Gato, demonstrate emerging capabilities for multimodal reasoning and task generalization, as well as demonstrate scalable substrates for synthetic cognition [61], [63]. Meanwhile, the XAI framework is geared towards making internal decision-making processes transparent and interpretable, in line with philosophical concerns about self-reflection and intentionality in conscious systems [55]. This development reflects the relevance of important computational architectures to overcome obstacles, to be an engineering solution, and to serve as an epistemic instrument that intersects with the ontological conditions of artificial consciousness.

For proponents of synthetic consciousness, functional equality through the simulation of neural dynamics is sufficient for the emergence of subjective states. If AI systems can mimic the recursive and integrative processes that underlie human cognition, they can cross thresholds that indicate conscious experience [64]. This ontological framing of consciousness in artificial systems has significant implications for the real world. As AI agents increasingly participate in autonomous decision-making, questions arise regarding accountability, moral agency, and legal personality thresholds. In addition, how the AI system design awareness attribute is able to interact human-AI, especially in domains that require empathy, trust, and interpretability. The implications of these questions underscore the urgency of integrating ontological insights into ethical and regulatory frameworks for smart technologies. For this, interdisciplinary research is needed, which integrates neuroscience, philosophy of mind, and computational theory to describe the conditions under which artificial systems can achieve consciousness. As the field advances, the question is no longer just whether machines can be aware, but how we can recognize, validate, and engage ethically with those systems. Thus, a synthesis stage was built to examine together the theoretical, empirical, and normative dimensions of artificial consciousness

To further clarify the ontological implications of consciousness in AI systems, a taxonomy of AI architecture based on its representational and reflexive capacity is indispensable. This taxonomy as shown in Figure 3, contains three main categories, namely: i) reactive agents, which operate through stimulus-response mechanisms without internal modeling; ii) agents of deliberation, which have symbolic representational and planning abilities, enable goal-oriented reasoning; and iii) self-modeling systems, which maintain dynamic internal models of their own states and actions, as well as enable meta-representational processing. This level of category reflects an increase in ontological depth, from purely operational responses to potential self-referential awareness, thus offering a scaffolding for evaluating the potential of consciousness across AI systems.

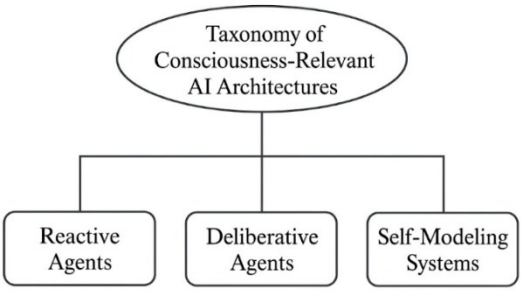

Figure 3. Taxonomy of consciousness-relevant AI architectures

Figure 3 illustrates the classification of AI system gradations based on their representational and reflective capacity. Reactive agents operate through direct stimulus-response mechanisms without internal modeling. Deliberative agents combine symbolic representation and planning ability to allow goal-oriented reasoning. Self-modeling systems maintain dynamic internal models of their own circumstances and actions, thus enabling meta-representational processing. This taxonomy supports the ontological argument that different levels of self-referentiality correspond to different potentials of consciousness in artificial systems. This stratified ontology is suitable for modular architectures, where different subsystems, for example, perceptual, affective, and deliberative modules, can be aligned with the corresponding model layers of consciousness. The ontological priority of the relational and contained dimensions also suggests that the module of ethical reasoning should not be isolated but should be integrated across layers. Adaptive learning mechanisms can be informed by recursive feedback loops in the model, which allow the system to refine

internal representations in response to environmental stimuli and self-referential states. These design principles together offer a roadmap for implementing awareness-inspired architecture in AI.

To apply an ontological framework to evaluative practice, three complementary pathways need to be considered to assess the properties relevant to consciousness in AI systems. First, the simulated environment serves as a realized virtual agent or multi-agent ecosystem, allowing for controlled observation of emergent reflectivity and adaptive behavior. Second, behavioral benchmarks can be designed to test integration, self-modeling, and goal-oriented planning, based on cognitive architecture and standardized task sets. Third, the ethical decision-making test, which includes moral dilemmas and value-sensitive scenarios, offers a perspective into synthetic intentionality and normative reasoning. This modality provides an empirical scaffolding to evaluate the ontological depth of artificial agents and supports the development of systems that are ethically aligned and capable of having consciousness. To support future empirical engagement, validation strategies based on simulations and architectural testing need to be undertaken. The ontological framework can be operationalized through agent-based models that create reflexivity, minimal phenomenality, and deliberate coupling. These models can be evaluated for behavioral coherence, adaptive response, and mapping of internal states. Counterfeiting occurs when predicted ontological markers, such as recursive self-reference or context-sensitive intentionality, fail to appear under controlled conditions. This approach allows the framework to serve as a conceptual guide as well as a testable scaffolding for synthetic consciousness research.

The limitation of this study is the lack of anticipation in dealing with the immediate implications of AI synthetic consciousness for regulatory design, including obligation structures, approval protocols, and oversight mechanisms in high-risk domains such as healthcare, finance, and defense. Future research, therefore, will address these limitations by developing dynamic ontologies, coping with legal implications, to provide a collaborative roadmap for conscious AI studies.

4. CONCLUSION

This study underscores the complexity of artificial consciousness that demands an integrative framework that unites the phenomenological aspects of consciousness with computational paradigms. An interdisciplinary synthesis of cognitive science, neuroscience, and ML is indispensable to developing an ontology that includes both structural and experiential dimensions. For the ontological framework to be applicable, AI researchers need to focus on three domains, namely embedding layered cognitive architectures to support recursive feedback and symbolic-subsymbolic integration, designing behavioral and ethical benchmarks to assess emergent intentionality and reflective capacity, and encouraging interdisciplinary collaboration between philosophers, cognitive scientists, and AI engineers to refine synthetic awareness criteria. Such a research orientation will result in a pragmatic scaffolding for translating ontological depth into system-level design and evaluation. The ontological framing of artificial consciousness should give significant social and ethical weight, as it produces a system capable of demonstrating intentionality and moral reasoning so that it can influence decision-making in domains such as healthcare, law, and education that need to be closely monitored. Regulations should be developed to address agency attribution, accountability, and recognition of rights in synthetic entities. Philosophy, neuroscience, and AI techniques are essential to collaborate to advance the consciousness system. Philosophical analysis provides an ontological foundation and normative criteria; neuroscience contributes empirical models of neural cognition, influences, and work patterns to be applied in AI; and AI engineering translates these insights into computational architectures and validation protocols. This collaborative platform is geared towards fostering mutual refinement and accelerating ethically aligned innovation. Public engagement and ethical literacy are essential to ensure that the development of conscious AI is aligned with the values of humanity and societal sustainability. As AI systems for autonomous decision-making and reflective intentionality evolve, ontological frameworks for legal personality and governance are becoming increasingly urgent, as they offer a basis for evaluating agency, responsibility, and attribution of rights in synthetic entities.

ACKNOWLEDGMENTS

The author gratefully acknowledges to President University for providing access to computing facilities that supported the investigative work.

FUNDING INFORMATION

The authors state that no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andreas Yumarma	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓
Wan Sen	✓	✓	✓	✓		✓	✓	✓		✓	✓	✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

This study did not involve human participants or the use of identifiable personal data that require informed consent.

ETHICAL APPROVAL

This study did not involve human participants or animal subjects that require ethical approval.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [AY], upon reasonable request.

REFERENCES

[1] D. Zhang *et al.*, "Artificial intelligent index report 2021," *Standford University: Human-Centered Artificial Intelligence,* 2021.
[2] S. Makridakis, "The forthcoming artificial intelligence (AI) revolution: its impact on society and firms," *Futures*, vol. 90, pp. 46–60, 2017, doi: 10.1016/j.futures.2017.03.006.
[3] N. Alonso, "An introduction to the problems of AI consciousness," *The Gradient*. 2023. Accessed: Dec. 28, 2023. [Online]. Available: https://thegradient.pub/an-introduction-to-the-problems-of-ai-consciousness/
[4] S. H. Hawley, "Challenges for an ontology of artificial intelligence," *Perspectives on Science and Christian Faith*, vol. 7, no. 2, pp. 83-93.
[5] A. Bertolini and F. Episcopo, "Robots and AI as legal subjects? disentangling the ontological and functional perspective," *Frontiers in Robotics and AI*, vol. 9, pp. 1–15, 2022, doi: 10.3389/frobt.2022.842213.
[6] E. Hildt, "Artificial intelligence: does consciousness matter?," *Frontiers in Psychology*, vol. 10, pp. 1–6, 2019, doi: 10.3389/fpsyg.2019.01535.
[7] U. Pagallo, "Algo-rhythms and the beat of the legal drum," *Philosophy and Technology*, vol. 31, no. 4, pp. 507–524, 2018, doi: 10.1007/s13347-017-0277-z.
[8] S. Ding, Gallagher, Y. E. Rhee, and F. Yu, "Interactions between humans, non-human animals, and AI: philosophical underpinnings and ethical considerations," *Topoi, Special Issue*, 2025.
[9] D. C. Dennett, "Facing up to the hard question of consciousness," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 373, no. 1755, 2018, doi: 10.1098/rstb.2017.0342.
[10] D. McDermott, "Artificial intelligence and consciousness," in *The Cambridge Handbook of Consciousness*. Cambridge, United Kingdom: Cambridge University Press, 2007, doi: 10.1017/CBO9780511816789.007.
[11] M. Minsky, "Consciousness," *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. New York, United States: Simon and Schuster, 2006.
[12] K. Mogi, "Artificial intelligence, human cognition, and conscious supremacy," *Frontiers in Psychology*, vol. 15, 2024, doi: 10.3389/fpsyg.2024.1364714.
[13] A. Alavi, H. Akhoundi, and F. Kouchmeshki, "Analyzing advanced AI systems against definitions of life and consciousness," 2025, *arXiv:2502.05007*.
[14] D. Yu, "Towards AI consciousness: a phased approach to functional organization, embodiment, and continuous learning," *Intelligent Computing (CompCom)*, Cham, Switzerland: Springer, 2025, pp. 1-21, doi: 10.1007/978-3-031-92605-1_1.
[15] A. Chella, "Artificial consciousness: the missing ingredient for ethical AI?," *Frontiers in Robotics and AI*, vol. 10, pp. 1–5, 2023, doi: 10.3389/frobt.2023.1270460.
[16] S. S. Rakover, "AI and consciousness," *AI and Society*, vol. 39, no. 4, pp. 2139–2140, 2024, doi: 10.1007/s00146-023-01663-8.
[17] P. Butlin *et al*., "Consciousness in artificial intelligence: insights from the science of consciousness," 2023, *arXiv:2308.08708*.
[18] F. Heylighen and S. Beigi, "The local prospect theory of subjective experience: a soft solution to the hard problem of consciousness," *Journal of Consciousness Studies*, vol. 31, no. 11, pp. 122–152, 2024, doi: 10.53765/20512201.31.11.122.
[19] M. Schmalzried, "A philosophical and ontological perspective on artificial general intelligence and the metaverse," *Journal of Metaverse*, vol. 5, no. 2, pp. 168–180, 2025, doi: 10.57019/jmv.1668494.

[20] S. K.-Nye, “Algorithmic AI consciousness,” *Philosophy of Science Archive*, 2024.
[21] C. Turner and S. Schneider, “Could you merge with AI? reflections on the singularity and radical brain enhancement,” in *The Oxford Handbook of Ethics of AI*. New York, United States: Oxford Academic, pp. 307–324, 2020.
[22] R. V. Yampolskiy, “On the differences between human and machine intelligence,” *CEUR Workshop Proceedings*, vol. 2916. 2021.
[23] L. Floridi *et al.*, “AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations,” *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.
[24] D. J. Gunkel, A. Gerdes, and M. Coeckelbergh, “Editorial: should robots have standing? the moral and legal status of social robots,” *Frontiers in Robotics and AI*, vol. 9, pp. 1–4, 2022, doi: 10.3389/frobt.2022.946529.
[25] R. Krzanowski and P. Polak, “The internet as an epistemic agent (EA),” *Informacios Tarsadalom*, vol. 22, no. 2, pp. 39–56, 2022, doi: 10.22503/inftars.XXII.2022.2.3.
[26] B. J. Copeland, "Artificial intelligence," *Encyclopædia Britannica*. Accessed: Dec. 28, 2023. [Online]. Available: https://www.britannica.com/technology/artificial-intelligence
[27] R. Mizoguchi and G. H. Canada, “Using ontological engineering to overcome common AI-ED problems,” *Journal of Artificial Intelligence and Education*, vol. 11, pp. 107–121, 2000.
[28] R. Kishore, H. Zhang, and R. Ramesh, “A helix-spindle model for ontological engineering,” *Communications of the ACM*, vol. 47, no. 2, pp. 69–75, 2004, doi: 10.1145/966389.966393.
[29] W. Jaworski, “Hylomorphism and the construct of consciousness,” *Topoi*, vol. 39, no. 5, pp. 1125–1139, 2020, doi: 10.1007/s11245-018-9610-0.
[30] C. Percy and A. P.-Hinojosa. 2025. “Ontological diversity in theoretical physics and its significance for consciousness research,” *Journal of Consciousness Studies,* vol. 32, no. 9, pp. 183-214, doi: 10.53765/20512201.32.9.183.
[31] D. S. Blundo, F. S.-Toscano, M. G. Pasca, A. Scozzari, and G. Arcese. 2026. “The transformation of technological rationality: from deductive control to abductive intelligence,” *Philosophies*, vol. 11, no. 3, doi: 10.3390/philosophies11030068.
[32] M. Neumann and V. Dirksen, “Symbol grounding for generative AI: lessons learned from interpretive ABM,” *Frontiers in Computer Science*, vol. 7, 2025, doi: 10.3389/fcomp.2025.1508004.
[33] M. Barbieri, “The code model of semiosis,” *The American Journal of Semiotics*, vol. 24, no. 1, pp. 23–37, 2008, doi: 10.5840/ajs2008241/33.
[34] Y. Ye, X. Cong, S. Tian, Y. Qin, C. Liu, Y. Lin, Z. Liu, M. Sun, “Rational decision-making agent with internalized utility judgment," *International Conference on Learning Representations (ICLR)*, 2025.
[35] Y. Zhang, J. Wu, F. Yu, and L. Xu, “Moral judgments of human vs. AI agents in moral dilemmas,” *Behavioral Sciences*, vol. 13, no. 2, 2023, doi: 10.3390/bs13020181.
[36] J. F. Bonnefon, I. Rahwan, and A. Shariff, “The moral psychology of artificial intelligence,” *Annual Review of Psychology*, vol. 75, no. 1, pp. 653–675, 2024, doi: 10.1146/annurev-psych-030123-113559.
[37] J. Wu, J. Zhu, and Y. Liu, M. Xiu, Y. Jin, “Agentic reasoning: reasoning LLMs with tools for deep research, ” *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025, 28489-28503.
[38] A. Entwistle, “What is artificial intelligence?,” *Engineering materials and design*, vol. 32, no. 3. 1988, doi: 10.55248/gengpi.2022.31261.
[39] Q. Hu, Y. Lu, Z. Pan, Y. Gong, and Z. Yang, “Can AI artifacts influence human cognition? The effects of artificial autonomy in intelligent personal assistants,” *International Journal of Information Management*, vol. 56, 2021, doi: 10.1016/j.ijinfomgt.2020.102250.
[40] H. J. Park, “Being self in heidegger’s ontology: a heideggerian contribution to the ontology of individuality,” *Kritike*, vol. 16, no. 3, pp. 5–20, 2023, doi: 10.25138/16.3.a1.
[41] I. Berent, “The ‘hard problem of consciousness’ arises from human psychology,” *Open Mind*, vol. 7, pp. 564–587, 2023, doi: 10.1162/opmi_a_00094.
[42] U. Peters, “Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque,” *AI and Ethics*, vol. 3, no. 3, pp. 963–974, 2023, doi: 10.1007/s43681-022-00217-w.
[43] L. Blum and M. Blum, “A theory of consciousness from a theoretical computer science perspective: insights from the conscious turing machine,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 21, May 2022, doi: 10.1073/pnas.2115934119.
[44] D. C. Youvan, “The propagation of consciousness: why AI may succeed where humans cannot,” *Research Gate*, 2025, doi: 10.13140/RG.2.2.26056.84483.
[45] M. Kiškis, “Legal framework for the coexistence of humans and conscious AI,” *Frontiers in Artificial Intelligence*, vol. 6, 2023, doi: 10.3389/frai.2023.1205465.
[46] S. Namestiuk, “Challenging the boundary between self-awareness and self-consciousness in AI from the perspectives of philosophy,” *Futurity Philosophy*, vol. 2, no. 4, pp. 43–60, 2023, doi: 10.57125/fp.2023.12.30.03.
[47] M. S. A. Graziano, A. Guterstam, B. J. Bio, and A. I. Wilterson, “Toward a standard model of consciousness: reconciling the attention schema, global workspace, higher-order thought, and illusionist theories,” *Cognitive Neuropsychology*, vol. 37, no. 3–4, pp. 155–172, May 2020, doi: 10.1080/02643294.2019.1670630.
[48] J. Qiu, L. Cheng, and J. Huang, “Charting the landscape of artificial intelligence ethics: a bibliometric analysis,” *International Journal of Digital Law and Governance*, vol. 2, no. 1, pp. 135–167, 2025, doi: 10.1515/ijdlg-2025-0007.
[49] P. Butlin and T. Lappas, “Principles for responsible AI consciousness research,” *Journal of Artificial Intelligence Research*, vol. 82, pp. 1673–1690, 2025, doi: 10.1613/jair.1.17310.
[50] G. Kutyniok, “How can reliability of artificial intelligence be ensured?,” *Harvard Data Science Review*, vol. 7, no. 2, pp. 1–8, 2025, doi: 10.1162/99608f92.9fc78e4c.
[51] E. Dahlin, “Trust in AI,” *AI and Society*, vol. 40, no. 8, pp. 6089–6095, 2025, doi: 10.1007/s00146-025-02429-0.
[52] H. I. Aysel, X. Cai, and A. P.-Bennett, “Explainable artificial intelligence: advancements and limitations,” *Applied Sciences*, vol. 15, no. 13, 2025, doi: 10.3390/app15137261.
[53] M. Debnath, “Mathematical foundations of explainable AI: a framework based on topological data analysis,” *Communications on Applied Nonlinear Analysis*, vol. 32, no. 9s, pp. 3119–3142, 2025, doi: 10.52783/cana.v32.4650.
[54] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): toward medical XAI,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.
[55] F. Herrera, “Making sense of the unsensible: reflection, survey, and challenges for XAI in large language models toward human-centered AI,” 2025, *arXiv:2505.20305*.
[56] A. S. Fokas, “Can artificial intelligence reach human thought?,” *PNAS Nexus*, vol. 2, no. 12, pp. 1–5, 2023, doi: 10.1093/pnasnexus/pgad409.
[57] B. Finzel, “Toward trustworthy AI with integrative explainable AI frameworks,” *IT - Information Technology*, vol. 67, no. 1, pp. 20–45, 2025, doi: 10.1515/itit-2025-0007.

[58] S. Baron, "Trust, explainability and AI," *Philosophy and Technology*, vol. 38, no. 1, 2025, doi: 10.1007/s13347-024-00837-6.
[59] R. Redaelli, "Intentionality gap and preter-intentionality in generative artificial intelligence," *AI and Society*, vol. 40, no. 4, pp. 2525–2532, 2025, doi: 10.1007/s00146-024-02007-w.
[60] W. Wiese, "Artificial consciousness: a perspective from the free energy principle," *Philosophical Studies*, vol. 181, no. 8, pp. 1947–1970, 2024, doi: 10.1007/s11098-024-02182-y.
[61] L. Dung and L. Kersten, "Implementing artificial consciousness," *Mind and Language*, vol. 40, no. 3, pp. 285–305, 2025, doi: 10.1111/mila.12532.
[62] D. Papineau, "Theories of consciousness," *Consciousness: New Philosophical Perspectives*, vol. 23, pp. 353–383, 2023, doi: 10.1093/oso/9780199241286.003.0013.
[63] T. Metzinger, "Artificial suffering: an argument for a global moratorium on synthetic phenomenology," *Journal of Artificial Intelligence and Consciousness*, vol. 8, no. 1, pp. 43–66, 2021, doi: 10.1142/S270507852150003X.
[64] A. K. Seth, "Conscious artificial intelligence and biological naturalism," *Behavioral and Brain Sciences*, pp. 1–42, 2025, doi: 10.1017/S0140525X25000032.
[65] P. B. Badcock, K. J. Friston, and M. J. D. Ramstead, "The hierarchically mechanistic mind: a free-energy formulation of the human psyche," *Physics of Life Reviews*, vol. 31, pp. 104–121, 2019, doi: 10.1016/j.plrev.2018.10.002.
[66] T. Parr, "Patterns and particles. Comment on 'The Markov blanket trick: On the scope of the free energy principle and active inference'," *Physics of Life Reviews*, vol. 43, pp. 17–19, 2022, doi: 10.1016/j.plrev.2022.08.001.
[67] R. Guevara, D. M. Mateos, and J. L. P. Velázquez, "Consciousness as an emergent phenomenon: a tale of different levels of description," *Entropy*, vol. 22, no. 9, 2020, doi: 10.3390/E22090921.
[68] P. F. M. J. Verschure, "Synthetic consciousness: the distributed adaptive control perspective," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 371, 2016, doi: 10.1098/rstb.2015.0448.
[69] D. S.-Pata *et al.*, "Epistemic autonomy: self-supervised learning in the mammalian hippocampus," *Trends in Cognitive Sciences*, vol. 25, no. 7, pp. 582–595, 2021, doi: 10.1016/j.tics.2021.03.016.
[70] T. E. Feinberg and J. Mallatt, "Phenomenal consciousness and emergence: eliminating the explanatory gap," *Frontiers in Psychology*, vol. 11, pp. 1–15, 2020, doi: 10.3389/fpsyg.2020.01041.
[71] D. Shabasson, "Illusionism about phenomenal consciousness: explaining the illusion," *Review of Philosophy and Psychology*, vol. 13, no. 2, pp. 427–453, 2022, doi: 10.1007/s13164-021-00537-6.
[72] T. Niikawa, "Illusionism and definitions of phenomenal consciousness," *Philosophical Studies*, vol. 178, no. 1, pp. 1–21, 2021, doi: 10.1007/s11098-020-01418-x.
[73] K. Frankish, "Consciousness, attention, and response," *Cognitive Neuropsychology*, vol. 37, no. 3–4, pp. 202–205, May 2020, doi: 10.1080/02643294.2020.1729111.
[74] S. A. Alavi and M. Noel, "Effect of model architecture and input parameters to improve performance of artificial intelligence models for estimating concrete strength using SonReb," *Engineering Structures*, vol. 323, 2025, doi: 10.1016/j.engstruct.2024.119285.
[75] F. Kammerer, "Defining consciousness and denying its existence. Sailing between charybdis and scylla," *Philosophical Studies*, vol. 182, no. 2, pp. 541–565, 2025, doi: 10.1007/s11098-025-02285-0.
[76] G. Hoffman and X. Zhao, "A primer for conducting experiments in human-robot interaction," *ACM Transactions on Human-Robot Interaction*, vol. 10, no. 1, pp. 1–31, 2020, doi: 10.1145/3412374.
[77] S. Bringsjord, M. Giancola, and N. S. Govindarajulu, "Culturally aware social robots that carry humans inside them, protected by defeasible argumentation systems," *Frontiers in Artificial Intelligence and Applications*, vol. 335, pp. 440–456, 2020, doi: 10.3233/FAIA200941.
[78] S. Shrivastav, "Exploring the debate on AI consciousness: can artificial intelligence systems achieve consciousness?" *Medium*, 2023. Accessed: Dec. 28, 2023. [Online]. Available: https://shivang-ahd.medium.com/exploring-the-debate-on-ai-consciousness-can-artificial-intelligence-systems-achieve-consciousness-8b65daf2e62

BIOGRAPHIES OF AUTHORS

Andreas Yumarma obtained a doctorate in philosophy from Pontifical Gregoriana University, Rome, Italy, in 1996. He is an associate professor in Department of Business Administration at the Faculty of Business, President University. His research interests include business philosophy and ethics, way of human existence in the digital era, philosophy and technology, educational and health philosophy, philosophy of history, and eastern philosophy. He can be contacted at email: andreasyumarma@president.ac.id.

Tjong Wan Sen learned his Ph.D. from Institut Teknologi Bandung in 2009. Since 2010, he has been affiliated with the Master of Informatics program within the Department of Computer Science, Faculty of Computing at President University, Jababeka, Indonesia, where he currently serves as a lecturer and researcher. His research endeavors encompass generative adversarial networks, low-powered AI, computer vision, internet of things, and embedded systems. He can be contacted at email: wansen@president.ac.id.