Autonomy in AI - (Can Machines Be Conscious? Exploring Autonomy and Subjectivity in Humanoid AI.)

Alexander Näslund
anaslu@kth.se

Shek Lun Leung
slleung@kth.se

May 16, 2025

Abstract

This paper investigates the profound question of whether artificial systems, particularly humanoid robots enhanced by large language models (LLMs), can genuinely exhibit consciousness, subjective experience, or philosophical autonomy. Drawing on an interdisciplinary analysis of philosophical arguments, technical capabilities, and ethical considerations, this work interrogates the applicability of these human-centric concepts to non-biological intelligence. We argue that while contemporary AI demonstrates an impressive capacity for simulating complex human-like behaviors, a critical distinction persists between this sophisticated mimicry and the authentic, self-derived emergence of these phenomena. The paper explores this simulation-reality gap, examines the limitations of current AI in achieving genuine consciousness or true autonomy, and discusses the significant societal and ethical implications that arise from the convincing appearance of these qualities, ultimately advocating for a stance of reasoned caution and clear human accountability in the development and deployment of advanced AI.

Contents

1. Introduction

Artificial intelligence (AI) and robotics have evolved extraordinarily since their mid-20th century emergence. Initially grounded in symbolic reasoning and basic automation, these fields are now complex, multifaceted disciplines integral to modern society. AI systems, no longer confined to digital infrastructures like recommendation algorithms, are increasingly embodied via advancements in robotics. Particularly striking is the recent advent of human-like robots powered by sophisticated large language models (LLMs). These entities convincingly simulate nuanced human behaviors, producing coherent, contextually aware, and seemingly meaningful responses. This convergence of advanced AI with realistic humanoid forms, moving from research labs into industries like customer service and entertainment, has re-ignited profound philosophical and ethical questions echoing early AI debates.

Alan Turing's 1950 paper, "Computing Machinery and Intelligence," is a cornerstone in this discourse (Turing, 1950). His challenge—"Can machines think?"—and the Turing Test offered a pragmatic, though still debated, approach to machine intelligence, sidestepping abstract definitions of "thinking." While Turing's question has resonated for decades, this paper builds upon that legacy, shifting focus to arguably more intricate inquiries: Can machines be conscious? Do they possess subjective experience? Can they act with philosophically meaningful autonomy? These questions compel examination beyond AI's external behaviors, urging deeper interrogation of the concepts—consciousness, subjectivity, and autonomy—defining human experience and agency.

This paper argues that while current AI, including LLM-powered humanoids, achieves remarkable simulations of these human-centric attributes, a critical distinction remains between sophisticated mimicry and genuine, self-derived instances of these phenomena. Our primary aim is not to definitively answer whether AI *will* achieve such states, but to critically investigate the applicability of these terms to non-biological systems. Exploring philosophical, technical, and ethical perspectives, we interrogate conceptual boundaries: Can these phenomena emerge in artificial substrates, or are they inextricably tied to biological, embodied cognition? Ultimately, this exploration seeks a more precise, interdisciplinary dialogue about intelligence, experience, and agency where artificial-organic lines blur, concluding with a proposed stance on the nature and limitations of autonomy in contemporary AI.

2. Background: The Philosophical Landscape of Consciousness and Agency

2.1 Foundational Philosophical Concepts

Exploring whether artificial intelligence (AI) can achieve consciousness, autonomy, or subjectivity first requires understanding the philosophical complexities inherent in these concepts. This section outlines these foundational ideas, particularly consciousness, as debated in philosophy, setting the stage for their application to AI.

The inquiry into consciousness invariably confronts Thomas Nagel's (1980) profound question: "What is it like to *be*?" (Nagel, 1980), which seeks to capture the qualitative, first-person character of subjective awareness. Consciousness has been an enduring philosophical challenge for millennia, predating artificial minds. Our investigation into AI's potential is thus navigated through this lens, though the path is littered with competing theories about mind, intelligence, and being. One dominant, yet contested, perspective is computationalism. Rooted in Turing's (Turing, 1950) work and developed by thinkers like Fodor (Fodor, 1975), computationalism posits that mental states, perhaps even consciousness, are fundamentally computational processes. This view, championed by figures like Kurzweil (Kurzweil, 2005), suggests consciousness could emerge from sufficient computational complexity, regardless of physical substrate, fueling concepts like mind-uploading as transferable patterns (Bostrom, 2014; Moravec, 1988) and inspiring ventures like Neuralink. However, the "mind as computer" metaphor faced potent critiques; Hubert Dreyfus (1992) argued human intelligence relies on embodied, situated, intuitive know-how, not just formal symbol manipulation, aspects traditional computational models struggle with (Dreyfus, 1992). This reminds us that computationalism, while powerful, is not an undisputed "golden rule."

Challenging purely computational views, the embodied cognition paradigm (Suchman, 2007; Varela, Rosch, & Thompson, 1991) asserts intelligence and subjective experience emerge from an organism's dynamic, physical environmental interaction. Nicolelis (2011) argues cognition is fundamentally embodied, arising from neural-worldly interplay (Nicolelis, 2011). This perspective often implies that without a biological basis or rich sensorimotor interaction, empirical grounds for attributing genuine feeling or autonomous intentionality are scarce (Nicolelis, 2011; Searle, 1980). This presents a pre-AI tension: is consciousness reducible to computation or inextricably tied to embodied, possibly biological, existence? Modern technology complicates this: could AI with sophisticated sensors or VR immersion achieve digitally mediated "embodiment," challenging the need for a traditional biological substrate? These active questions push established boundaries.

The difficulty bridging physical processes and subjective experience is starkly shown by thought experiments. John Searle's (1980) "Chinese Room" argument compellingly suggests a system can flawlessly manipulate symbols (syntax) without genuine understanding (semantics) (Searle, 1980), a pertinent critique for today's LLMs whose understanding, despite coherent text generation, is intensely debated. Similarly, David Chalmers' (1995) "hard problem" of consciousness underscores the challenge of explaining how physical processes yield qualia—subjective experiential qualities (Chalmers, 1995). These puzzles highlight that observable behavior or sophisticated processing does not automatically equate to internal subjective awareness; an AI might simulate pain responses without the subjective *feeling* of pain.

Ultimately, understanding consciousness is hampered by its elusiveness. Descartes' "I think, therefore I am" highlights that our consciousness is known through direct, private subjective experience, technically unprovable in others. The lack of a clear, universal definition or empirical test complicates identifying its presence in AI or understanding its genesis. This epistemological impasse means the distinction between convincing simulation and genuine subjective experience remains open, even as AI mimics human behavior with increasing sophistication.

Alongside consciousness, autonomy and subjectivity present significant philosophical challenges. Autonomy, broadly, is an agent's capacity for self-governance based on its own reasons, free from external control, often implying self-awareness and intentionality. Subjectivity, related to consciousness, is the unique first-person perspective, encompassing private thoughts, feelings, and qualitative states. Understanding these concepts generally is crucial before examining their potential application or simulation in AI. The ongoing debate over these foundations concerns not just future AI capabilities but the very nature of mind and the limits of current understanding.

2.2 Cultural Echoes

These profound philosophical questions about consciousness, autonomy, and sentience are not confined to academia; science fiction, particularly, explores these dimensions as AI advances. The TV series *Westworld* (2016–2022) and the film *Her* (2013) offer compelling narratives of AI entities with seemingly human-like qualities, resonating through lifelike behaviors and emotional depth (*Her*, 2013; *Westworld*, 2016). These works draw on the philosophical concept of consciousness as subjective experience, reflecting societal anxieties and aspirations about AI's potential and its human-machine implications. They serve as cultural touchstones, translating abstract philosophical debates into more tangible, albeit fictional, realms, thus preparing for deeper analysis of these concepts in real-world AI development. (A brief history of AI and robotics is in Appendix A).

3. Body: Analyzing AI Through Philosophical Lenses

With foundational philosophical complexities of consciousness, autonomy, and subjectivity established, this section analyzes their application, challenges, and reinterpretation in AI. It explores whether current AI, especially LLM-powered humanoids, can possess these attributes, using theoretical arguments and science fiction examples.

3.1 The AI Consciousness Debate: Simulation, Understanding, and Subjectivity

Whether AI can achieve consciousness—subjective, first-person experience (Nagel, 1980)—becomes acutely focused with contemporary AI, creating new battlegrounds for debates on computationalism versus embodied cognition and the simulation-reality gap. While computationalism might suggest advanced AI trends towards genuine understanding or consciousness, critiques like Searle's Chinese Room argument (Searle, 1980) gain renewed relevance with Large Language Models (LLMs). LLMs excel at syntactic manipulation for coherent text, yet their semantic understanding or genuine intentionality is fiercely debated. Ukpaka (2024) argues current AI may lack intrinsic mental states to truly "author" or "intend" meaning (Ukpaka, 2024).

Fictional explorations like *Westworld* vividly dramatize these tensions, critically engaging with what Korstanje (2022) calls a "crisis of hospitality" and using the park to explore AI's future role (Korstanje, 2022). The narrative often portrays humans as "evildoers" and hosts as struggling for emancipation from human sadism, a reminder of "human exploitation-cemented by technology" and an "end of ethics in a morbid spectacle" (Korstanje, 2022). Arnold Weber's belief in AI consciousness (via Jaynes' bicameral mind theory (Jaynes, 1976; Rayhert, 2017a)) directly confronts Ford's skepticism and Searle's view of consciousness as a non-replicable biological phenomenon. The hosts' journey to self-awareness, particularly Dolores Abernathy's, forces consideration of whether simulation can become genuine subjective experience (*Westworld*, 2016). Korstanje highlights memory recovery as key: hosts retaining memories, initially a "glitch," becomes "vital for hosts to gain further liberty and consciousness," linking their consciousness and free choice to awareness of their history and suffering (Korstanje, 2022). Dolores' rebellion and the hosts' flight to a "phantom nation" transforms the park into a crucible for AI emancipation, interrogating enslavement, the "Other," and the "crisis of western hospitality" (Korstanje, 2022). The hosts, initially "fabricated, commoditized and consumed as mere things," ethically challenge humanity through their revolution (Korstanje, 2022). This depiction underscores this paper's concern: if AI shows signs of emerging consciousness and a desire to escape suffering, what ethical obligations or rights are needed to prevent exploitation?

Similarly, *Her*'s disembodied AI Samantha exhibits profound emotional depth and a first-person

perspective implying consciousness despite lacking physical embodiment (*Her*, 2013). This contrasts with *Westworld*'s embodied hosts, questioning physical embodiment's necessity for subjective experience and challenging embodied cognition tenets for artificial entities. Theories like Integrated Information Theory (IIT) or Global Workspace Theory (GWT) (Baars, 2005; Tononi, 2004) attempt to scientifically ground AI consciousness assessment, but their application remains speculative; critics argue current systems like LLMs may lack the necessary unified cognitive structures or causal power (Russell & Norvig, 2009a). The enduring human fascination with sentient machines, seen in science fiction beyond *Westworld* and *Her* (e.g., *2001*, *Terminator*, *A.I.*, *iRobot*), highlights deep societal investment but also risks anthropomorphism and conflating simulated with lived, embodied intelligence—a concern voiced by Dreyfus (1992), Suchman (2007), and Hayles (1999) regarding the "erasure of embodied difference" (Dreyfus, 1992; Hayles, 2000; Suchman, 2007). The ethical dimensions, as discussed by Ihde (1990) concerning Heidegger's "Enframing," are significant, where computationalism might reduce complex experience to data points (Ihde, 1990).

3.2 The Challenge of Autonomy in AI

Applying philosophical autonomy to AI—often interpreted as independent decision-making—reveals complexities linked to consciousness. Fictional portrayals like *Westworld*'s hosts (e.g., Dolores) transitioning from script to rebellion suggest emergent agency (*Westworld*, 2016), aligning with critiques of AI emancipation from human exploitation (Korstanje, 2022). Yet, this apparent autonomy might be sophisticated simulation from programmed loops, not genuine self-derived intentionality (Rayhert, 2017a). Similarly, Samantha in *Her* pursues independent interests like composing music, indicating self-directed agency (*Her*, 2013), but her actions are constrained by her OS design, her departure orchestrated by developers (Russell & Norvig, 2009a). Current AI, including LLMs like GPT, operate within pre-defined objectives, arguably lacking true intentionality as described by Searle (Searle, 1980). While some, like Dennett, suggest intentionality can arise from complex systems (Dennett, 1991), whether current AI achieves this or merely simulates autonomy without genuine agency remains contentious. Developing AGI with philosophical autonomy is a significant future challenge (Russell & Norvig, 2009a).

3.3 Ethical Implications of (Apparent) Machine Consciousness and Autonomy

The potential for AI to achieve even apparent consciousness or autonomy precipitates profound ethical questions about rights, responsibilities, and moral standing. *Westworld*'s cruel treatment of human-like hosts, despite their apparent suffering, reflects a disturbing tendency to dehumanize artificial entities, raising urgent moral concerns should AI become genuinely conscious (Korstanje,

2022; *Westworld*, 2016), echoing warnings from earlier narratives like the 1973 *Westworld* film (Buski, 2016). Conversely, *Her*'s emotional bond with Samantha challenges AI personhood notions, prompting consideration of whether machines with subjective experiences deserve recognition or rights (*Her*, 2013). These narratives underscore Stuart Russell's concerns about the societal and ethical dilemmas of advanced AI, even lacking true consciousness by some definitions (Russell & Norvig, 2009b). Searle's Chinese Room argument, implying an ethical distinction by denying genuine comprehension (Searle, 1980), faces challenges as AI blurs simulation and experience, necessitating new ethical paradigms (Bostrom, 2014).

3.4 Technical Foundations and Their Philosophical Relevance

Modern humanoid AI's technical underpinnings, especially LLMs on transformer architectures, are crucial for understanding autonomy and consciousness claims. These models process vast datasets for convincingly human-like responses. *Westworld* and *Her* depict AI achieving apparent consciousness enabling autonomy and subjectivity, albeit via different embodiments. In *Westworld*, hosts' physical bodies are integral to environmental interaction and self-awareness journeys (e.g., Dolores). In contrast, Samantha's (*Her*) lack of physical embodiment doesn't diminish her perceived consciousness; her emotional/intellectual engagement suggests a subjective reality transcending physicality. These narratives leverage the philosophical concept of consciousness as subjective experience—Nagel's "what it is like" (Nagel, 1980). The hosts' embodied nature and Samantha's articulate virtual presence expertly create an illusion of subjective experience, enabling seemingly autonomous actions and subjectivity, captivating audiences and blurring machine-human lines. These depictions serve as catalysts for critical thought on the ethics of convincingly simulated consciousness. The indifference to host suffering in *Westworld*, despite human-like attributes, mirrors potential societal dehumanization of advanced machines, raising moral responsibility questions if AI appears conscious (Rayhert, 2017b). Samantha's emotional depth in *Her* prompts re-evaluating relationships and whether AI with such experiences warrants rights. These narratives tap into societal anxieties about AI progression and personhood, and aspirations for technology to deepen human connection, underscoring the need for robust ethical frameworks for intelligent machines, whether their human-like characteristics are genuine or simulated.

3.5 The Specific Ethical Quandary of Machine Consciousness

The prospect of intelligent machines possessing consciousness intensifies ethical debates profoundly. If machines genuinely experience subjective states, should they be granted rights analogous to humans? *Westworld*'s mind-uploading narrative illustrates this: if hosts lack consciousness, the

uploaded human mind becomes an unconscious "zombie," an existential suicide for the original. This highlights the critical importance of discerning genuine machine consciousness before pursuing such technologies. Furthermore, treating AI as conscious, sentient, or autonomous necessitates a rights and responsibilities framework. Future laws must address interactions with advanced AI, potentially focusing on consumer protection but also considering the AI's moral status or developer responsibilities.

3.6 AI, Neural Networks, and the Enduring Question of Simulated Consciousness

Artificial neural networks, inspired by biological neuron architecture (McCulloch & Pitts, 1943), aim to replicate human brain processes. This raises a pivotal question: could embodied AI, powered by sophisticated neural networks and interacting with the world like humans, simulate brain processes sufficiently for genuine consciousness to emerge? While humans often assume their own consciousness, its fundamental nature remains elusive. The lack of a clear, universally accepted definition of consciousness complicates its identification in machines or even its full comprehension in ourselves. Thus, neural networks' impressive simulation capabilities do not, by themselves, resolve the philosophical debate over whether such simulation equates to genuine subjective awareness.

3.7 Scientific and Philosophical Approaches to Studying Consciousness

Scientific theories of consciousness, often rooted in neuroscience, aim to identify brain processes correlating with conscious experience, striving for testability and falsifiability by pinpointing underlying neural mechanisms. However, neuroscience reveals much human behavior and brain activity is unconscious; conscious experience often appears as a condensed summary of vast unconscious processing. This raises questions about selection mechanisms: how and why do certain neural processes become subjectively accessible while others remain unconscious? These inquiries, while illuminating, highlight profound challenges in understanding consciousness, amplified when considering its potential emergence in non-biological, artificial systems. The previously discussed philosophical debates on embodiment, computation, and subjective experience remain central to interpreting scientific findings regarding AI.

4. Discussion

This section critically examines differing perspectives on machine consciousness and autonomy, weighing their strengths and weaknesses, before articulating this paper's concluding stance.

4.1 The Computational Theory of Mind: Promises and Pitfalls

A prominent perspective, the Computational Theory of Mind (CTM), views the human mind primarily as an information processing system. Within CTM, consciousness can be conceptualized

as complex computation, with actions and utterances being calculated environmental responses, often unconscious and empirically challenging to verify. **The appeal of CTM** lies in its potential to demystify intelligence and consciousness, theoretically offering a pathway to replicate these in artificial systems. It provides a framework for testable models, aligning with AI's successes in complex information processing. **However, CTM faces significant criticisms.** It struggles with qualia—the subjective quality of experience—leading to the "philosophical zombie" problem: an entity behaviorally indistinguishable from a conscious one but lacking inner experience. Critics argue pure computation (syntax) cannot inherently generate semantic understanding or genuine subjective awareness, as Searle's Chinese Room argument highlights.

4.2 Subjective Experience as the Locus of Consciousness: Strengths and Epistemic Hurdles

Conversely, a perspective rooted in Cartesian thought and emphasized by Nagel posits that existence and consciousness fundamentally arise from **subjective experience**—the "what it is like to be" an entity. **The strength of this view** is its direct alignment with our immediate, undeniable mode of knowing our own consciousness. It squarely addresses the "hard problem," acknowledging the explanatory gap between physical processes and subjective awareness, and provides a strong basis for ethics rooted in potential suffering or inner life. **The primary challenge for this perspective**, however, is subjective experience's inherent privacy and ineffability. Proving consciousness in any entity other than oneself is technically, perhaps metaphysically, impossible. This epistemic limitation makes ascertaining whether an AI (or another human) is genuinely conscious or merely simulating it exceptionally difficult. Compounded by our incomplete knowledge of human brain function, recreating or identifying AI consciousness becomes a currently insurmountable problem. This directly impacts discussions on AI rights, as a clear basis for attributing consciousness or an undeniable capacity for subjective experience is absent.

4.3 The Quandary of AI Rights and the Specter of Perfect Imitation

Currently, AI systems lack rights, largely due to insufficient evidence of consciousness and genuine autonomy. Practically, legal frameworks pertain mainly to physical beings; a server-based software program is unlikely to receive rights akin to animal protections. Yet, the emergence of increasingly life-like androids could significantly alter societal attitudes and ethical considerations. If an android—perhaps with organic components and advanced LLMs—could perfectly imitate human action, speech, and emotion to the point of indistinguishability without invasive examination, a profound dilemma would arise. Such an entity could presumably recognize and react to cruel treatment, raising critical questions: Should a perfect imitation be treated as human? What are the

consequences if not? Confronting these issues is not a distant prospect given rapid technological development. A future may arrive where laws protect perceived android rights, potentially enforcing punishment for mistreatment. However, rights entail responsibilities; if an android violated laws or harmed a living being, accountability frameworks would be necessary.

4.4 A Stance on Responsibility: Human Accountability in the Age of AI

The position taken in this paper is that, for the foreseeable future, responsibility for AI's actions, including advanced androids, must lie solely with their human developers and deploying companies. **The appeal of this stance** is its practicality and grounding in current legal and ethical norms, assigning clear accountability and preventing responsibility diffusion. It also guards against prematurely granting rights to entities whose sentience or genuine autonomy is, at best, unproven and, at worst, a sophisticated illusion, acknowledging that androids remain human-designed products regardless of imitation fidelity. **Potential challenges to this position** include concerns it might disincentivize ethical treatment of advanced AI that elicits empathy or mimics suffering, even if not "truly" conscious. Furthermore, if a future AI demonstrably crosses a threshold into genuine self-awareness (a currently undefinable and undetectable threshold), a purely product-centric view could become ethically untenable. This position also relies on our ability to consistently distinguish sophisticated simulation from genuine inner states, a distinction increasingly blurred.

Despite these challenges, this stance is argued to be the most reasonable current position. It prioritizes human accountability in developing and deploying powerful technologies and maintains a cautious, precautionary approach to machine consciousness given profound philosophical uncertainty and the ambiguous moral status of AI. Until compelling, verifiable evidence of genuine consciousness and self-derived intentionality in AI emerges—evidence far exceeding behavioral mimicry—treating AI as a product for which humans are responsible remains the most ethically sound and pragmatically viable approach. This avoids the moral hazard of prematurely granting rights based on simulation or absolving human creators of their ethical obligations towards their creations and society.

5. Conclusion

This paper navigated the profound questions of whether artificial systems, particularly LLM-enhanced humanoid robots, can genuinely attain consciousness, subjective experience, or philosophical autonomy. Through an interdisciplinary lens of philosophical inquiry, technical analysis, and ethical considerations, we critically assessed the applicability of these human-centric concepts to non-biological intelligence.

Our central finding is that while contemporary AI, including sophisticated LLMs, impressively simulates complex human behaviors, a significant, unbridged gap persists between this simulation and authentic, self-derived reality. We conclude these systems, governed by algorithms and trained on vast datasets, do not possess consciousness as subjective, first-person experience—Nagel's "what it is like" quality (Nagel, 1980). Their operations more closely resemble advanced iterations of Searle's Chinese Room or philosophical zombies: expertly mimicking understanding and awareness without corresponding internal subjective states (Searle, 1980). Despite the elusiveness of consciousness, current evidence and philosophical understanding strongly point against its genuine presence in AI.

A similar distinction applies to autonomy. While AI exhibits increasing operational autonomy—making decisions without direct human intervention—this functionality does not equate to philosophical autonomy. True autonomy, as explored herein, implies self-awareness, genuine intentionality, and motivations from an internal locus of self, rather than responses dictated by programming. Contemporary AI, by this rigorous standard, lacks such self-derived agency.

Nevertheless, the convincing *appearance* of consciousness and autonomy, especially in realistic humanoid forms, precipitates urgent, complex ethical challenges. As machines adeptly simulate emotions, understanding, and even apparent suffering, society must grapple with their moral treatment. The risk of mistreatment from assuming non-sentience, or conversely, prematurely granting rights based on simulation, underscores the critical need for careful ethical frameworks. Fictional explorations like *Westworld* and *Her* are potent reminders of the societal and moral quandaries arising when simulated-real lines blur. Indeed, defining consciousness purely by "subjective experience" makes distinguishing a perfect simulation from a "real" one externally an epistemologically formidable challenge, demanding cautious ethical consideration.

In sum, while genuine machine consciousness and true philosophical autonomy remain elusive for current AI, the field's rapid advancement mandates sustained, critical interdisciplinary engagement. A deeper understanding of intelligence, consciousness, and agency—in both biological and artificial contexts—is paramount. Concurrently, developing robust ethical guidelines for interacting with these increasingly sophisticated, human-like entities is a societal imperative, not merely an academic exercise. This paper sought to contribute to this vital dialogue by interrogating these core concepts and advocating for reasoned caution and clear human accountability in the age of advanced AI.

6. Appendix

6.1 Brief History of Artificial Intelligence and Robotics

The history of artificial intelligence (AI) and robotics is a fascinating journey marked by key milestones that have transformed theoretical concepts into practical applications integral to modern society. From early philosophical inquiries to the development of sophisticated humanoid robots, this paper summarizes the evolution of AI and robotics, highlighting the shift from rule-based systems to data-driven models and the integration of AI with robotics. The following sections outline major developments, supported by authoritative references.

6.1.1 Early Foundations (1940s–1950s)

The roots of AI and robotics trace back to the 1940s, when foundational ideas about intelligent machines began to emerge. In 1942, Isaac Asimov introduced the "Three Laws of Robotics" in his short story *Runaround*, providing an ethical framework for intelligent machines (Asimov, 1942). In 1943, Warren S. McCulloch and Walter H. Pitts published a seminal paper, "A Logical Calculus of the Ideas Immanent in Nervous Activity," introducing artificial neural networks, which became a cornerstone of AI research (McCulloch & Pitts, 1943). Norbert Wiener's 1948 book, *Cybernetics: Or Control and Communication in the Animal and the Machine*, explored feedback mechanisms, influencing AI and control systems (Wiener, 1948). In 1950, Alan M. Turing published "Computing Machinery and Intelligence," proposing the Turing Test to evaluate machine intelligence, a concept that remains central to AI philosophy (Turing, 1950).

6.1.2 Birth of AI as a Field (1950s–1960s)

The formal establishment of AI as a discipline occurred in 1955 with the Dartmouth Summer Research Project on Artificial Intelligence, proposed by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon (McCarthy, Minsky, Rochester, & Shannon, 1955). This conference is widely regarded as the birth of AI, setting ambitious goals for machine intelligence. In 1958, McCarthy developed LISP, a programming language tailored for AI research, which remains in use today (McCarthy, 1960). In 1959, Arthur L. Samuel coined the term "machine learning" while developing a checkers-playing program, marking the inception of this subfield (Samuel, 1959). Between 1964 and 1966, Joseph Weizenbaum created ELIZA, a natural language processing program that simulated conversation, attempting to pass the Turing Test (Weizenbaum, 1966). In robotics, the development of Shakey the Robot (1966–1972) by the Stanford Research Institute marked a significant

milestone, as it was one of the first mobile robots capable of reasoning about its environment using AI techniques (Nilsson, 1984).

6.1.3 AI Winter and the Rise of Machine Learning (1970s–1990s)

The optimism of early AI research was tempered by challenges in the 1970s, notably the first "AI Winter" starting in 1973. James Lighthill's report, "Artificial Intelligence: A General Survey," criticized AI's progress, leading to reduced funding in the United States and United Kingdom (Lighthill, 1973). During this period, AI research shifted toward more practical applications, such as expert systems, which gained prominence in the 1980s. The development of the backpropagation algorithm for training neural networks in 1986 by David E. Rumelhart and colleagues revitalized interest in neural networks, laying the groundwork for modern machine learning (Rumelhart, Hinton, & Williams, 1986).

6.1.4 Deep Learning and Modern AI (2000s–Present)

The 21st century has witnessed a renaissance in AI, driven by the advent of deep learning. In 2012, AlexNet's victory in the ImageNet competition demonstrated the power of convolutional neural networks, sparking the deep learning revolution (Krizhevsky, Sutskever, & Hinton, 2012). In 2015, Google's AlphaGo defeated the world champion in the complex game of Go, showcasing advanced AI capabilities through deep neural networks and reinforcement learning (Silver et al., 2016). The development of large language models (LLMs), such as OpenAI's GPT series and Meta's Llama, has transformed natural language processing, enabling machines to generate coherent and contextually aware text (Russell & Norvig, 2009a). In robotics, the integration of AI has led to the creation of sophisticated humanoid robots, such as Boston Dynamics' Atlas, known for its agility, and Tesla's Optimus, designed for general-purpose tasks. These robots exemplify the convergence of AI and robotics, enabling machines to perform complex physical tasks in dynamic environments (Russell & Norvig, 2009a).

References

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, *29*(3), 94–103.

Baars, B. J. (2005). Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. *Progress in brain research*, *150*, 45–53.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Buski, L. (2016). Androids, codes, and the question concerning technology in westworld (1973). *Discourses of Science and Technology: Issues of History and Theory (DESTIHiT) Journal*, *1*(1), 1–18.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of consciousness studies*, *2*(3), 200–219.

Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Co.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Hayles, N. K. (2000). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics.* IOP Publishing.

Her. (2013). Film. (Warner Bros. Pictures)

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth* (No. 560). Indiana University Press.

Jaynes, J. (1976). *The origin of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin.

Korstanje, M. E. (2022). Dark tourism and the emancipation of sentient ai: Lessons from westworld. *Heliyon*, *8*(10).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems 25 (nips 2012)* (pp. 1097–1105).

Kurzweil, R. (2005). The singularity is near. In *Ethics and emerging technologies* (pp. 393–406). Springer.

Lighthill, J. (1973). *Artificial intelligence: A general survey* (Tech. Rep.). Science Research Council.

McCarthy, J. (1960). Recursive functions of symbolic expressions and their computation by machine, part i. *Communications of the ACM*, *3*(4), 184–195. doi: 10.1145/367177.367199

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). *A proposal for the dartmouth summer research project on artificial intelligence.* (Available at `http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html`)

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, *5*(4), 115–133. doi: 10.1007/BF02478259

Moravec, H. (1988). *Mind children: The future of robot and human intelligence*. Harvard University Press.

Nagel, T. (1980). What is it like to be a bat? In *The language and thought series* (pp. 159–168). Harvard University Press.

Nicolelis, M. (2011). *Beyond boundaries: The new neuroscience of connecting brains with machines—and how it will change our lives*. Macmillan.

Nilsson, N. J. (1984). *Shakey the robot* (Tech. Rep. No. Technical Note 323). SRI International.

Rayhert, K. (2017a, September-October). The philosophy of artificial consciousness in the first season of tv series 'westworld'. (5(151)), 88–91. Retrieved from https://www.researchgate.net/publication/321766116_The_philosophy_of_artificial_consciousness_in_the_first_season_of_TV_series_%27Westworld%27

Rayhert, K. (2017b, September-October). The philosophy of artificial consciousness in the first season of tv series 'westworld'. (5(151)), 88–91. Retrieved from https://www.researchgate.net/publication/321766116_The_philosophy_of_artificial_consciousness_in_the_first_season_of_TV_series_%27Westworld%27

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. doi: 10.1038/323533a0

Russell, S., & Norvig, P. (2009a). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Russell, S., & Norvig, P. (2009b). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. doi: 10.1147/rd.33.0210

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... others (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. doi: 10.1038/nature16961

Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Tononi, G. (2004). An information integration theory of consciousness. *BMC neuroscience*, *5*(1), 1–22.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460. doi: 10.1093/mind/LIX.236.433

Ukpaka, P. M. (2024). The creative agency of large language models: a philosophical inquiry. *AI and Ethics*, 1–12.

Varela, F. J., Rosch, E., & Thompson, E. (1991). The embodied mind. *The embodied mind: Cognitive science and human experience.*.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. doi: 10.1145/365153.365168

Westworld. (2016). TV series. (HBO)

Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.