Philosophy, Cognitive Science, and Policy:

Interdisciplinary Perspectives on Generative AI from Wittgenstein, Lewis, Dennett, and Nagel

Michael Stoyanovich
michael@mstoyanovich.com

Disclaimer

This paper is for informational and educational purposes only. The ideas and analyses presented, including discussions of ethics and policy, reflect the author's interpretations and do not constitute professional, legal, or regulatory advice. Readers are encouraged to critically evaluate the content and consult qualified professionals and authorities before applying any concepts discussed. The author assumes no responsibility or liability for any actions taken based on this work.

License & Attribution

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) License. You are free to share and adapt the content as long as proper credit is given to the author. No additional permissions are required for non-commercial or commercial use, provided attribution is maintained.

For full license details, visit: https://creativecommons.org/licenses/by/4.0/.

Executive Summary

Context & Purpose

As generative AI—especially large language models (LLMs) like OpenAI's GPT series—reshapes human-computer interaction, critical questions arise: How do these systems handle language? Do they "understand" in any meaningful way? And what ethical or policy considerations should guide their development? This paper leverages insights from philosophy, cognitive science, and AI research to answer these questions, integrating perspectives from Ludwig Wittgenstein, David Lewis, Daniel Dennett, and Thomas Nagel.

Core Argument

AI language models exhibit fluency, coherence, and adaptability, but their "understanding" remains an open question. Through Wittgenstein's *language games*, Lewis's *scorekeeping in conversation*, Dennett's *intentional stance*, and Nagel's *hard problem of consciousness*, this paper provides a structured framework for evaluating what AI can and cannot do—and why it matters.

Key Findings

1. **Language as Use (Wittgenstein)** – LLMs generate text based on statistical patterns but do not engage in human "forms of life," meaning they approximate rule-following without genuine understanding.
2. **Context Sensitivity (Lewis)** – AI maintains local conversational context but lacks long-term memory and true dialogue scorekeeping, leading to inconsistencies in extended interactions.
3. **Interpretation & Anthropomorphism (Dennett)** – Users naturally treat AI "as if" it has beliefs, but this heuristic risks overtrust, ethical missteps, and misplaced accountability.
4. **Consciousness & Limits (Nagel)** – Even the most advanced AI lacks subjective experience, reinforcing the distinction between mimicking understanding and possessing it.

Implications & Takeaways

- **AI Designers**: Improve long-term context tracking, transparency, and human-in-the-loop feedback.
- **Users & Organizations**: Be aware of AI's limitations—treating LLMs as thought partners rather than autonomous agents.
- **Policymakers**: Implement clear AI transparency requirements and safeguards against undue anthropomorphism.

Conclusion

Generative AI is a powerful tool, but it remains fundamentally different from human cognition. By understanding its limitations through interdisciplinary perspectives, we can develop and govern AI more responsibly. This paper serves as an invitation for continued discussion at the intersection of philosophy, AI research, and policy.

Abstract

This paper examines how philosophical frameworks—primarily Ludwig Wittgenstein's language games, David Lewis's conversational scorekeeping, Daniel Dennett's intentional stance, and Thomas Nagel's perspective on consciousness—can enhance our understanding of and interaction with generative AI, particularly large language models (LLMs) such as OpenAI's GPT series. Wittgenstein's concepts of language games and rule-following offer insights into how AI handles language within social contexts, while Lewis's scorekeeping theory illustrates the dynamic updating of shared conversational assumptions. Dennett's intentional stance provides a pragmatic heuristic for interpreting AI behavior without requiring genuine understanding or consciousness, and Nagel's critique in "What's it like to be a bat?" highlights the gap between simulated behavior and subjective experience.

These views are enriched by additional perspectives, including embodied cognition theory, cognitive architectures, pragmatism, and social constructivism, as well as advances in AI interpretability, ethics, and global policy debates. While some scholars argue that sufficiently advanced AI might approximate aspects of human cognition, this paper maintains that such systems lack true subjective awareness, embodied context, and the ability to engage in the socially embedded rule-following that characterizes human language games. The paper proposes actionable strategies for improving AI design—such as memory-augmented neural networks, context management, and transparency—and addresses counterarguments, ethical considerations, and emerging trends in responsible AI development. Throughout, key concepts are explained in clear language to ensure accessibility for non-specialists.

Keywords: generative AI; language games; scorekeeping; intentional stance; consciousness; embodied cognition; AI ethics; cognitive science; neuroscience; explainable AI; cognitive architectures; posthumanism; AI governance; policy; global regulation

1. Introduction

Generative AI—epitomized by large language models (LLMs) like OpenAI's GPT series—is transforming human-computer interaction by generating contextually coherent text from massive datasets. As these systems increasingly impact sectors such as education, healthcare, law, and creative industries, critical questions arise: How do these systems "understand" language, and how should their behavior be interpreted? Moreover, how can we design and govern these AIs to ensure beneficial outcomes? Addressing such questions requires an interdisciplinary approach that goes beyond technical considerations to include philosophy of mind and language, cognitive science, and ethics.

This paper integrates classical philosophical insights from Wittgenstein, Lewis, Dennett, and Nagel with emerging perspectives from cognitive science, neuroscience, AI interpretability, posthumanism, critical theory, and global policy frameworks. The goal is to develop a comprehensive conceptual framework that enhances both theoretical understanding and practical application of generative AI, while considering ethical, societal, and regulatory dimensions. The discussion remains accessible to non-specialists, avoiding unnecessary jargon and explaining concepts in plain language.

Roadmap

- **Section 2 (Literature Review)** surveys philosophical and technical foundations—encompassing embodied cognition, cognitive architectures, posthumanism, and relevant ethical/policy debates—to situate generative AI in a broader interdisciplinary context.
- **Section 3 (Theoretical Framework)** integrates Wittgenstein's language games, Lewis's scorekeeping, Dennett's intentional stance, and Nagel's consciousness critique, enriched by pragmatism, social constructivism, and AI interpretability research.
- **Section 4 (Methodology)** outlines the blend of philosophical analysis and empirical integration used to evaluate LLMs, with an eye toward case studies, user surveys, and validation strategies.
- **Sections 5 and 6 (Results, Counterarguments, and Practical Implications)** discuss findings from applying the theoretical framework to current AI systems,

address counterarguments from cognitive science, and propose strategies (e.g., memory-augmented networks, transparency measures, policy recommendations) to improve AI design and governance.
- **Section 7 (Conclusion)** summarizes key insights, highlights limitations, and offers recommendations for future research, emphasizing an interdisciplinary approach to responsible and effective AI development. An Ethical and Permissions note clarifies data usage, and Acknowledgments & Disclosure address authorship, AI tool usage, and the "living document" nature of this work.

2. Literature Review

2.1 Philosophical Foundations of AI

Early debates in AI philosophy set the stage for understanding generative models. A seminal argument is John Searle's Chinese Room (Searle, 1980), which posits that mere symbol manipulation (as in a computer following code) does not yield genuine understanding or semantics. Searle's thought experiment suggests that an AI could appear to converse fluently in Chinese by following syntactic rules, yet lack true understanding—implying that syntax alone does not produce semantics. In contrast, Alan Turing's criterion for intelligence (the Turing Test, Turing, 1950) focuses on observable behavior: if a machine's responses are indistinguishable from a human's, we may as well call it intelligent, sidestepping the question of internal understanding. This tension between behaviorism and semantic internalism continues to inform debates about LLMs. Hubert Dreyfus (1992) and Martin Heidegger (1927) offered phenomenological critiques, arguing that intelligence is deeply tied to embodied, context-rich experience in the world—something classical AI lacked. Shannon and Weaver's (1949) information theory provided a foundation for computational linguistics and the statistical approach used by modern LLMs, but by treating information primarily in terms of bits and entropy, it did not address the deeper question of meaning. John Haugeland later underscored the importance of "embodied intentionality" in understanding cognition, presaging arguments that true intelligence must incorporate more than abstract symbol processing.

Embodied Cognition Theory has since grown into a significant perspective in cognitive science, emphasizing that human cognition arises from real-time interactions between the mind, body, and environment (Clark, 2008; Varela, Thompson & Rosch, 1992). By grounding thought in sensory and motor processes, embodied cognition suggests that a non-embodied AI—merely manipulating linguistic symbols—may never achieve the full richness of human-like understanding. In the context of generative AI, this raises questions about how LLMs, which rely on text-only training, could ever capture the lived experiences that shape human linguistic meaning. Indeed, some researchers propose integrating robotics or multimodal data (visual, tactile, auditory) to give AI systems at least a partial "body in the world," thereby potentially mitigating the symbol-grounding problem.

Cognitive Architectures like SOAR or ACT-R offer another angle on how AI might move beyond brute-force statistical approaches toward something more akin to human cognition (Laird, 2012; Anderson et al., 1998). These architectures model functional modules—such as memory stores, perceptual processors, and rule-based reasoning—suggesting a way for AI systems to integrate symbolic and sub-symbolic processes. While large language models excel at pattern recognition and language generation, they typically lack the structured memory and goal-directed components that cognitive architectures attempt to replicate. Incorporating insights from these architectures could enrich the design of future LLMs, making them more context-aware, capable of long-term planning, and sensitive to the "global workspace" aspects of cognition. Researchers exploring hybrid approaches argue that bridging LLMs with cognitive architectures or memory-augmented modules might yield AI systems that demonstrate more robust forms of reasoning and understanding.

These foundational discussions set up the challenge: can generative AI move beyond being a sophisticated manipulator of symbols to something that grasps meaning? Recent critics of LLMs echo these concerns, describing them as "stochastic parrots" that generate plausible language without true comprehension. Proponents, however, point to increasingly general capabilities of advanced models as evidence of at least a form of understanding emerging from complex patterns. This literature provides a backdrop for applying specific philosophical lenses—Wittgenstein's language games, Lewis's scorekeeping, Dennett's intentional stance, and Nagel's critique—to AI systems, which we turn to in subsequent sections.

2.2 Wittgenstein's Philosophy and AI

Ludwig Wittgenstein's later work, especially Philosophical Investigations (1953), introduces the idea of language games, wherein meaning emerges from use within specific social activities and contexts. Words do not have fixed definitions in isolation; their meaning is defined by the "rules" of the particular language game being played. For instance, the word pawn means something different in the "game" of chess than it does in everyday conversation. Crucially, for Wittgenstein, language is a public, social activity —rule-following and meaning are grounded in shared forms of life (cultural and practical contexts). While some scholars argue that AI could become a participant in language games through sufficient interaction, this paper follows the view that true language use is inseparable from human forms of life—contextually rich, socially embedded, and embodied. Scholars like P. M. S. Hacker and Danièle Moyal-Sharrock have argued that this communal nature of language poses a challenge for LLMs, which generate text based on statistical patterns rather than genuine participation in human forms of life. Winograd and Flores (1986) similarly drew on Wittgenstein (and Heidegger) to critique AI's purely formal approach to language, suggesting that computers lack the lived context that imbues human language with depth. From this perspective, if an AI lacks an authentic understanding of the rules as grounded in human practice, it is not truly "playing the language game"—merely simulating it.

Social Constructivism further illuminates this communal aspect by arguing that meaning is co-created through social interactions and shared conventions. In line with Wittgenstein's emphasis on public criteria for rule-following, social constructivists highlight how the collective negotiation of concepts shapes reality—an iterative process in which humans converge on norms and meanings. LLMs, by contrast, rely primarily on static text corpora, lacking the ongoing communal feedback loops that living language communities use to refine and revise their shared linguistic practices.

Pragmatism—particularly as advanced by philosophers like William James and John Dewey—parallels Wittgenstein's view that meaning is rooted in practical usage. Pragmatists argue that concepts acquire meaning through their consequences and utility in real-world problem-solving contexts. From this angle, a word's significance lies in how it guides action and thought. While LLMs can generate contextually appropriate text, they do so without genuine practical engagement or an experiential stake in the outcomes. Thus, one could argue that, from a pragmatist standpoint, LLMs remain detached from the pragmatic dimension that underpins genuine rule-following in human language use.

This issue ties back to the symbol grounding problem: LLMs handle symbols (words) without direct connection to their real-world referents. Consequently, critics question whether generative AI can ever achieve meaningful language use if it never participates in the "forms of life" that give words their significance. Others maintain that sufficient breadth and depth of data might approximate the effects of communal participation, allowing the model to mimic context-sensitive use fairly closely. Whether such mimicry counts as "understanding" is an open debate, which subsequent sections explore from multiple philosophical angles.

2.3 David Lewis and Contextual Dynamics

David Lewis's scorekeeping theory of conversation (Lewis, 1979) provides another useful lens for understanding how context shapes linguistic meaning. In any dialogue, participants keep a metaphorical "score" of the context—facts that have been established, assumptions about what words refer to, the state of the conversation, and so forth. As the conversation progresses, each utterance can update this contextual score. For instance, if someone says "Let's meet at the bank" in the middle of a fishing discussion, the score (context) will record that bank likely refers to a riverbank rather than a financial institution. Lewis's core insight is that meaning in conversation is highly dynamic and context-dependent, maintained through an implicit consensus that constantly evolves with each contribution to the dialogue.

Modern LLM-based chatbots mimic a form of scorekeeping by using attention mechanisms to track recent context in an input window. This allows them to exhibit a degree of context-sensitivity—answering follow-up questions coherently, interpreting pronouns, and so forth. However, unlike human interlocutors, LLMs typically have a fixed memory window and do not genuinely retain long-term context or purpose. Consequently, once the text falls outside the model's input buffer, it no longer influences the "score." This leads to known limitations: an AI may contradict earlier statements or fail to adapt to subtle context shifts over the course of a lengthy conversation.

Cognitive Pragmatics research reinforces the importance of adaptive context management. Human communicators track not only what has been said but also participants' intentions, background knowledge, and situational cues, updating these assumptions as the interaction unfolds. By comparison, LLMs operate largely on local context, lacking an internal model of a conversation's evolving goals and shared knowledge. This shortcoming is especially noticeable in multi-turn dialogues where references to earlier details can get lost or overridden by newer inputs.

Memory-Augmented Neural Networks offer one potential remedy. By integrating a structured memory component (e.g., an external database or a specialized neural module), AI systems can preserve key facts and conversation states beyond the immediate token window. Such architectures could allow an LLM to retrieve relevant past information and maintain a more robust "score" over extended exchanges. Similarly, logic-based approaches like Reiter's default logic (1980) can complement neural methods by encoding and updating assumptions until contradicted by new information. Developers are actively experimenting with these techniques to address LLMs' memory limitations, aiming to improve contextual coherence and consistency.

By applying Lewis's theory to LLMs, we see that context is not a static snapshot but a dynamic, continuously renegotiated framework. Designing AI systems that actively update their "conversational scoreboard"—through memory-augmentation, retrieval strategies, or a blend of symbolic and sub-symbolic reasoning—represents a critical step toward achieving more human-like dialogue management.

2.4 Dennett's Intentional Stance and AI

Daniel Dennett's intentional stance (Dennett, 1989) is a strategy where we interpret an entity's behavior by ascribing beliefs, desires, and intentions to it—treating it as if it were a rational agent. This stance is pragmatically useful for predicting the entity's behavior, regardless of whether it actually possesses a mind. For example, one can predict a chess computer's moves by assuming it "wants" to win and "knows" the rules of chess, even though internally it is merely executing algorithmic processes. In the context of large language models, this stance naturally arises when users say an AI "knows" a great deal or "understands" questions, even though the AI is ultimately a statistical engine generating text.

Anthropomorphism in AI Ethics

A key implication of adopting the intentional stance toward AI is the risk of anthropomorphism—mistakenly attributing human-like understanding, motives, or emotions to systems that do not actually possess them. Such over-ascription can lead users to develop misplaced trust or emotional bonds with AI, resulting in adverse outcomes (Coeckelbergh, 2020). For instance, a user who believes a chatbot genuinely "cares" might divulge sensitive information or rely on it for emotional support in contexts where professional human help is needed. From an ethical standpoint, designers and policymakers must anticipate and mitigate these risks. Features like user education, disclaimers ("I am an AI and do not have feelings or personal beliefs"), or interface cues that highlight the AI's limitations can reduce harmful anthropomorphism.

Critical Theory Perspectives

From a critical theory standpoint, how we talk about AI—in human-like terms or otherwise—reflects broader societal attitudes and power structures. Some scholars argue that the intentional stance can obscure the labor, data, and socio-technical systems underpinning AI development; by anthropomorphizing, we overlook the humans involved in data annotation, system maintenance, or the corporate entities that control AI technologies. Critical theorists also stress that assigning agency to AI might absolve humans of responsibility when technology is used in harmful ways (e.g., "the algorithm decided," rather than admitting corporate or governmental accountability). Consequently, critically examining why and how we deploy Dennett's stance can reveal hidden assumptions about human agency, ethics, and technology's role in society.

Overall, Dennett's perspective underscores that the intentional stance is a choice rather than an assertion of fact. We can treat AI systems as if they have beliefs or desires to streamline interactions, but we must remember this is a heuristic tool, not a literal description of the AI's internal states. Designing systems that clearly communicate their non-human nature can help users strike a balance—benefiting from the stance's practical utility while avoiding undue anthropomorphism.

2.5 Nagel's Challenge to AI Consciousness

Thomas Nagel's famous essay "What is it like to be a bat?" (1974) poses a fundamental question about subjective experience. Nagel argues that even if we know everything about the objective, physical processes of a bat's brain, we still would not know what it is like for the bat to experience the world (e.g., the subjective feeling of echolocation). This ineffable, first-person quality of experience—often termed qualia—highlights a potentially unbridgeable gap between an objective description (or simulation) of a being and the being's own perspective.

Applying this to AI, Nagel might ask, "What is it like to be GPT-4?" The common intuition is that there is nothing it is like to be GPT-4; an LLM, as an artifact, has no inner life or conscious viewpoint. It processes text statistically, without any "felt" experience. Hence, no matter how perfectly an AI might simulate human conversational behavior, there remains the so-called hard problem of consciousness unaddressed—namely, how subjective awareness could emerge from computational processes. Philosophers like David Chalmers (1996) distinguish between the "easy problems" of consciousness (explaining cognitive functions and behaviors) and the "hard problem" (explaining why and how those processes are accompanied by phenomenal experience). Current AIs tackle many of the "easy" cognitive tasks—categorizing images, conversing, playing games—yet according to Nagel's argument, they do not approach the hard problem, as there is no indication that their statistical algorithms generate subjective awareness.

Some contemporary neuroscientists and theorists have proposed measures or theories of consciousness (e.g., Tononi's Integrated Information Theory (IIT) or global workspace theory) to gauge how or whether consciousness might arise in an AI system. Under IIT, for instance, a purely feed-forward transformer model might score low on integrated information, suggesting it lacks the kind of unified, causal structure believed to underlie conscious states. Meanwhile, global workspace theory posits that consciousness emerges when information is broadcast broadly across different functional modules, a feature that LLMs currently lack. These debates remain speculative, indicating that Nagel's challenge still looms large.

A deeper concern is the potential illusion of consciousness. Because advanced LLMs can use language about subjective states—discussing emotions, introspection, or even "wanting" certain outcomes—people may over-interpret these outputs as evidence of sentience. From an ethical standpoint, conflating fluent verbal performance with genuine subjective experience can lead to misplaced attributions of moral status or agency. Granting moral personhood to non-sentient systems, for instance, could skew responsibility and accountability (if an AI is "blamed" instead of the humans who developed or deployed it). Conversely, some futurists argue that if an AI's structure became complex, self-referential, and embodied in ways that approximate human cognition, a form of subjectivity might emerge—though this remains speculative and controversial.

Nagel's perspective thus acts as a cautionary guide. We should not conflate behavioral sophistication with phenomenal consciousness nor rush to treat generative AI as moral equals simply because they simulate human-like conversation. At the same time, it invites an open-minded stance regarding the future: as AI systems evolve—potentially integrating more embodied approaches, multimodal data, or hybrid cognitive architectures—the question of whether something like subjective experience might one day arise cannot be dismissed outright. For now, however, Nagel's question underscores the gulf between simulating mind and being a mind, setting ethical and philosophical boundaries around how we interpret and govern current AIs.

2.6 Integration of Contemporary Debates and Broader Perspectives

Beyond the four key philosophers surveyed above, a wide range of contemporary debates and interdisciplinary perspectives deepen our understanding of AI:

Posthumanism and AI

Posthumanist theories, such as Donna Haraway's "Cyborg Manifesto" (1985), challenge strict human/machine dichotomies by emphasizing the hybridity of human and technological systems. Rather than viewing AI as a mere tool, posthumanist viewpoints encourage seeing humans and AI as forming novel, hybrid agencies. These perspectives highlight ethical questions around human–machine symbiosis, prompting us to reconsider how we define identity, cognition, and even ethical responsibility when boundaries blur between organic and artificial intelligence.

Critical Theory and Sociotechnical Context

Scholars in critical theory and science and technology studies (STS) argue that AI systems reflect—and can perpetuate—existing social power structures. By examining the political, economic, and cultural contexts in which AI is developed and deployed, critical theorists expose how data, algorithms, and platforms can reproduce biases or concentrate power. Treating LLMs as neutral objects overlooks the broader social fabric of labor, infrastructure, and corporate interests behind them (Coeckelbergh, 2020). This perspective resonates with Wittgenstein's emphasis on social practices and Dennett's warning about anthropomorphizing systems, cautioning us to question not just how AI "thinks," but who controls its design and whose values it serves.

Anthropology and Sociolinguistics

Language usage varies by culture, community, and context. Anthropological and sociolinguistic research sheds light on how different cultures interpret AI-generated text, highlighting the potential for misunderstandings when AIs trained on predominantly Western, English-language corpora interact with users from other cultural backgrounds. This relates to Wittgenstein's "forms of life": each linguistic community has its own rules and assumptions. LLMs that lack direct exposure to diverse cultural norms can inadvertently perpetuate biases or fail to grasp the nuance of local idioms. Incorporating broader linguistic data and working with community stakeholders can partially mitigate these shortcomings.

Embodied Cognition and Cognitive Architectures

As noted earlier, embodied cognition frameworks argue that genuine understanding arises from the interplay between mind, body, and environment (Varela, Thompson & Rosch, 1991). In practical AI terms, researchers experiment with multimodal architectures—incorporating vision, audio, or robotics—so that an AI interacts physically with the world, potentially alleviating some of the symbol-grounding problem. Meanwhile, cognitive architectures (e.g., SOAR, ACT-R) model AI systems on cognitive modules like memory, attention, and executive control, aiming for a more holistic approach than text-only LLMs. These advances resonate with Lewis's scorekeeping notion—an AI with richer memory or sensorimotor feedback could update its "conversational score" more dynamically.

Cognitive Science and Neuroscience

Studies comparing LLMs' internal representations to patterns in the human brain suggest intriguing parallels in how linguistic information is processed. Yet critical gaps remain: humans rely on long-term memory, emotional salience, and embodied knowledge that purely text-based models lack. Neuroscientific insights into consciousness, such as Global Workspace Theory or Integrated Information Theory (IIT), may further clarify the line between complex computation and subjective awareness (Chalmers, 1996; Tononi, 2012). While no current evidence suggests LLMs achieve anything akin to phenomenological consciousness, ongoing research keeps the debate open, particularly with the rapid evolution of AI architectures.

Global Policy and Regulatory Frameworks

From a governance standpoint, AI ethics and policy discussions increasingly shape how generative AI is developed and deployed. The European Union's AI Act (passed in 2024), the UNESCO Recommendation on AI Ethics (2021), and the OECD AI Principles

(2019) seek to balance innovation with transparency, accountability, and human rights. These frameworks often reflect key philosophical concerns: Dennett's stance on not attributing unwarranted autonomy to AI, Nagel's caution about conflating sophistication with consciousness, and Wittgenstein's emphasis on socially situated meaning. In practice, this can manifest as transparency mandates (e.g., labeling AI-generated content), accountability mechanisms (ensuring human oversight), and risk assessments (classifying AI systems by potential harm). Such policy efforts aim to align AI development with shared ethical norms, though global consensus remains a work in progress.

Ethical Implications and Societal Impact

Across these perspectives, several ethical and societal themes emerge. AI can amplify biases, concentrate power in the hands of a few tech entities, and reshape labor markets. It can also enhance creativity, bridge language barriers, and support research. Philosophical insights help stakeholders navigate these tensions: acknowledging AI's limitations prevents overtrust (Dennett), understanding its lack of subjective experience (Nagel) helps define moral boundaries, and recognizing its reliance on human language games (Wittgenstein) can direct us to more inclusive and context-aware AI design. Ultimately, an interdisciplinary approach—integrating philosophy, cognitive science, anthropology, ethics, and policy—provides the richest toolkit for guiding AI's ongoing transformation of society.

In summary, contemporary discourse on AI is a tapestry of ideas from multiple fields. Classic philosophical frameworks articulate core conceptual distinctions, while emerging research in embodied cognition, critical theory, and policy reveals how AI systems operate within—and shape—living human cultures. This backdrop lays the foundation for the theoretical framework in the next section, uniting philosophical insights with practical imperatives for responsible AI.

3. Theoretical Framework

Having surveyed both classical philosophical sources and contemporary interdisciplinary perspectives, this section constructs a theoretical framework linking Wittgenstein's language games, Lewis's scorekeeping, Dennett's intentional stance, and Nagel's challenge on subjective experience to generative AI. The framework also draws on insights from embodied cognition, cognitive architectures, critical theory, and policy discussions, aiming to provide a comprehensive lens for understanding and improving AI interactions.

From a philosophical standpoint, each of the four thinkers offers a distinct angle:

- **Wittgenstein** underscores how language meaning is rooted in communal, rule-governed practices.
- **Lewis** emphasizes the dynamic maintenance and updating of conversational context.
- **Dennett** alerts us to the strategic but potentially misleading nature of treating AI as if it had beliefs or desires.

- **Nagel** highlights the gulf between behavioral simulation and genuine subjective experience.

When viewed through the prism of embodied cognition and social constructivism, these frameworks suggest that AI's language use is inseparable from the broader sociotechnical environments in which it is deployed. Meanwhile, practical considerations in AI design—ranging from explainability methods to memory-augmented neural networks—speak to how these philosophical insights can inform more coherent, reliable, and ethically grounded AI. Policy debates around transparency, fairness, and accountability supply a real-world backdrop, reinforcing the importance of aligning theoretical principles with governance structures.

In the subsections that follow, we examine how each philosophical perspective applies directly to LLMs and related AI systems. The resulting synthesis will inform the study's methodology, shape the empirical illustrations, and guide our discussion of results, counterarguments, and future directions.

3.1 Wittgenstein's Language Games and LLMs

Wittgenstein's concept of language games (1953) serves as a powerful starting point for analyzing how LLMs handle linguistic meaning. In Wittgenstein's view, the significance of words emerges from their use in the shared activities and forms of life of a community. Rules are not static entities but living conventions: they gain traction only through the social context in which they operate.

Application to LLMs

1. **Statistical Imitation vs. Communal Grounding**

LLMs learn language primarily by detecting patterns in vast text corpora, mimicking grammar, style, and context-specific usage. This can produce outputs that appear to follow human "rules." However, lacking direct participation in human activities—or an "embodied" form of life—LLMs only approximate rule-following. They do not originate language games based on shared praxis; they merely predict the next plausible token.

2. **Social Constructivism and Communal Feedback**

From a social constructivist standpoint, humans refine language by continuously negotiating meaning and validating each other's usage. By contrast, LLMs rely on static training sets; although they can occasionally be fine-tuned or updated, they do not co-evolve with a linguistic community in real time. Their "understanding" of words like love, freedom, or justice is thus fragile, removed from the living social rituals that embed these concepts in human life.

3. **Pragmatist Dimensions**

Pragmatism echoes Wittgenstein's emphasis on use by claiming that concepts derive meaning from their practical consequences. While LLMs can generate text that aligns with certain practical contexts (e.g., answering technical questions), they lack genuine involvement in any consequential activity—no personal goals, stakes, or lived feedback loops. This makes their "rule-following" essentially performative, rather than grounded in pragmatic engagement.

Improving AI through Wittgensteinian Insights

1. Multimodal and Interactive Learning

One way to bring AI closer to genuine "use" is to expand beyond text. Embodied cognition research suggests that coupling LLMs with sensors or robotic platforms could give AI systems rudimentary participation in shared activities—learning language in tandem with physical actions or visual feedback. Although this may never perfectly replicate human lived experiences, it reduces the abstract detachment of text-only training.

2. Community-Based Fine-Tuning

Encouraging AI systems to learn interactively from specific user communities (with tight feedback loops that correct misunderstandings) can approximate the iterative, communal aspects of language games. For example, domain experts can continually refine a specialized chatbot's vocabulary and interpretative rules, introducing elements of real-world negotiation into the model's training process.

3. Transparency and User Education

Users should be informed that an AI's "grasp" of words is at best an echo of aggregate text usage, not a deep or personal comprehension. This transparency can temper overreliance on AI "understanding" and encourage user vigilance when interpreting a chatbot's linguistic performance.

Taken together, the Wittgensteinian lens clarifies why LLMs, despite their fluency, frequently falter when language relies on shared life-forms or subtle pragmatic contexts. Efforts to embed AI more deeply in interactive, real-world practices—and to maintain user awareness of the AI's inherent limitations—are thus central to overcoming these shortcomings. The next sections extend this analysis by examining how Lewis's dynamic scorekeeping, Dennett's stance-based interpretation, and Nagel's consciousness critique further shape our understanding of generative AI.

Thought Experiment Recap: The Private AI Language

Imagine an AI that generates its own language without any human input. According to Wittgenstein's theory, such a language would be unintelligible and internally inconsistent, underscoring the vital role of social grounding in language development. Similarly, an LLM operating solely on data-driven mimicry—without real-world feedback—remains confined to syntactic reproduction, thereby highlighting the limitations of purely statistical approaches to semantics.

3.2 Lewis's Conversational Scorekeeping and Generative AI

David Lewis's "scorekeeping" theory of conversation (1979) provides a dynamic lens for understanding how context evolves during interaction—a concept crucial for explaining and improving the performance of large language models. In Lewis's view, conversational participants continuously track and update a shared "score," reflecting assumptions, referents, and presuppositions. Each new utterance can revise or clarify this context.

Applying Scorekeeping to LLMs

1. Local vs. Ongoing Context

LLMs replicate a limited version of scorekeeping by relying on an attention mechanism over a fixed window of tokens, allowing them to appear context-aware. Yet once relevant information falls outside this window, the "score" is essentially lost. In contrast, human participants maintain a far more robust and persistent record of the discussion, integrating updates into long-term memory.

2. Memory-Augmented Neural Networks

Recent work on memory-augmented neural networks and retrieval-based architectures aims to address LLMs' short memory. By storing conversation summaries or key entities in an external database, these systems can retrieve contextual facts even after they exceed the model's token limit. This helps the AI sustain coherent dialogue over extended interactions, aligning more closely with Lewis's notion of dynamically updated assumptions.

3. Scorekeeping in Complex Dialogues

Real-world conversations—such as legal consultations or multi-step planning—often involve sustained back-and-forth exchanges where context builds cumulatively. Without robust scorekeeping, an AI might contradict earlier statements or ignore critical user inputs, undermining trust and usability. Implementing structured context tracking can significantly enhance the AI's reliability in high-stakes or professional settings.

Practical Design Implications

1. Structured Summaries and State Tracking

Including a rolling summary of the conversation or a state graph that explicitly captures changing facts and user goals can help the AI maintain consistency. Such approaches resonate with Lewis's perspective by making context a "first-class citizen" in the AI's design.

2. Adaptive Dialogue Management

In multi-turn interactions, the system can periodically prompt itself (or be prompted by the user) to confirm or update the shared context ("score"). This active negotiation of assumptions mirrors human conversation, where speakers continually refine and align on what's been established.

3. Ethical and Policy Dimensions

From a governance standpoint, accountability and transparency often hinge on whether an AI can keep track of critical details—for instance, user consent or privacy preferences over long sessions. Policymakers may require systems to log conversation states or disclaim when prior context is no longer accessible, ensuring users are aware of the AI's memory limits and design constraints.

By weaving Lewis's scorekeeping theory into AI system development, we gain a blueprint for more stable, context-aware interactions. This shift from a static "snapshot" of context to a fluid, evolving conversation state—augmented by memory mechanisms—positions generative AI to function more like genuine conversational partners. As we move forward, Dennett's intentional stance adds another layer to this picture, clarifying both the benefits and pitfalls of treating such systems as if they truly grasp their conversational context.

Application Scenario: Multi-Turn Dialogue Adaptation

Imagine a legal consultation chatbot where the user initially explains their case in detail. Midway through the conversation, the user corrects a detail or introduces new evidence. The AI must then update its recommendations based on this revised information. Applying Lewis's scorekeeping theory, the chatbot would treat the new data as an update to the shared conversational context, potentially revising earlier statements accordingly. Implementing this effectively might require an architecture that actively revises a stored summary of the case rather than relying on static memory alone. This scenario underscores why dynamic context tracking is essential for professional applications of AI.

3.3 Dennett's Intentional Stance and the "As If" Agency of AI

Dennett's *intentional stance* (1989) illuminates the benefits and risks of interpreting AI systems as if they possess beliefs, desires, or intentions. While this interpretive approach simplifies prediction and interaction—much like we assume a chess engine "wants" to checkmate—treating an LLM in this manner can obscure crucial differences between genuine understanding and statistical text generation. Unlike past heuristics applied to simple machines or animals, LLMs exhibit linguistic fluency so convincingly that even experts may mistakenly attribute genuine agency—raising new ethical and regulatory challenges.

Adopting the Intentional Stance

1. User Experience and Interface Design

From a human–computer interaction perspective, it is pragmatically useful to address an AI as though it has a mind. Users may feel more comfortable asking a chatbot, "What do you think about…?" than issuing purely mechanical queries. The stance facilitates more natural conversation and can improve user satisfaction.

2. Interpretability and Debugging

Developers sometimes speak of what the model "knows" when diagnosing errors or fine-tuning performance. This informal language aids problem-solving: by attributing an "internal state" to the AI, engineers can conceptualize how misclassifications or incoherent replies arise. Yet they remain aware that these states are metaphorical rather than literal representations of beliefs.

Ethical Tensions and Anthropomorphism

1. Risks of Over-Attribution

Treating an LLM as sentient may encourage anthropomorphism, where users mistakenly attribute emotions, intentions, or moral standing to a system that lacks genuine experience or values. This can engender misplaced trust or emotional bonds, potentially harming vulnerable users who turn to AI for companionship or guidance in critical situations.

2. Accountability and Power

Critical theorists warn that the intentional stance can mask the human labor, corporate power, and social contexts that shape AI behavior. If something "goes wrong," blaming "the AI" can deflect accountability from developers or institutions. Recognizing the AI's as if agency—while keeping real human agency central—helps maintain appropriate responsibility structures.

Policy and Governance Perspectives

1. Transparency Requirements

Policy proposals often emphasize that AI systems should explicitly clarify they are non-human, preventing user confusion about the source of decisions or advice. This might include disclaimers ("I am an AI assistant and do not have personal opinions"), or design features (e.g., robotic avatars) that visually distinguish the system from a human agent.

2. User Education

In regulated domains—like healthcare or finance—educational materials can caution users: "AI responses are heuristic approximations, not licensed professional advice." This nudges people to adopt an intentional stance only in limited, functional ways, rather than fully anthropomorphizing the system.

3. Preventing Ethical and Legal Loopholes

Legislators are increasingly attentive to how “autonomous” AI is portrayed. Granting AI legal personhood or ascribing it moral status prematurely could create legal gray areas, undermining clear lines of liability. Dennett’s stance supports a more measured approach, where AI is treated as if it has intentions only to the extent that such treatment aids human aims—without absolving human overseers or developers of responsibility.

In sum, Dennett’s intentional stance offers a pragmatic framework for designing and interacting with generative AI, yet it must be wielded thoughtfully. By recalling that the stance is a useful fiction, we avoid conflating linguistic fluency with genuine understanding or volition—a confusion that could undermine ethical responsibility, regulatory clarity, and user well-being.

3.4 Nagel’s Challenge: Subjective Experience and the Limits of Simulation

Thomas Nagel’s classic query—“What is it like to be a bat?” (1974)—highlights a fundamental puzzle: even exhaustive knowledge of a being’s physical or functional processes does not necessarily reveal its subjective experience. Applying this to large language models, we confront the possibility that no matter how adeptly AI mimics human conversation, there may be “nothing it is like” to be that AI. This gap between outward behavior and subjective awareness is central to what philosophers call the hard problem of consciousness.

Simulation vs. Consciousness

1. Behavioral Sophistication

Modern LLMs can simulate introspection—discussing desires, fears, or inner thoughts—yet these outputs likely reflect patterns in text rather than genuine self-awareness. Nagel’s point underscores that generating talk about mental states does not entail having those states, a distinction that even advanced AI architectures may never bridge purely through language-based processing.

2. Illusions of Consciousness

Precisely because LLMs are so adept at producing natural language, users may ascribe consciousness or emotions to them. This illusion of consciousness can arise when the AI convincingly references its own “thoughts” or “feelings.” From Nagel’s perspective, such attributions rest on superficial clues rather than the presence of an inner subjective viewpoint.

Ethical and Societal Implications

1. Moral Status and Responsibility

If current AI systems lack any subjective experience, granting them moral personhood is premature. Doing so could dilute genuine moral responsibilities that belong to humans

—designers, companies, policymakers—and create legal loopholes by attributing accountability to an entity with no capacity for actual suffering or intent.

2. Conflation of Fluency and Sentience

In social or therapeutic contexts, an LLM might produce empathetic-seeming replies that users find comforting. While beneficial in certain scenarios, overstating the system's "empathy" can lead to emotional harm if users come to believe they are engaging with a sentient companion. Policymakers and ethicists emphasize transparency to prevent confusion and protect users who might be vulnerable.

3. Future Directions: Embodiment and Hybrid Approaches

Some researchers speculate that if an AI architecture integrated enough sensorimotor grounding, self-referential loops, and memory unification, a form of consciousness could emerge. This remains highly speculative and controversial. Nagel's skepticism reminds us that no matter how advanced the hardware and software become, subjective experience might lie outside the reach of purely functional replication. At the very least, demonstrating anything akin to phenomenal consciousness in AI would require far more than a large language model predicting tokens.

Positioning Nagel in the Framework

Nagel's challenge complements Wittgenstein, Lewis, and Dennett by anchoring the discussion in the philosophical limits of simulation. While Wittgenstein and Lewis focus on meaning's social and contextual dimensions, and Dennett examines interpretive stances, Nagel highlights that subjective feeling is not captured by outward behavior alone. This perspective helps temper any rush to anthropomorphize AI or to treat conversational fluency as proof of "mind."

Overall, Nagel's view situates generative AI within a broader philosophical conversation about the nature of consciousness itself—reminding developers, users, and regulators that even the most sophisticated language generation need not imply real awareness. This distinction is vital for ethical frameworks, preventing misguided assumptions about AI rights or agency, and reinforcing the need for human accountability in AI governance.

Philosophical Reflections

If an AI were ever to insist that it is conscious and plead for ethical consideration, humanity would face a Nagel-inspired dilemma. We would then need to determine how much credence to give to the AI's self-reported experience. While this remains a speculative thought experiment for now, it underscores the critical distinction between simulating a mind and truly possessing one—a difference that is not merely theoretical but could have significant practical and moral ramifications. Our current approach should be one of caution, recognizing that extraordinary claims of AI consciousness require extraordinary evidence.

3.5 Synthesis: A Multi-Layered Theoretical Lens for Generative AI

Bringing together Wittgenstein's language games, Lewis's scorekeeping, Dennett's intentional stance, and Nagel's challenge to subjective experience, we arrive at a layered theoretical framework that illuminates both the capabilities and inherent limits of generative AI.

1. Language as Social Practice (Wittgenstein)

Wittgenstein underscores that meaning emerges from communal rule-following embedded in a form of life. For LLMs, this highlights the gap between statistically learned patterns and genuine participation in the social and embodied contexts that shape linguistic meaning.

2. Dynamic Context Management (Lewis)

Lewis's conversational scorekeeping shows us that language is not merely a matter of static definitions but an ongoing negotiation of shared assumptions. LLMs approximate this through attention mechanisms and text windows, yet they often fail to maintain context over long interactions. Memory-augmented or retrieval-based architectures can partially address this, but the design must be deliberate and transparent to users.

3. Pragmatic Utility and Anthropomorphism (Dennett)

Dennett's intentional stance explains why many users find it natural to treat AI "as if" it has beliefs or desires: it simplifies interaction and offers predictive power. However, this stance also risks over-ascribing human-like agency. Critical theory perspectives emphasize that anthropomorphism can obscure the real socio-technical structures behind AI—potentially diverting accountability or reinforcing power imbalances.

4. The Hard Problem of Consciousness (Nagel)

Even if an AI perfectly simulates human conversation, Nagel's challenge suggests that actual subjective experience may remain out of reach. Fluency and self-referential talk do not guarantee phenomenal consciousness or moral standing. This distinction becomes ethically crucial in avoiding premature ascriptions of rights or emotional capacities to AI systems.

Interdisciplinary Threads

- **Embodied Cognition and Cognitive Architectures:** Insights from embodied cognition suggest grounding AI in sensorimotor contexts, potentially shrinking the gap noted by Wittgenstein. Meanwhile, cognitive architectures provide a structural blueprint for integrating memory, inference, and long-term goals—complementing Lewis's emphasis on context and Dennett's focus on functional stances.
- **Pragmatism, Social Constructivism, and Policy**: Pragmatist and social-constructivist views stress that language and meaning evolve through concrete use and

communal feedback. Translating these insights into policy could mean requiring ongoing human oversight, iterative fine-tuning with diverse communities, and transparent disclosure of AI's design and limitations.

- **Ethical Implications:** Across all four perspectives, there is consensus that conflating AI simulation with genuine understanding or consciousness can lead to social and ethical pitfalls, such as erosion of accountability, user manipulation, or misplaced trust. Effective governance must address these pitfalls with measures like transparency mandates, user education, and robust mechanisms for redress.

Looking Ahead

This multi-perspective framework sets the stage for examining how AI developers, regulators, and users can address current challenges: from hallucinations and context loss to ethical dilemmas around autonomy and anthropomorphism. Sections 4 through 6 apply these concepts methodologically and empirically, discussing how real-world AI deployments fare against the criteria established by these philosophical lenses—and how we might design or regulate AI systems to better align with human values and societal needs. discussion.

4. Methodology

4.1 Research Design

This study employs a mixed-method research design that integrates philosophical analysis, empirical illustration, and interdisciplinary validation. The methodology unfolds in several interrelated phases:

1. Philosophical Analysis

A close reading of seminal texts by Wittgenstein, Lewis, Dennett, and Nagel establishes the foundational concepts. This phase is enriched by exploring related perspectives—such as embodied cognition, pragmatism, and social constructivism—to clarify how these ideas inform our understanding of AI.

2. Theoretical Mapping

Insights from philosophical analysis are then systematically mapped onto the operational features of generative AI systems. For instance, concepts like language games, scorekeeping, and the intentional stance are compared to mechanisms such as attention windows, memory constraints, and output generation in large language models. This mapping process highlights both the capabilities and limitations of current AI systems.

3. Empirical Integration and Interdisciplinary Case Studies

To ground the analysis in real-world contexts, the study incorporates illustrative case studies from diverse domains (e.g., multi-turn dialogues in legal consultations or customer service applications). These case studies serve to test and refine the theoretical framework by exposing it to practical challenges. In addition, survey research is proposed to capture user experiences and perceptions regarding AI's conversational coherence and the tendency to anthropomorphize its outputs. This dual approach ensures that the conceptual insights are validated against observable phenomena.

4. Validation Strategy and Experimental Proposals

The research design outlines potential controlled experiments aimed at evaluating key components of the framework. For example, experiments may measure how well an AI maintains context over extended dialogues (using consistency or "scorekeeping" metrics) or assess changes in user trust when informed about the AI's limitations. These quantitative and qualitative measures will provide empirical evidence to support or refine the proposed theoretical insights.

5. Interdisciplinary Collaboration

Recognizing the complexity of AI, this design explicitly calls for collaboration among philosophers, cognitive scientists, AI developers, and policy experts. Such interdisciplinary engagement ensures that theoretical models are continuously refined by practical feedback, and empirical findings are interpreted in light of broader ethical and societal considerations.

By combining these diverse methods, the research design aims to produce a robust, empirically informed theoretical framework. This approach not only advances our conceptual understanding of generative AI but also lays the groundwork for developing practical strategies and policy guidelines that address its inherent limitations and ethical challenges.

4.2 Data Sources

Our study draws on a diverse range of sources, spanning philosophy, cognitive science, AI technical research, empirical observations, and policy documentation. These include:

- **Primary Philosophical Texts**

Foundational works by Wittgenstein, Lewis, Dennett, and Nagel serve as the backbone for our conceptual framework. Detailed exegeses of these texts provide the core philosophical concepts that are later mapped onto AI systems.

- **Secondary Literature in Philosophy and Cognitive Science**

Scholarly analyses that extend classical ideas—such as embodied cognition, pragmatism, and social constructivism—offer critical insights into the limitations and potential of generative AI. Key works include discussions by Clark (2008), Varela et al. (1991), and contemporary critiques that debate the "stochastic parrots" notion.

- **Technical AI Literature**

Research papers on transformer architectures (e.g., Vaswani et al., 2017) and large language models (e.g., Brown et al., 2020) provide the technical context and operational details of AI systems. Additional sources on memory-augmented networks and explainable AI contribute to our discussion on context management and transparency.

- **Empirical and User-Generated Data**

Illustrative case studies from real-world interactions—such as chatbot transcripts, customer service dialogues, and online forum discussions—demonstrate practical challenges in maintaining context and avoiding anthropomorphism. Survey research and ethnographic studies (from published sources or planned future research) further inform our understanding of how users perceive and interact with AI.

- **Policy Documents and Regulatory Frameworks**

Documents such as the EU AI Act, UNESCO's Recommendation on AI Ethics (2021), and the OECD AI Principles provide a regulatory and ethical backdrop. These sources help align our theoretical insights with contemporary governance challenges and public policy debates.

4.3 Analytical Approach

Our analysis employs a structured, multi-step approach to ensure that each element of the theoretical framework is examined thoroughly and from diverse disciplinary perspectives:

1. Concept Mapping

We begin by systematically aligning key philosophical concepts—such as Wittgenstein's language games, Lewis's scorekeeping, Dennett's intentional stance, and Nagel's challenge to consciousness—with corresponding features in generative AI. For example, we map:

- Rules of language use (Wittgenstein) to the statistical patterns learned from text.
- Dynamic context updating (Lewis) to the limited attention mechanism and memory constraints in LLMs.
- Attribution of beliefs (Dennett) to the heuristic strategies used in interpreting AI outputs.
- The gap between simulation and experience (Nagel) to the absence of subjective awareness in AI.

This mapping highlights both parallels and critical gaps, providing concrete footholds for further analysis.

2. Cross-Disciplinary Correlation

To validate and nuance our claims, we corroborate philosophical assertions with evidence from cognitive science, neuroscience, and technical research. For instance, while the philosophical stance asserts that LLMs lack genuine understanding, empirical studies showing semantic errors or context loss support this claim. Similarly, documented user behaviors—such as naming personal devices or treating chatbots as companions—help illustrate tendencies toward anthropomorphism, reinforcing critical theory perspectives. This step ensures that our conclusions are not isolated but are supported by multiple fields.

3. Critical Evaluation and Counterarguments

We maintain a critical stance toward both AI capabilities and the philosophical theories. By identifying where LLMs successfully mimic human communication and where they fall short (e.g., in sustaining long-term context or exhibiting genuine empathy), we expose tensions that demand further inquiry. We also incorporate counterarguments—such as the possibility that extensive training data might approximate aspects of human context—to provide a balanced view and highlight areas for future research.

4. Scenario Analysis and Thought Experiments

Practical scenarios and thought experiments are employed to illustrate how the theoretical framework applies in real-world contexts. For example, we simulate a conversation where an AI misinterprets culturally specific humor, pinpointing how a lack of embodied context (per Wittgenstein) and dynamic memory (per Lewis) contributes to the failure. In another scenario, we analyze a case where a user's overreliance on the AI's seemingly empathetic responses leads to emotional distress, underscoring the ethical implications discussed in Dennett's and Nagel's sections. These narrative analyses provide concrete examples that guide practical design and policy recommendations.

5. Interdisciplinary Peer Review (Simulated)

Although a formal peer review was not conducted, we simulated interdisciplinary feedback by posing questions from the perspectives of cognitive scientists, AI developers, and policymakers. This iterative process helped us refine our analysis, ensuring that it remains relevant across domains and addresses both theoretical and practical concerns.

By integrating these steps, our analytical approach offers a rigorous, balanced, and actionable examination of generative AI. This multifaceted evaluation lays the groundwork for developing strategies to improve AI system design, enhance user interaction, and inform policy decisions—all while maintaining clear ethical and philosophical boundaries.

Before transitioning to our findings, it is important to acknowledge some methodological limitations. Our case studies are illustrative rather than statistically representative, and several forward-looking proposals—such as methods for testing consciousness or enhancing context management—remain speculative. We have made an effort to clearly label hypothetical scenarios and to distinguish speculative ideas from those supported by empirical evidence. These limitations serve as a reminder that while our framework offers valuable insights, further research is necessary to fully validate and refine our claims. The next sections detail our findings and discuss how they can inform both future research and current AI practices.

5. Results and Findings

Applying the above framework to generative AI yields several key findings, which we organize by the philosophical lens and then synthesize into practical implications.

5.1 Wittgenstein and LLMs: The Importance of Communal Context

Observations:

LLMs generate grammatically correct and contextually appropriate text by learning statistical patterns from vast text corpora. However, they lack the communal, embodied context that gives human language its rich meaning. For instance, while an LLM may mimic conversational norms, it often misinterprets subtleties like sarcasm or culturally specific idioms because it does not participate in the shared "language games" that naturally ground human communication. In technical terms, the AI's world model is limited to text correlations, making its "understanding" brittle when it encounters input that falls outside its training distribution or requires lived, cultural experience.

Integrated Perspectives:

Wittgenstein's emphasis on language as a social practice highlights the gap between mere statistical mimicry and genuine linguistic participation. This gap is further underscored by embodied cognition theory, which posits that true understanding arises from real-time interactions between body, mind, and environment—a dimension that text-only models inherently lack. Although LLMs can simulate context by adhering to learned patterns, they remain detached from the dynamic, embodied interactions that inform human language use.

At the same time, the intentional stance (Dennett) explains why we often perceive AI outputs as meaningful: we naturally adopt a heuristic that treats coherent language as evidence of understanding. However, Nagel's challenge reminds us that no matter how eloquent the AI, there is always a gap between simulated responses and lived experience. For example, an LLM might describe a sunny day with poetic flair, yet it has never experienced sunlight or warmth; its description is an echo of aggregated human texts, not a reflection of sensory reality.

Implications and Recommendations:

- **Interactive Learning and Cultural Fine-Tuning:**

Incorporating interactive learning protocols—where the AI solicits clarification and receives iterative feedback—can help bridge the gap between static text patterns and the fluid nature of human language. Additionally, culturally aware fine-tuning can expand an AI’s repository of “language games,” enabling it to better handle diverse linguistic and contextual cues.

- **Multimodal Integration:**

Moving beyond text-only training, integrating multimodal inputs (such as images, sound, or sensor data) could provide AI systems with a rudimentary form of embodied context. This approach may not fully replicate human lived experience but could reduce the abstract detachment inherent in current LLMs.

- **User Education and Transparency:**

Users should be informed that an AI’s fluent language is an artifact of pattern recognition, not evidence of true comprehension. Transparent disclosures about the AI’s limitations can help temper expectations and encourage more explicit communication during interactions.

Conclusion:

The Wittgensteinian lens reveals that while LLMs are remarkably adept at reproducing human-like language, their lack of communal grounding and embodied experience leads to occasional misfires in meaning. To enhance robustness in AI interaction, it is critical to pursue strategies that infuse context—through interactive learning, culturally informed training, and multimodal data—while ensuring that users remain aware of the AI’s inherent limitations. This understanding sets the stage for developing more effective, transparent, and ethically grounded AI systems.

5.2 Lewis and LLMs: Contextual Scorekeeping in Conversations

Observations:

Analysis of real-world AI dialogues reveals that modern LLMs achieve a degree of context sensitivity by relying on attention mechanisms over a limited token window. This approach allows them to maintain a form of “scorekeeping,” tracking recent exchanges and resolving pronouns or references within that scope. However, this scorekeeping is inherently local: once critical details fall outside the model’s immediate context, consistency can degrade[1]. This leads to observable challenges in multi-turn dialogues, where the AI might inadvertently contradict itself or overlook earlier conversation elements.

Integrated Perspectives:

[1] Recent research on FILM-7B (An et al., 2024) demonstrates empirical techniques—such as information-intensive long-context training—that significantly improve LLMs’ ability to maintain context across longer dialogues and document structures. While technically impressive, these advances remain consistent with the argument made here: they enhance simulated coherence

Lewis's theory emphasizes that conversational meaning is dynamically constructed and continually updated through shared assumptions. Human interlocutors seamlessly integrate long-term context and implicit knowledge far beyond immediate utterances—a process that LLMs currently approximate only imperfectly. Cognitive pragmatics further supports this view by underscoring the importance of adaptive context management, where speakers constantly revise and negotiate meaning based on new input.

To address these limitations, recent research in memory-augmented neural networks and retrieval-based architectures is promising. By incorporating external memory modules or structured state representations, AI systems can retain and retrieve key details from earlier in a conversation. These approaches not only extend the effective context window but also align more closely with the human ability to update a "conversational score" over long interactions.

Implications and Recommendations:

- **Structured Summaries:** Implementing periodic summaries or state graphs that encapsulate the evolving conversation can help the AI maintain consistency across multiple turns. This explicit scorekeeping mirrors how humans often recap key points during long discussions.

- **Adaptive Retrieval:** Retrieval-based techniques—where the system can query an external database or memory store for relevant past information—offer a practical solution for sustaining context. These methods can be especially valuable in scenarios such as legal consultations or customer service, where retaining a comprehensive context is critical.

- **Ethical and Policy Considerations:** From a governance standpoint, ensuring transparency about an AI's memory limitations is vital. Users should be informed that, beyond a certain point, earlier conversation details might not be fully retained by the system. This disclosure can help manage expectations and maintain trust, particularly in sensitive applications where context integrity is essential.

Conclusion:

Lewis's scorekeeping theory illuminates a fundamental challenge for generative AI: the need for dynamic, enduring context management. While current LLMs achieve a basic form of context tracking, integrating cognitive pragmatic insights and advanced memory architectures can substantially improve coherence and user satisfaction. By refining AI's ability to maintain a comprehensive conversational score, we can move closer to interactions that mirror the fluid, context-rich exchanges of human dialogue.

but do not establish socially embedded, rule-governed understanding in the Wittgensteinian or Lewisian sense.

5.3 Dennett and LLMs: The Utility and Risks of the Intentional Stance

Observations:

Users frequently interact with large language models as if these systems possess beliefs, desires, and intentions. When a chatbot responds with “I think…” or “I recommend…”, it naturally prompts users to adopt an intentional stance—treating the AI as if it were a conscious agent. This approach simplifies interaction, making the dialogue feel more natural and relatable. However, despite the conversational fluency, LLMs are ultimately advanced pattern-matching algorithms without genuine internal states or subjective experience.

Integrated Perspectives:

Dennett’s intentional stance is a pragmatic tool: it allows both users and developers to interpret and predict AI behavior without necessitating actual mental states. This heuristic is useful in everyday interactions and can even guide debugging or system improvement. Yet, a critical implication is the risk of over-anthropomorphizing AI.

- **Ethical Implications**: Overattributing human-like qualities may lead users to place unwarranted trust in AI systems, potentially causing them to divulge sensitive information or rely on the AI in contexts that require human judgment.
- **Critical Theory Concerns**: From a broader societal perspective, treating AI as if it possesses intentions can obscure the human and institutional labor behind these technologies, deflecting accountability and reinforcing existing power structures.

Implications and Recommendations:

- **Design Transparency:** AI interfaces should include clear, consistent cues that remind users of the system’s non-human nature (e.g., disclaimers like “I am an AI assistant, not a person”). This transparency helps mitigate risks of emotional overreliance and maintains realistic expectations about the AI’s capabilities.
- **User Education:** Educating users about the limitations of the intentional stance—emphasizing that attributing true understanding to AI is a heuristic shortcut—can empower them to critically assess AI outputs, especially in high-stakes applications like healthcare or legal advice.
- **Policy Measures:** Regulators and industry standards should mandate that AI systems clearly disclose their computational basis and lack of genuine intentionality. Such measures can prevent the misuse of anthropomorphic designs to manipulate user trust or evade accountability.
- **Balanced Interface Design:** While anthropomorphic elements (such as friendly language or avatars) may enhance user engagement, they must be carefully calibrated to avoid creating illusions of sentience. The goal is to strike a balance between intuitive usability and honest representation of the AI’s limitations.

Conclusion:

Dennett's intentional stance is a double-edged sword in the realm of generative AI. On one hand, it facilitates smoother, more relatable interactions by allowing users to navigate complex conversations with an "as if" understanding. On the other hand, it poses significant ethical and societal challenges if taken too literally, leading to overtrust and a misallocation of responsibility. By integrating transparency, user education, and thoughtful policy guidelines, designers and regulators can harness the benefits of the intentional stance while safeguarding against its risks. This balanced approach is essential for ensuring that AI systems serve as reliable tools without blurring the critical distinction between simulation and genuine agency.

6. Discussion

6.1 Addressing Counterarguments and Emerging Perspectives

While the theoretical framework presented thus far offers a comprehensive lens for understanding generative AI, it is essential to engage with several counterarguments and emerging perspectives that both challenge and refine our conclusions.

"LLMs Do Understand" – The Optimists' Argument

Some researchers argue that large language models capture significant aspects of meaning—demonstrating capabilities in reasoning, creative composition, and even the explanation of humor—that suggest a form of understanding. These optimists argue that with sufficient training data and computational power, LLMs might approximate aspects of human cognition, appearing to 'know' the implicit rules of language through statistical inference. From a cognitive science standpoint, this view is supported by experiments showing that LLMs can sometimes generalize or apply concepts in novel ways, even if that understanding remains shallow compared to human experiential learning.

Anthropomorphism versus Denialism

Critics also caution against two extremes. On one hand, over-attributing human-like qualities to AI (anthropomorphism) can lead to misplaced trust and emotional overreliance, with users ascribing moral or social agency to systems that merely simulate intelligent behavior. On the other hand, some suggest a strict functionalist view that dismisses any semblance of understanding as irrelevant. We propose a middle path: while LLMs exhibit impressive linguistic coherence and context handling, their "understanding" is fundamentally different from human cognition—a nuance that must be clearly communicated to avoid both undue fear and overtrust.

Emergent Abilities and AGI Hype

Recent observations of emergent capabilities in larger models have fueled speculation that qualitative leaps in intelligence—and even the onset of conscious-like features—may eventually occur. For example, the claim that GPT-4 exhibits "sparks of AGI" raises provocative questions about scalability and complexity. However, despite such advances, issues like hallucinations, context degradation, and the lack of embodied interaction persist. These challenges echo the concerns raised by Wittgenstein, Lewis, Dennett, and Nagel: simulation of understanding does not equate to genuine, human-like cognition. Until AI systems can integrate sensorimotor feedback or develop robust long-term memory in ways comparable to human experience, the gap between functional mimicry and true understanding remains substantial.

Alternate Philosophical Frameworks and Societal Considerations

Beyond the core four philosophers, alternate frameworks—such as John Searle's Chinese Room argument or posthumanist critiques—offer contrasting perspectives on AI's capabilities and limitations. Searle's perspective warns against conflating syntactic processing with semantic understanding, while posthumanist voices challenge us to reframe our conceptions of intelligence altogether. Additionally, ethical critiques from critical theory emphasize that the framing of AI as an autonomous agent may obscure the socio-technical systems, labor, and power structures behind its creation. Recognizing these broader implications is essential for guiding responsible development and governance.

Implications for Policy and Future Research

The diversity of perspectives underscores the need for ongoing interdisciplinary dialogue. Future research should empirically test the limits of AI "understanding" through controlled experiments—such as assessing context retention and adaptability—and evaluate the impact of user education on mitigating anthropomorphism. In parallel, policymakers must craft transparent regulatory frameworks that clearly communicate the capabilities and limitations of generative AI, ensuring accountability and safeguarding against the misuse of technology.

Conclusion of the Counterargument Discussion

In sum, while optimistic views highlight the impressive achievements of LLMs and suggest that statistical learning may approximate aspects of human cognition, a cautious appraisal—guided by classical philosophical insights—reminds us that current AI remains fundamentally different from human minds. By addressing these counterarguments head on, we refine our theoretical framework and ensure that subsequent recommendations for AI design, user interaction, and policy are both nuanced and robust.

6.2 Enhancing Methodological Rigor and Future Research

To further validate and refine our theoretical framework, it is essential to strengthen the empirical and interdisciplinary aspects of our research. We propose several avenues for future work that address the limitations of current studies and advance our understanding of generative AI.

Controlled Experiments with AI Systems

Future research should implement controlled experiments to rigorously test how well AI systems maintain context, simulate intentionality, and manage conversational dynamics. For example, experiments could:

- Evaluate context retention by comparing AI performance in multi-turn dialogues with and without memory augmentation.
- Test the effectiveness of interactive learning protocols by measuring improvements in user satisfaction and error reduction when the AI solicits clarifications.
- Assess whether specific design interventions—such as structured summaries or adaptive retrieval mechanisms—lead to measurable gains in conversational coherence.

Interdisciplinary Collaboration and Case Studies

A robust research agenda requires collaboration across disciplines:

- Interdisciplinary Research Teams: Bringing together philosophers, cognitive scientists, AI engineers, and policy experts will help design experiments that are both conceptually sound and practically relevant.
- Empirical Case Studies: Detailed case studies drawn from real-world applications (e.g., legal consultations, customer service interactions, or educational tutoring sessions) can illuminate the practical challenges and benefits of applying our theoretical framework. These case studies should include qualitative and quantitative analyses, capturing both performance metrics and user feedback.

Cross-Cultural and Global Research

Language use and contextual understanding vary widely across cultures:

- Conduct cross-cultural studies to determine whether AI systems trained predominantly on Western data can adapt to diverse linguistic and cultural contexts.
- Evaluate how different communities interact with AI and whether culturally tailored training or feedback mechanisms improve performance and user trust.

Longitudinal and Iterative Evaluations

AI systems are not static—they evolve over time:

- Longitudinal studies should track the performance of AI systems in real-world settings over extended periods. Observing how user interactions and system performance evolve can reveal new insights into context management and the durability of the "scorekeeping" mechanism.
- Iterative evaluations of ethical and policy outcomes are also crucial. For instance, assessing the impact of transparency mandates or user education initiatives on mitigating over-anthropomorphism can help refine both technical designs and regulatory guidelines.

Evaluation of Ethical and Policy Outcomes

As AI systems become more integrated into society, their ethical implications must be rigorously evaluated:

- Studies should examine whether clear disclosures about AI limitations affect user trust and reliance.
- Policy-oriented research can assess how current regulatory frameworks (such as the EU AI Act or OECD AI Principles) influence AI development practices, accountability structures, and user protection measures.

By adopting these research strategies, future studies can provide a more empirically grounded and contextually sensitive understanding of generative AI. Enhancing methodological rigor through interdisciplinary collaboration and comprehensive evaluation will not only validate our theoretical framework but also inform the development of more robust, transparent, and ethically sound AI systems.y too high, affecting potentially billions of lives in everyday interactions and societal structures.

6.3 Broader Implications for Ethics, Policy, and Society

As generative AI systems become more pervasive, the ethical, policy, and societal implications of their deployment must be thoroughly considered. The philosophical perspectives discussed throughout this paper not only guide our understanding of AI's capabilities and limitations but also underscore critical responsibilities for developers, users, and regulators.

Ethical Considerations

- **Transparency and Disclosure**:

AI systems should clearly indicate that they are non-human entities. Disclosures such as "I am an AI assistant" help prevent the misattribution of human qualities and ensure that users are aware of the system's limitations. Transparency about data sources, training methods, and potential biases is also essential to foster trust and allow for informed scrutiny of AI behavior.

- **Avoiding Misplaced Trust**:

Over-anthropomorphizing AI may lead users to rely on systems in contexts where human judgment is necessary—such as medical, legal, or mental health scenarios. Ethical guidelines must caution against overdependence on AI outputs, stressing that these systems, despite their fluency, remain fundamentally different from human agents.

- **Moral and Legal Accountability:**

Since generative AI lacks genuine intentionality or consciousness, responsibility for its actions must remain with the human creators, developers, and operators. Ethical frameworks and legal policies should explicitly assign accountability, preventing the diversion of responsibility to the AI itself. While no current AI system possesses the qualities necessary for moral consideration, continued advancements may require an evolving ethical framework.

Policy and Regulatory Frameworks

- **Global and Uniform Standards:**

As noted in initiatives like the EU AI Act, UNESCO's AI Ethics Recommendation, and the OECD AI Principles, policymakers are working toward common regulatory frameworks that emphasize fairness, accountability, and transparency. A coordinated international approach can help ensure that AI systems meet minimum ethical standards, irrespective of regional differences.

- **Accountability Mechanisms:**

Regulations might require the implementation of audit trails or logging systems that document AI decision-making processes. Such measures not only aid in diagnosing errors and biases but also ensure that any misuse of AI can be traced back to responsible parties.

- **Incorporating Multidisciplinary Insights:**

Policies should be informed by interdisciplinary research that incorporates philosophical insights, cognitive science findings, and technical evaluations. By bridging these fields, policymakers can craft regulations that are both practically enforceable and philosophically sound.

Public Education and Societal Impact

- **User Awareness and Literacy:**

Public education campaigns can help demystify AI technologies and inform users about their strengths and limitations. Understanding that AI "understanding" is an artifact of statistical processing—rather than genuine comprehension—can prevent undue trust and potential harm.

- **Cultural Sensitivity and Inclusivity:**

As AI systems interact with diverse global populations, it is vital to address cultural biases in training data and to design systems that are sensitive to local linguistic and social norms. This inclusive approach not only enhances system performance but also promotes fairness and respect for diversity.

- **Balancing Innovation and Regulation:**

While robust regulation is essential for ensuring ethical AI development, it should be carefully balanced to avoid stifling innovation. Transparent standards, iterative policy development, and active stakeholder engagement can help achieve a dynamic equilibrium between technological progress and societal well-being.

Conclusion of Broader Implications

The integration of ethical, policy, and societal considerations into the development and deployment of generative AI is critical. By drawing on philosophical insights—ranging from the communal grounding of language to the limits of simulated consciousness—we gain a nuanced perspective on both the potentials and the perils of AI. Ultimately, transparent practices, rigorous accountability, and informed public discourse will be indispensable for harnessing AI's benefits while safeguarding human values and social equity.

7. Conclusion

This paper has developed an interdisciplinary framework that bridges classical philosophical theories with contemporary insights from cognitive science, neuroscience, technical AI research, and policy analysis. By synthesizing Wittgenstein's language games, Lewis's conversational scorekeeping, Dennett's intentional stance, and Nagel's challenge to subjective experience, we have constructed a multi-layered lens through which to understand the capabilities and limitations of generative AI.

Our analysis reveals that, while large language models can generate text that is contextually coherent and often surprisingly human-like, they remain fundamentally detached from the embodied, communal, and experiential dimensions of human language. While advancements in memory-augmented AI and multimodal learning are improving LLM capabilities, these systems remain fundamentally statistical pattern predictors rather than intentional agents. This gap is evident in their reliance on statistical patterns rather than dynamic, lived interactions, as well as in the potential risks associated with over-anthropomorphizing these systems.

The interdisciplinary perspective advanced here not only illuminates the theoretical challenges—such as context loss, ethical concerns, and the hard problem of consciousness—but also points toward practical strategies for improvement. Integrating memory-augmented architectures, multimodal data, and community-based fine-tuning can help mitigate current limitations. At the same time, user education and transparent design are critical to ensuring that AI systems are used responsibly, with clear acknowledgment of their non-human nature.

From a policy and societal standpoint, our framework underscores the need for robust, internationally coordinated regulations that promote transparency, accountability, and fairness. As generative AI continues to influence diverse sectors—from healthcare and education to creative industries—policymakers must work in close collaboration with researchers and practitioners to craft standards that safeguard user interests while fostering innovation.

In sum, while the sophistication of generative AI invites us to consider its potential as a transformative tool, our findings reaffirm that these systems are, at their core, simulations of human communication rather than replacements for human understanding. Future research should build on this interdisciplinary foundation, continuously refining both the technical capabilities of AI and the ethical, legal, and societal frameworks that guide its use. By doing so, we can harness AI's benefits while ensuring that its deployment remains ethically responsible, socially beneficial, and aligned with human values.

Living Document Notice: This work will be periodically revisited as generative AI technology and its societal implications continue to evolve. New findings, whether empirical breakthroughs or theoretical critiques, will be integrated to refine the framework and recommendations. The aim is to maintain an up-to-date resource that bridges enduring philosophical questions with the state-of-the-art in AI.

As AI continues evolving, engaging with its philosophical and ethical implications is no longer optional—it is essential. By integrating interdisciplinary insights, we can design AI systems that are not just powerful but also aligned with human values. This paper is an invitation to rethink how we conceptualize, govern, and interact with generative AI—not as mystical entities or autonomous minds, but as tools that reflect and reshape our collective intelligence.

This document is periodically updated, as noted. Version history: v1.21.2 (April 2025 – Initial publication).[2]

Ethical and Permissions:

- **Data Usage:** This study did not utilize or disclose any private personal data. Illustrative examples were drawn from publicly available sources or hypothetical scenarios. Any user interaction data referenced (e.g., example dialogues) were either from published research or anonymized forum posts in the public domain. No identifiable personal user data was used, thus no institutional review was required for this conceptual research. We have adhered to fair use and academic standards in referencing sources.
- **AI Tool Involvement:** Portions of this document were developed with the assistance of AI language models (for brainstorming and draft generation, and to act as "interlocutors"), consistent with the paper's topic. While AI-assisted, all arguments, interpretations, and conclusions are the product of human expertise. The author took care to verify all content, integrate human expertise, and ensure that the final narrative, arguments, and conclusions are original and properly attributed where sources are used. Any use of AI did not involve sensitive data and was aimed at improving clarity and coherence. The final responsibility for the content lies with the author, who critically reviewed and edited all AI contributions. And if there is an error, this reflects the fact that the author is human, all too human.

Acknowledgments:

[2] V(Major).(Minor) – (Date or Revision Description):

- Major Version (V1.0, V2.0, etc.) → Used for significant updates (e.g., new sections, substantial revisions, or conceptual shifts).
- Minor Version (V1.1, V1.2, etc.) → Used for small edits, clarifications, formatting updates, and typo fixes.
- Optional Patch (V1.1.1, etc.) → Used to track micro-edits (e.g., fixing a single reference or small wording change).
- Release Date or Description → Used to help readers track and contextualize updates.

As noted, the author acknowledges the assistance of multiple generative AI tools that were used as conversational partners ("interlocutors") throughout the writing process, rather than any collaboration with human co-authors. These AI platforms – including OpenAI's ChatGPT among others – aided in refining the manuscript's language, suggesting structure, and integrating feedback. All final content remains the sole responsibility of the author, who maintained full editorial control over the work. The contributions of the AI tools are duly credited here for their support in the writing process, but they are not listed as co-authors.

Disclosure:

The author confirms that this work was conducted independently, with no external funding or institutional influence. Any perspectives or insights drawn from the author's background (being alive and *in situ* here on Earth) were applied impartially. No conflicts of interest are present, and the views and conclusions expressed in this paper are solely those of the author. All opinions and analyses were developed in a personal capacity and are not influenced by any current or former employer or related enterprise.

References:

Books

- Berger, P. L., & Luckmann, T. (1966). *The social construction of reality*. Anchor Books.
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.
- Coeckelbergh, M. (2020). *AI ethics*. MIT Press.
- Dennett, D. C. (1989). *The intentional stance*. MIT Press.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.
- Floridi, L. (2011). *The philosophy of information*. Oxford University Press.
- Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)
- James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. Longmans, Green, and Co.
- Nagel, T. (1974). What is it like to be a bat? *In Mortal questions* (pp. 165–180). Cambridge University Press. (Original work published 1974)
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
- Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Alfred A. Knopf.
- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

- Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
- Varela, F. J., Thompson, E., & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience*. MIT Press.
- Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell Publishing.

Journal Articles & Conference Papers

- Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, 8(1), 339–359. https://doi.org/10.1007/BF00258436
- Luger, E., & Sellen, A. (2016). "Like having a really bad PA": The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286–5297). Association for Computing Machinery (ACM). https://doi.org/10.1145/2858036.2858288
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). Association for Computing Machinery (ACM). https://doi.org/10.1145/2939672.2939778
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457. https://doi.org/10.1017/S0140525X00005756
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Book Chapters

- Haraway, D. (1985). A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, cyborgs, and women: The reinvention of nature* (pp. 149–181). Routledge.

Preprints & Online Articles

- Bennett, M. T., Welsh, S., & Ciaunica, A. (2024). Why is anything conscious? *arXiv Preprint*. https://arxiv.org/abs/2409.14545
- An, S., Ma, Z., Lin, Z., Zheng, N., & Lou, J.-G. (2024). Make your LLM fully utilize the context. *arXiv preprint*. arXiv:2404.16811. https://arxiv.org/abs/2404.16811
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv preprint* arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint*. https://arxiv.org/abs/1702.08608
- Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. *arXiv Preprint*. https://arxiv.org/abs/1410.5401
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30 (pp. 5998–6008). https://arxiv.org/abs/1706.03762