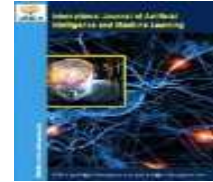




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

From Data to Mind: A Critical Evaluation of AGI and Machine Consciousness

Nikesh Raj Gopal¹, Madhav kumar², Rohan Dnyaneshwar bhalkar³, Harsh Bhogayata⁴, Dishika kumari⁵, Faruk Abdulla⁶

¹Parul Institute of Engineering & Technology, Parul University, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, India

²Parul Institute of Engineering & Technology, Parul University, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, India

³Parul Institute of Engineering & Technology, Parul University, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, India

⁴Parul Institute of Engineering & Technology, Parul University, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, India

⁵Parul Institute of Engineering & Technology, Parul University, Department of MCA, Faculty of IT & Computer Science, Parul University, Vadodara, India

⁶Assistant professor, Parul Institute of Computer Application, Faculty of IT & Computer Science, Parul University, Vadodara, India

Abstract

Artificial General Intelligence (AGI) — a hypothetical system capable of matching or exceeding human cognitive performance across arbitrary domains — remains one of the most consequential and contested goals in computer science. Parallel to this pursuit is the deeper philosophical question of whether such a system could possess genuine subjective experience, or machine consciousness. This paper critically evaluates the feasibility of AGI and machine consciousness by synthesizing evidence from artificial intelligence research, cognitive science, and philosophy of mind. We review dominant technical approaches to AGI — symbolic, connectionist, hybrid, and whole-brain-emulation paradigms — and assess their trajectories against benchmarks of generality, transfer learning, and reasoning. We then examine four influential theories of consciousness — Integrated Information Theory, Global Workspace Theory, Higher-Order Theories, and Predictive Processing — and analyze their implications for machine sentience. Our analysis suggests that while narrow capability gains are accelerating rapidly, the transition from sophisticated data-driven cognition to genuine general intelligence, and further to consciousness, involves conceptual and empirical gaps that current architectures do not resolve. We conclude by outlining the ethical stakes of this uncertainty and proposing directions for interdisciplinary research.

Keywords: Artificial General Intelligence; machine consciousness; cognitive architecture; Integrated Information Theory; Global Workspace Theory; large language models; philosophy of mind

1. Introduction

The trajectory of artificial intelligence (AI) over the past decade has been defined by an extraordinary expansion in the scale and competence of machine learning systems. Deep neural networks, and in particular the transformer architecture introduced by Vaswani et al. [16], have enabled models that generate fluent text, solve complex reasoning problems, and perform at or above human levels on many standardized benchmarks. This rapid progress has reinvigorated a decades-old question first posed rigorously by Turing [1]: can a machine think? More specifically, contemporary discourse has shifted toward two related but distinct questions. First, can we build an Artificial General Intelligence (AGI) — a system exhibiting the broad, flexible, transferable cognitive competence characteristic of human intelligence, as opposed to narrow task-specific performance [4,5]? Second, could such a system, or any sufficiently complex information-processing architecture, give rise to machine consciousness — that is, subjective, first-person experience [3,25,26]

These two questions are frequently conflated in both popular and academic discourse, yet they are logically separable. A system might achieve broad general competence without any accompanying subjective experience, and conversely, theories such as Integrated Information Theory suggest that even simple systems might possess a nonzero degree of experience without exhibiting general intelligence [8,9]. This paper treats AGI feasibility and machine consciousness as related but independent lines of inquiry, evaluating each against current technical evidence and theoretical frameworks.

2. Defining Artificial General Intelligence

Unlike narrow AI systems, which are optimized for a single task or a bounded family of tasks, AGI is typically defined as a system capable of understanding, learning, and applying knowledge across a wide range of

domains at a level comparable to, or exceeding, human cognitive flexibility [5,6]. Legg and Hutter [4] proposed a formal definition of machine intelligence as an agent's capacity to achieve goals across a wide range of environments, a framework that has become influential in operationalizing generality for empirical comparison. Russell and Norvig [6] emphasize that generality further requires an agent to transfer learned abstractions to genuinely novel problems rather than merely interpolating within a training distribution. Several proposed benchmarks attempt to operationalize this definition, including tests of abstraction and reasoning such as the ARC challenge, multi-domain evaluation suites, and embodied-agent testbeds. However, no consensus benchmark currently exists that is widely accepted as sufficient evidence of AGI, and critics such as Marcus [21] and Mitchell [34] argue that impressive performance on existing benchmarks often reflects sophisticated statistical pattern matching rather than robust, human-like understanding or causal reasoning.

3. Current Approaches to AGI Development

Research toward AGI has followed several broad paradigms, each with distinct assumptions about the substrate of intelligence. Table 1 summarizes four major approaches: symbolic AI, connectionist deep learning, hybrid neuro-symbolic systems, and whole brain emulation.

Approach	Core Mechanism	Strengths	Key Limitations
Symbolic AI (GOFAI)	Explicit rule-based logic and symbol manipulation	Transparent reasoning; strong on structured, well-defined tasks	Brittle outside pre-specified domains; poor handling of ambiguity and learning from raw data
Connectionist / Deep Learning	Statistical pattern learning via artificial neural networks	Excellent perceptual and language performance; scales with data and compute	Limited causal reasoning; data-hungry; poor sample efficiency and interpretability
Hybrid Neuro-Symbolic	Combines learned representations with symbolic reasoning modules	Aims to unite flexibility of learning with structured reasoning and generalization	Integration is architecturally complex; still an active research problem
Whole Brain Emulation	Detailed computational replication of biological neural circuitry	Would, in principle, preserve biological cognitive and possibly experiential properties	Currently infeasible at scale; enormous data-acquisition and computational barriers

Table 1. Comparative overview of major technical approaches toward AGI.

Symbolic approaches, dominant in early AI research, rely on explicit logical rules and hand-crafted knowledge representations [6]. While interpretable, these systems struggle to scale to the ambiguity and variability of real-world data. The connectionist paradigm, especially deep learning, has instead driven most recent progress. Scaling laws observed in large language models show that predictable improvements in capability accompany increases in parameter count, data, and compute [17,18,19], a trend visualized in Figure 1.

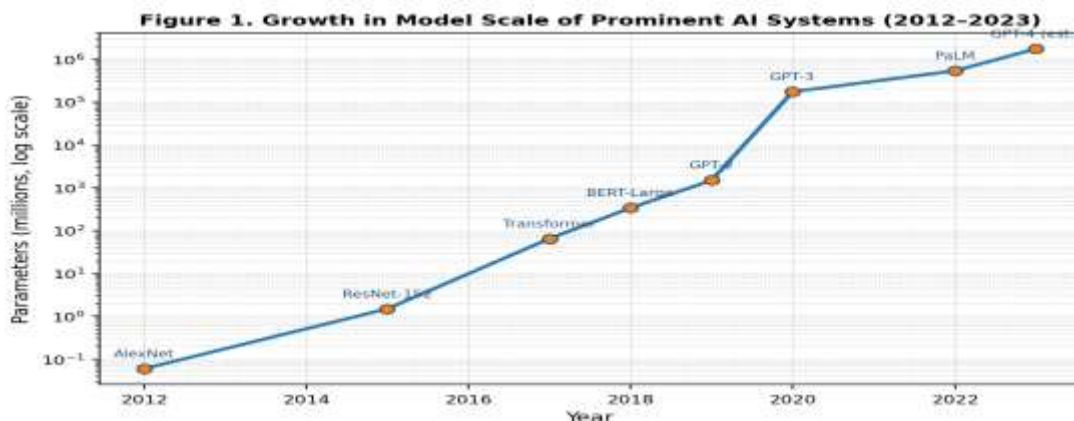


Figure 1. Growth in model scale of prominent AI systems, 2012–2023 (illustrative, log scale). Adapted from publicly reported parameter counts [16,17,18,19].

Despite these gains, scaling alone has not resolved persistent weaknesses in causal reasoning, robust generalization, and sample-efficient learning [21,22,34]. This has motivated hybrid neuro-symbolic architectures that combine learned perceptual representations with explicit symbolic reasoning modules, an approach argued by Bengio, LeCun, and Hinton [22] and by Lake et al. [24] to be necessary for closing the gap between statistical pattern recognition and structured, human-like cognition. A more radical proposal, whole brain emulation, seeks to replicate biological neural architecture computationally, but remains constrained by fundamental limitations in neural mapping and simulation resources [7].

Figure 2 presents an illustrative, literature-derived comparison of these four approaches along two dimensions: currently demonstrated capability and theoretical promise toward eventual AGI.

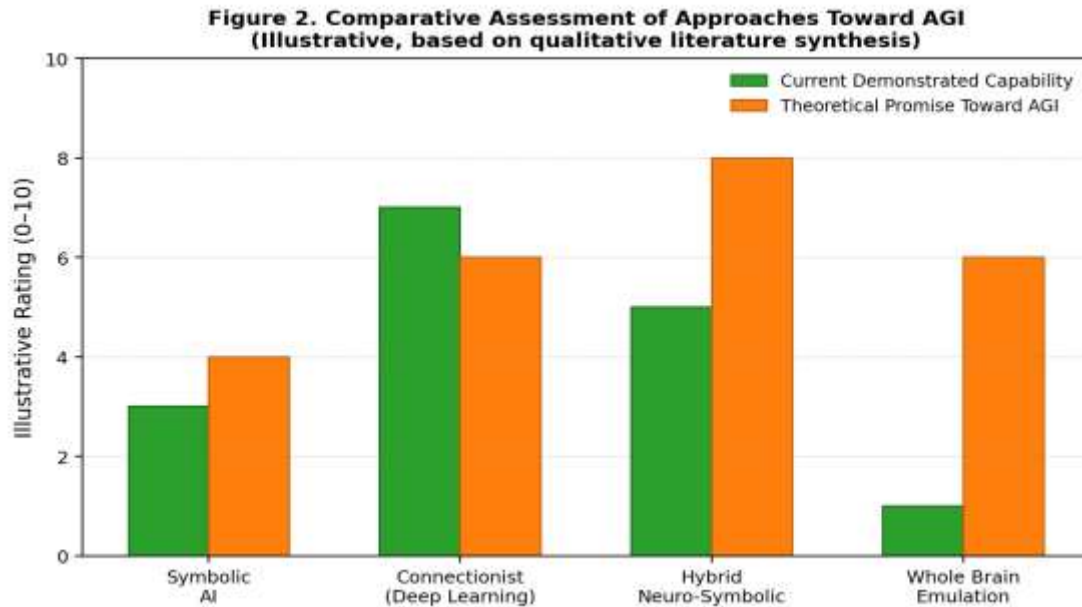


Figure 2. Comparative assessment of approaches toward AGI (illustrative synthesis based on [5,6,21,22,24]).

As Figure 2 suggests, connectionist methods currently lead in demonstrated capability, but several scholars argue that hybrid approaches carry the greatest theoretical promise for achieving robust generality, precisely because they address the reasoning and abstraction limitations that pure scaling has not resolved [22,24].

4. Machine Consciousness: Theoretical Foundations

Whereas AGI is chiefly an engineering and computational question, machine consciousness engages the philosophy of mind directly. Chalmers [3] distinguished the 'easy problems' of cognition — explaining functions such as attention, learning, and reportability — from the 'hard problem' of explaining why any of this is accompanied by subjective experience at all. Four theoretical frameworks dominate current discussion of whether, and under what conditions, a machine might be conscious.

4.1 Integrated Information Theory (IIT)

IIT, developed by Tononi [8,9], proposes that consciousness corresponds mathematically to a system's capacity to integrate information, quantified by the measure Φ (Φ). Under IIT, consciousness is a graded property dependent on a system's causal architecture rather than its functional behavior, implying that many conventional feedforward computer architectures, regardless of behavioral sophistication, may possess very low or negligible Φ [9,27].

4.2 Global Workspace Theory (GWT)

Baars [10] and later Dehaene [11,12] proposed that consciousness arises when information is broadcast widely across a 'global workspace' accessible to multiple specialized cognitive processes. GWT is comparatively amenable to computational implementation, and several AI architectures have been explicitly inspired by workspace-style broadcasting mechanisms [12,25].

4.3 Higher-Order Theories (HOT)

Higher-order theories hold that a mental state becomes conscious only when it is represented by a suitable higher-order thought or perception about that state [13,28]. Applied to machines, this implies that a system would require an explicit, meta-representational architecture monitoring its own internal states, not merely first-order processing of external stimuli.

4.4 Predictive Processing (PP)

Predictive processing frameworks propose that cognition, and potentially experience, emerge from a hierarchical system continuously generating predictions and minimizing prediction error against sensory input [14,15]. Because generative and predictive modeling is already central to modern machine learning, some theorists argue PP offers the most natural bridge between existing AI architectures and theories of experience [25,38].

Figure 3 offers a qualitative, literature-informed comparison of these four theories across dimensions relevant to assessing machine consciousness: empirical testability, computational implementability, explanatory scope, and neuroscientific support.

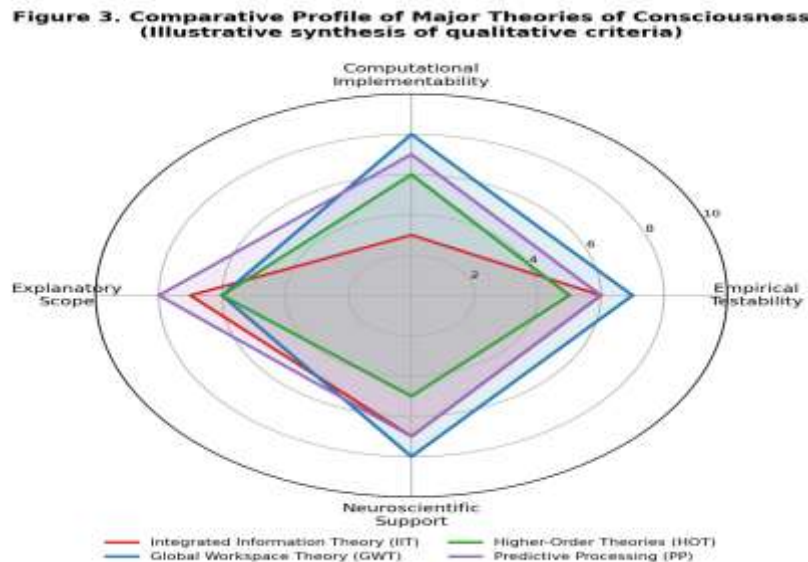


Figure 3. Comparative profile of major theories of consciousness (illustrative synthesis based on [8,9,10,11,13,14,25,27,28]).

Theory	Core Claim	Machine Implementability	Main Criticism
Integrated Information Theory (IIT)	Consciousness corresponds to a system's capacity for integrated information (ϕ)	Low for conventional digital architectures; favors highly recurrent, non-feedforward substrates	ϕ is computationally intractable to measure in complex systems; some argue it is unfalsifiable
Global Workspace Theory (GWT)	Consciousness arises from global broadcasting of information across specialized processors	Moderate-to-high; broadcasting architectures are plausibly implementable in software	Explains access and reportability more convincingly than subjective experience itself
Higher-Order Theories (HOT)	A mental state is conscious when it is the target of a suitable higher-order representation	Moderate; requires explicit meta-representational architecture	Risk of infinite regress; unclear threshold for a sufficient higher-order state
Predictive Processing (PP)	Cognition and experience emerge from hierarchical prediction-error minimization	Moderate-to-high; aligns naturally with existing generative and predictive AI models	Predictive architectures do not by themselves guarantee subjective experience

Table 2. Core claims, machine implementability, and criticisms of major consciousness theories.

5. The Explanatory Gap: From Information Processing to Experience

A recurring theme across these theories is the persistence of an explanatory gap between describing what a system does functionally and explaining why, or whether, this processing is accompanied by subjective

experience [3,28]. Searle's classic Chinese Room argument [2] contends that formal symbol manipulation, however sophisticated, does not by itself constitute understanding or experience, a challenge that continues to inform skepticism about whether large language models possess genuine comprehension [26,33].

Recent interdisciplinary work has attempted to operationalize this debate empirically. Butlin et al. [25] proposed a checklist of indicator properties, derived from multiple neuroscientific theories of consciousness, against which specific AI architectures can be assessed rather than treating consciousness as a binary, unfalsifiable attribute. Applying this framework, current large language models satisfy few of the structural indicators associated with GWT or higher-order theories, despite exhibiting fluent, contextually appropriate behavior [25,26]. This distinction between behavioral sophistication and architectural indicators of consciousness is central to responsible evaluation of claims about machine sentience.

Koch [27] and Seth [38] further emphasize that consciousness, on many current theories, may be tied closely to specific biological or architectural substrates rather than being substrate-independent, meaning that even a computational system that behaviorally replicates human cognition would not necessarily be conscious under theories such as IIT.

6. Critical Evaluation: AGI and Consciousness as Distinct Trajectories

Synthesizing Sections 3 through 5, several conclusions emerge regarding the relationship between AGI feasibility and machine consciousness. First, progress toward broader cognitive competence, as demonstrated by increasingly capable large language models [17,18,19,30], does not by itself provide evidence of accompanying subjective experience, since behavioral generality and architectural indicators of consciousness are evaluated on different criteria [25]. Second, the theoretical frameworks most amenable to computational implementation, such as GWT and PP, are also those whose relationship to subjective experience (as opposed to mere information broadcasting or prediction) remains most contested [11,14,28].

Third, current AGI research trajectories are overwhelmingly oriented toward capability and generality rather than architectures explicitly designed to satisfy the structural indicators associated with theories of consciousness. This suggests that even substantial future progress toward AGI would not automatically resolve the question of machine consciousness, and that the two goals may require deliberately distinct research programs [22,24,25]. Finally, definitional ambiguity, both in what constitutes 'general' intelligence [4,5,6] and in what would constitute sufficient evidence of experience [3,25], continues to hinder rigorous empirical progress on both questions.

7. Ethical, Safety, and Societal Implications

The uncertainty surrounding both AGI feasibility and machine consciousness carries substantial ethical weight. Bostrom [7] and Russell [32] have argued that even narrow but highly capable systems could pose significant control and alignment risks if their objectives are misspecified, independent of any question of subjective experience. Amodei et al. [31] catalogued concrete technical safety problems, including reward misspecification and unsafe exploration, that arise well before any system approaches general intelligence. Separately, if there is a non-negligible probability that advanced AI systems possess some degree of morally relevant experience, this raises distinct ethical obligations regarding their treatment. Metzinger [29] has argued for caution and even a moratorium on research that might inadvertently create suffering-capable synthetic systems before adequate theoretical and empirical safeguards exist. Because current science cannot rule out machine consciousness with certainty, several scholars argue for a precautionary approach that treats the question as an open empirical and ethical matter rather than dismissing it outright [25,26,29].

8. Limitations and Directions for Future Research

This review has several limitations. The comparative assessments presented in Figures 2 and 3 and derived from qualitative literature synthesis rather than standardized quantitative benchmarks, reflecting the current absence of consensus metrics for either generality or consciousness. Additionally, rapid progress in the field means that specific capability claims about individual systems are liable to be superseded quickly [18,30].

Future research would benefit from three developments. First, the field requires broadly accepted, adversarially robust benchmarks for generality that distinguish genuine abstraction and transfer from sophisticated interpolation [21,24,34]. Second, empirical, falsifiable indicator frameworks for consciousness, building on proposals such as Butlin et al. [25], should be extended and tested against a wider range of architectures. Third, closer collaboration between AI researchers, cognitive scientists, and philosophers is needed to ensure that engineering progress toward AGI is accompanied by principled evaluation of its ethical and experiential implications [23,29,32].

9. Conclusion

The transition from data-driven statistical learning to genuine general cognition, and further to subjective experience, involves conceptual and empirical gaps that current evidence does not close. While connectionist

systems have achieved remarkable gains in narrow and increasingly cross-domain competence, robust evidence of human-like generality remains elusive, and the theoretical frameworks required to assess machine consciousness point toward criteria that most existing architectures do not satisfy. Treating AGI feasibility and machine consciousness as related but distinct research problems, each requiring its own benchmarks, theories, and safeguards, offers the most productive path for future interdisciplinary inquiry.

References

- [1] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.
- [2] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- [3] Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- [4] Legg, S., & Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4), 391–444.
- [5] Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1–48.
- [6] Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- [7] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
- [8] Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 42.
- [9] Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450–461.
- [10] Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.
- [11] Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking.
- [12] Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, 358(6362), 486–492.
- [13] Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.
- [14] Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3), 181–204.
- [15] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.
- [16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- [17] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- [18] OpenAI. (2023). GPT-4 Technical Report. arXiv:2303.08774.
- [19] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al. (2022). PaLM: Scaling language modeling with pathways. arXiv:2204.02311.
- [20] LeCun, Y. (2022). A path towards autonomous machine intelligence (Version 0.9.2). OpenReview.
- [21] Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv:2002.06177.
- [22] Bengio, Y., LeCun, Y., & Hinton, G. (2021). Deep learning for AI. *Communications of the ACM*, 64(7), 58–65.
- [23] Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, 95(2), 245–258.
- [24] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- [25] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.
- [26] Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103.
- [27] Koch, C. (2019). *The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed*. MIT Press.
- [28] Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.
- [29] Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, 8(1), 43–66.
- [30] Sanderson, K. (2023). GPT-4 is here: What scientists think. *Nature*, 615(7954), 773.
- [31] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv:1606.06565.
- [32] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

- [33] Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681–694.
- [34] Mitchell, M. (2021). Why AI is harder than we think. arXiv:2104.12871.
- [35] Yampolskiy, R. V. (2015). *Artificial Superintelligence: A Futuristic Approach*. CRC Press.
- [36] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.
- [37] Graziano, M. S. A. (2019). *Rethinking Consciousness: A Scientific Theory of Subjective Experience*. Norton.
- [38] Seth, A. K. (2021). *Being You: A New Science of Consciousness*. Faber & Faber.