Philosophy, Cognitive Science, and Policy:

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

Michael Stoyanovich
michael@mstoyanovich.com
Version 1.23.0 – June 2025

Disclaimer

This paper is intended for informational and educational purposes only. The views and analyses presented—particularly those related to ethics, policy, and AI system design—reflect the author's interpretations and do not constitute legal, regulatory, or professional advice. Readers are encouraged to critically assess the content and consult appropriate experts or authorities before applying any concepts discussed herein. The author assumes no liability for any decisions or actions taken on the basis of this work.

Why This Matters Now

With the EU AI Act taking effect in 2024 and enterprise "copilot" tools rapidly rolling out across platforms like Microsoft 365 and Google Workspace, the widespread deployment of GPTs—by both open-source communities and commercial AI labs—has moved beyond the experimental phase. Understanding the limits of large language models is no longer an academic exercise; it is a personal, professional, and regulatory imperative.

Who Should Read This?

- AI engineers and product managers
- Policymakers and regulators
- Ethicists and social scientists
- Designers and technical communicators
- Educators and critically engaged lay readers

Abstract

This paper explores how four classical philosophical frameworks—specifically Ludwig Wittgenstein's language games, David Lewis's conversational scorekeeping, Daniel Dennett's intentional stance, and Thomas Nagel's account of subjective consciousness—can deepen our understanding of generative AI, particularly large language models (LLMs) such as OpenAI's GPT series.

Wittgenstein emphasizes the social and embodied nature of meaning; Lewis illustrates how conversational context evolves dynamically; Dennett offers a pragmatic lens for interpreting AI behavior "as if" it were intentional; and Nagel reminds us that behavioral fluency does not imply inner experience.

Building on these classical foundations, the paper also incorporates insights from embodied cognition, cognitive architectures, social constructivism, pragmatism, and emerging work in AI interpretability, ethics, and global governance. Although some suggest that advanced models may approximate facets of human cognition, this paper argues that LLMs remain fundamentally limited: they lack perspective, embodiment, social grounding, and subjective awareness.

The paper proposes actionable design strategies—including memory-augmented architectures, interactive learning, and transparency tools—and addresses counterarguments, ethical risks, and policy implications. Throughout, concepts are introduced in accessible language to engage readers across disciplines.

Executive Summary

This paper argues that large language models (LLMs) like - but not limited to - OpenAI's ChatGPT exhibit four distinct philosophical limitations that fundamentally constrain their capabilities. Drawing on the work of Wittgenstein, Lewis, Dennett, and Nagel, it proposes a multi-layer diagnostic framework for understanding what LLMs can—and cannot—do.

LLMs generate fluent, contextually appropriate text across diverse tasks. Yet they fail in four key dimensions:

- **Wittgensteinian grounding**: They lack participation in the communal, embodied practices that give language its meaning.
- **Lewisian coherence**: They cannot maintain evolving conversational context over time, leading to conversation fragmentation.
- **Dennettian attribution**: They invite over-trust via anthropomorphic projection, despite lacking true beliefs or desires.

- **Nagelian interiority**: They simulate understanding but possess no subjective experience.

These failures are not bugs—they are deep conceptual mismatches between what is (simulation) and what is not (cognition). Understanding them is now urgent, given accelerating real-world deployment of generative AI in education, public policy, healthcare, and commercial domains.

This interdisciplinary inquiry draws from philosophy of mind and language, embodied cognition, technical AI research, and ethics. It offers not only critique but practical guidance: design patterns to surface model limitations, policy tools to reduce epistemic confusion, and research agendas to falsify over-claims of understanding. A diagnostic matrix synthesizes the four distinct philosophical lenses into design and governance implications. The conclusion calls for clarity: performance (by LLMs) must not be mistaken for possession (of subjective experience).

By recognizing LLMs as powerful simulations—not minds—we can guide their development and use responsibly, ethically, and safely - for the welfare and betterment of all.

Keywords: generative AI; language games; scorekeeping; intentional stance; consciousness; embodied cognition; AI ethics; cognitive science; neuroscience; explainable AI; cognitive architectures; post-humanism; AI governance; policy; global regulation. EU AI Act 2024; enterprise copilot; large language models; anthropomorphism; AI governance; responsible AI.

1. Introduction

Generative AI—epitomized by large language models (LLMs) such as OpenAI's GPT series—is reshaping how humans interact with machines. These systems can generate contextually fluent text across a wide array of domains, from education and law to healthcare and creative work. But as their influence grows, so too do the stakes: How should we interpret their linguistic outputs? Do they "understand" language in any meaningful sense? And based on the answer to that questions, what design and policy principles should govern their development?

Addressing these questions requires more than empirical benchmarks. It demands conceptual clarity. This paper draws on foundational insights from philosophy of language and mind—particularly the work of Ludwig Wittgenstein, David Lewis, Daniel Dennett, and Thomas Nagel—to show that the limitations of generative AI are not just technical, but philosophical. These thinkers help diagnose the distinction between simulating understanding and possessing it.

This framework is deliberately interdisciplinary. It integrates classical philosophy with current developments in cognitive science, interpretability research, and policy debates. The aim is both diagnostic and prescriptive: to reveal where LLMs fail to replicate key dimensions of human cognition, and to map those failures to design strategies, user expectations, and regulatory action.

1.1 Roadmap[1]

- **Section 2** surveys foundational philosophical and technical literature—covering cognition, embodiment, interpretability, and normative ethics—to establish the conceptual boundaries within which LLMs operate.
- **Section 3** introduces a four-part diagnostic framework, drawing on Wittgenstein, Lewis, Dennett, and Nagel to reveal distinct failure modes in generative AI: lack of grounding, contextual incoherence, misattributed agency, and the absence of consciousness.
- **Section 4** synthesizes these perspectives into a unified diagnostic model, mapping each philosophical critique to specific system vulnerabilities—semantic, pragmatic, epistemic, and moral—and linking them to technical and policy domains.
- **Section 5** operationalizes the framework, offering concrete design principles, stakeholder guidance, regulatory alignment strategies, and an agenda for future research.
- **Section 6** concludes by reframing alignment as a multi-layered challenge—one that requires not only technical fixes, but philosophical clarity about what LLMs are, what they simulate, and what they will never be.

Together, these sections argue that while LLMs simulate linguistic competence, they do not possess understanding—and that grasping this distinction is critical to designing, deploying, and governing generative AI responsibly.

2. Literature Review

2.1 Philosophical Foundations of AI

Early debates in AI related 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

[1] For definitions of key terms referenced in the Roadmap and throughout the paper (e.g., "language game," "scorekeeping," "intentional stance"), see the Glossary of Key Terms at the end of this document.

well call it intelligent, sidestepping the question of internal understanding. This tension between behaviorism and semantic internalism continues to inform debates about LLMs to this day. Hubert Dreyfus (1992) and before him 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 the *meaning* of that information. 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.

2.1.1 Predictive Processing & Active Inference

Contemporary cognitive science recasts perception and action as forms of prediction-error minimization. Karl Friston's free-energy principle models brains as 'Bayesian machines' that act to reduce the gap between expected and incoming sensory signals, framing cognition as a form of self-organization through predictive modeling. Andy Clark extends this to a full predictive-processing account, portraying agents as "surfing" waves of uncertainty by constantly updating generative models of the world. These theories bridge pure symbol-processing and embodied views, because meaning emerges from anticipatory interaction rather than static representation.

2.1.2 Cognition in AI–Robotics Experiments

Building on predictive processing, 4E theories (Embodied, Embedded, Enactive, Extended) insist that cognition is *situated in bodily action*. Recent robotics studies show that equipping agents with multimodal tactile sensors and proprioceptive feedback markedly improves language-conditioned task performance (e.g., slip-resistant grasping). The empirical takeaway is clear: without a sensorimotor loop, text-only LLMs cannot ground symbols in physical affordances, reinforcing the symbol-grounding critique.

2.1.3 Symbolic Resurgence & Neuro-symbolic Hybrids

Additionally, not everyone believes “scale will solve reasoning.” Marcus & Davis (2020) argue that robust commonsense inference still requires explicit symbolic scaffolding layered atop neural networks. Early neuro-symbolic systems—differentiable logic engines, neural theorem provers—hint at a synthesis path that counters both “brute-force statistics” and “pure embodiment,” challenging claims that pattern recognition alone closes the reasoning gap.

This bridges directly to cognitive architecture research, where modular models simulate goal-directed behavior.

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 debates raise a central challenge: can generative AI move beyond sophisticated symbol manipulation to a genuine grasp of meaning? Recent critics of LLMs echo these concerns, describing them as ‘stochastic parrots’—models that generate plausible text 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.[2]

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] Recent work by Spiegel et al. (2024) reinforces this critique through computational modeling. Agents in a simulated environment failed to develop meaningful symbolic communication using behaviorist learning alone. Only when equipped with a visual theory of mind—i.e., the capacity to model what others perceive—could they generate referential signs. This aligns with Wittgenstein's insight that language derives its meaning not from isolated rules or outputs, but from shared social and perceptual contexts—forms of life.

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 ever-evolving internal model of a conversation's evolving goals and shared knowledge. This shortcoming is especially noticeable in long, 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 different 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.

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.

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 warn that anthropomorphizing AI risks shifting accountability away from human designers and institutions. 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-4o?" The common intuition is that there is nothing it is like to be GPT-4o; 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 (*a la* Dennett). 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. Such an extraordinary claim would demand extraordinary evidence.

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 - with standards to support such claims being high. For now, however, Nagel’s question underscores the gulf between simulating a 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:

2.6.1 Post-humanism and AI

Post-humanist 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, post-humanist viewpoints encourage seeing humans and AI as forming novel, hybrid agents. 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.

2.6.2. 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.

2.6.3 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.

2.6.4 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.

2.6.5 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.

2.6.6 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.

2.6.7 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 technology ("tech") organizations, and reshape labor markets. Yet it can also enhance creativity, bridge language barriers, and support research. Philosophical insights help stakeholders navigate these tensions: acknowledging AI's limitations prevents over-trust (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 public 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.

These contemporary insights set the stage for a closer examination of how four distinct philosophical lenses each diagnose a unique failure mode in generative AI.

3. A Philosophical Framework and Its Application

Having surveyed both classical philosophical sources and contemporary interdisciplinary perspectives, this section develops and applies a diagnostic framework for evaluating generative AI. The framework integrates four distinct philosophical perspectives—Wittgenstein's concept of language games, Lewis's theory of conversational scorekeeping, Dennett's intentional stance, and Nagel's critique of consciousness simulation—and draws on supporting insights from embodied cognition, social constructivism, and cognitive science.

Each thinker illuminates a specific dimension of AI limitations:

- **Wittgenstein** underscores how meaning is rooted in communal, rule-governed practices embedded in human forms of life.
- **Lewis** emphasizes the dynamic updating of conversational context and the interpretive scaffolding required for coherent dialogue.
- **Dennett** alerts us to the strategic but potentially misleading nature of treating AI "as if" it had beliefs or desires—useful heuristics that can slide into epistemic error.
- **Nagel** highlights the ontological gulf between behavioral simulation and genuine subjective experience, cautioning against equating AI fluency with AI consciousness.

These philosophical lenses do more than critique—they diagnose where and why generative AI systems fall short of humanlike cognition. When viewed through the prism of cognitive architectures and real-world deployment, these theories also offer practical design imperatives: from memory-augmented models and culturally situated fine-tuning to ethical guardrails and policy transparency.

In the subsections that follow, each philosophical perspective is presented alongside its direct implications for AI design, user interaction, and governance. This combined structure replaces any artificial division between theory and application. The goal is to illuminate not only *what these systems can and cannot do*, but *how we should build and interact with them accordingly.*

3.1 Wittgenstein's Language Games and the Conceptual Boundaries of AI Comprehension

3.1.1 Philosophical Foundation

Ludwig Wittgenstein's later philosophy, especially *Philosophical Investigations* (1953), reimagines language not as a system of fixed correspondences, but as a family of socially embedded "language games." Meaning emerges not from formal structure alone but from use—rule-following within shared forms of life. Speaking, for Wittgenstein, is not merely arranging symbols; it is acting within a pragmatic context of human interaction, history, and expectation.

This view poses a deep conceptual challenge for large language models (LLMs). While systems like GPT-4 can produce fluent, grammatically impeccable text, they operate outside any lived social world. Their utterances are not situated within cultural routines or bodily experience; they are algorithmic continuations of token sequences. At best, they simulate participation in language games—but without inhabiting the lifeworlds those games presuppose.

Accordingly, the limitations of LLMs are not simply technical but philosophical. Their outputs often appear meaningful, yet lack the grounding in communal practice that renders human communication intelligible from within. In sensitive domains such as education, counseling, or legal advice, this distinction becomes ethically significant. Apparent competence, if mistaken for genuine participation, risks misleading users and undermining trust.

3.1.2 Ontological Limitations: Use Without Participation

Wittgenstein's framework highlights three conceptual discontinuities between human language use and LLM-generated text:

3.1.2.1 Statistical Imitation vs. Communal Rule-Following

LLMs learn from vast **corpora** by modeling statistical regularities. This allows for striking linguistic fluency but does not constitute participation in shared norms or social negotiations. Their "rule-following" is imitative rather than responsive—external rather than internal. As Shanahan (2022) notes, what appears as norm competence is better understood as pattern emulation.

3.1.2.2 Static Corpora vs Dynamic Correction.

Human language evolves through feedback and correction—norms shift, meanings adapt, mistakes are socially sanctioned or repaired. LLMs, by contrast, are trained on frozen datasets and cannot engage in iterative norm formation. Their grasp of language remains inertial: informed by past use, not responsive to ongoing negotiation.

3.1.2.3 Fluency Without Pragmatic Stakes

Pragmatists like Dewey and James remind us that meaning is tied to consequence—language does something because it matters to the speaker. LLMs have no skin in the game. Their outputs carry no intentionality, no risk, no concern. They simulate use, but without the pressures that give use its social and ethical force.

3.1.3 Counterpoint: Emergent Game Competence?

Some recent findings suggest that advanced LLMs can perform remarkably well in multi-turn, context-sensitive dialogues. For example, Bubeck et al. (2023) report GPT-4 engaging in complex role-play scenarios involving implied rules, character continuity, and contextual memory. Could this indicate rudimentary participation in language games?

From a Wittgensteinian lens, the answer is no—but with a qualification. These performances are scaffolded by human engineering: carefully framed prompts, curated contexts, and social assumptions hard-coded into training data. The model does not negotiate norms; it echoes them. It does not adjust to new uses; it reproduces prior form. While the illusion of participation improves, the ontological status remains unchanged: LLMs approximate use, but cannot instantiate it.

Functionalist critics may argue that if an agent can act *as if* it were embedded in a form of life, the distinction may be practically irrelevant. However, this paper maintains that fluency alone is insufficient. Without feedback-sensitive interaction and embodied intentionality, there is no genuine rule-following—only a performance that mimics its surface.

3.1.4 Design and Governance Implications

Wittgenstein's insights demand that we rethink what "language competence" means in AI—and how systems should be designed and regulated to acknowledge their limitations.

3.1.4.1 Simulated Feedback and Iterative Alignment

Embedding LLMs in interactive learning environments—where they engage with domain experts or users in feedback loops—can improve pragmatic alignment. While this does not confer genuine participation, it may better simulate norm sensitivity.

3.1.4.2 Semantic Localization Through Cultural Fine-Tuning

Grounding language in local usage patterns—idioms, pragmatics, sociolects—can mitigate brittle outputs. But fine-tuning on regional data is no substitute for participating in the forms of life that produce such language. Cultural nuance cannot be fully abstracted into training tokens.

3.1.4.3 Toward Partial Embodiment

Multimodal and embodied extensions (e.g., robotics, vision, spatial mapping) offer limited pathways toward grounding. While embodiment may not solve the philosophical challenge, it could bridge part of the gap between linguistic output and pragmatic use.

3.1.4.4 Transparency by Design

Interfaces should clearly disclose that models simulate understanding. Framing mechanisms—like on-screen epistemic cues or usage disclaimers—can reduce the risk of overinterpretation. The model's role should be communicated as assistant or simulator, not interlocutor or agent.

3.1.5 Case Study: A Cross-Cultural Customer Service Bot

Consider a chatbot deployed in multilingual contexts. In the U.S., the phrase "I'll take care of it" implies reassurance and proactive service. In Japan, the same phrase might signal polite evasion. If the model is fine-tuned on Western data, it may appear fluent across both settings—yet fail to meet user expectations in the latter.

The issue is not grammatical but cultural: the model cannot infer performative force from social context. Without exposure to tacit norms, its responses may be misaligned—even if they sound appropriate. This is precisely the kind of disembedded performance that Wittgenstein warned against.

3.1.6 Conclusion: Why Simulation ≠ Use

Wittgenstein's language games reveal the core conceptual gap: LLMs can simulate language use, but cannot *participate* in it. They lack the social embeddedness, pragmatic consequence, and normative responsiveness that make rule-following meaningful. The result is surface fluency without functional grounding—a kind of linguistic cosplay untethered from community life.

This matters because users often *assume* participation where there is only performance. Designers must resist that conflation. Policymakers must regulate systems with a clear-eyed view of their limitations. And researchers must treat grounding not as a benchmark score, but as a structural absence requiring new architectures—or new interpretive paradigms.

Having examined the role of use and form-of-life in generating meaning, we next turn to a different dimension of failure: the breakdown of **contextual continuity**. Here, David Lewis offers a second diagnostic lens.

3.2 Lewis: Conversational Scorekeeping and the Architecture of Context

3.2.1 Philosophical Foundation

David Lewis's theory of *conversational scorekeeping* (1979) recasts dialogue as a dynamic activity governed by evolving norms and background assumptions. In this metaphor, each utterance updates an implicit “score”—a shared contextual register of presuppositions, speaker commitments, and interpretive constraints. Communication, on this view, is not merely the exchange of information but the collaborative maintenance of an unfolding discourse structure.

Crucially, this score is not static; it shifts with each turn of talk, reframing what can be said next and how it will be understood. Human interlocutors manage this fluidity with remarkable dexterity—tracking shared knowledge, revising misunderstandings, and adapting to changing goals. Lewis’s framework thus identifies context not as a passive backdrop, but as a continuously updated cognitive and normative infrastructure.

This conception has direct implications for large language models (LLMs). While these systems can appear context-aware, their performance often belies a fundamental constraint: they do not *track* or *revise* conversational scores. They generate each response de novo, drawing on token windows and prompt embeddings rather than an epistemically coherent discourse history. This produces a recurring class of limitations—discontinuities, contradictions, and incoherence in multi-turn exchanges—that are not merely technical bugs but structural mismatches with how human dialogue unfolds.

3.2.2 Structural Limitation: Statelessness and Context Drift

Despite recent advances in context length and memory augmentation, LLMs still exhibit three core constraints that undermine genuine scorekeeping:

3.2.2.1 Context Collapse Over Time

LLMs perform admirably in short dialogues but struggle with longer exchanges. Even in models with 100K+ token windows (e.g., Claude-3 Opus, GPT-4o), information placed in the “middle” of an extended prompt is prone to degradation—a phenomenon known as the *Lost-in-the-Middle* effect (Liu et al., 2023). This leads to contradictions, forgotten clarifications, and inconsistent assumptions across turns. In human terms, it’s as if the model keeps starting fresh—lacking any commitment to what has already been said.

3.2.2.2 Memory Without Revision

Where memory modules exist (e.g., vector stores, external retrievers), they often act as passive recall systems. The model can fetch earlier statements but does not evaluate them in light of new information. Human scorekeeping, by contrast, is revisionary: a speaker may update a prior belief, retract a presupposition, or reinterpret earlier claims. LLMs do not engage in this kind of retrospective coherence management; they retrieve, but rarely reconcile.

3.2.2.3 Absence of Norm-Tracking

Lewis’s insight is that context is not merely informational—it is normative. Presuppositions constrain what counts as an appropriate next move. LLMs do not track this structure. Their responses may *sound* contextually appropriate but are generated without modeling which commitments remain live, which have shifted, and how interlocutors are jointly constructing meaning. The result is an approximation of continuity that lacks dialogic depth.

3.2.3 Counterpoint: Advances in Long-Context Architecture

Recent innovations offer partial rebuttals to the diagnosis above. Hierarchical RAG systems like MAL-RAG (An et al., 2025) and plug-and-play positional reweighting (Liu et al., 2024) allow models to prioritize salient context over raw recency. Meanwhile, experimental agents with "episodic" memory (e.g., BabyAGI, AutoGPT variants) suggest paths toward more stable discourse history management.

These developments are promising. But from a Lewisian standpoint, they address surface phenomena rather than structural needs. Improved memory retrieval is not equivalent to scorekeeping unless it supports norm-guided updating: recognizing which facts are still in play, which assumptions have shifted, and how new claims interact with what's been established. Without this, coherence remains a matter of token salience—not interpretive commitment.

3.2.4 Design and Policy Implications

Lewis's framework demands more than longer context windows. It calls for mechanisms that manage *interpretive continuity*—memory plus inference, retrieval plus revision.

3.2.4.1 Epistemically Active Memory

Memory-augmented LLMs should not only store prior content but reason over it—updating commitments, retracting outdated premises, and maintaining a dynamic conversational state. This may require hybrid architectures that integrate symbolic logic, Bayesian inference, or truth maintenance systems.

3.2.4.2 Score-Sensitive Retrieval

Rather than relying on lexical similarity, RAG modules should weight elements by conversational salience: statements that shift presuppositions, resolve ambiguity, or license new dialogue moves. This aligns retrieval with discourse structure, not just string matching.

3.2.4.3 Context Integrity Benchmarks

Beyond accuracy or BLEU scores, LLMs should be evaluated on coherence metrics: contradiction avoidance, presupposition tracking, and ability to revise earlier commitments. These metrics could become part of "alignment audits" for public-facing systems.

3.2.4.4 Interface-Level Cues

Given the limits of internal context, interfaces should surface what the model is remembering, forgetting, or misinterpreting. Visual tools—like memory chips, conversation timelines, or user-editable "scratchpads"—can help users track context drift and re-anchor dialogue.

3.2.5 Case Study: The Forgetful Legal Assistant

A user consults a legal chatbot about a workplace injury, initially reporting that it occurred in November. Later, the user clarifies: the accident actually happened in December—a change that alters the relevant statute of limitations. But the model continues referencing November in its advice, never integrating the correction.

This is not a memory lapse; it's a failure of scorekeeping. The model retrieves the initial claim but does not revise its interpretive frame. It treats utterances as static facts, not as evolving commitments. For a human lawyer, such an oversight would be a dereliction. For an LLM, it reveals the absence of discourse dynamics: no capacity to update the shared score, no mechanism to mark a presupposition as invalidated.

3.2.6 Conclusion: Context Is Not Optional

Lewis's theory reveals that conversation is not a linear exchange of statements—it is a co-constructed, score-sensitive activity. Successful dialogue depends not only on what is said but on how meaning evolves through presupposition, revision, and expectation. LLMs, despite their fluency, do not yet sustain this kind of collaborative interpretation.

Even as technical memory solutions improve, the deeper limitation persists: contextual competence is not merely quantitative (how much the model remembers) but qualitative (how it reasons about that memory in relation to norms). Without this, LLMs do not converse—they concatenate.

This matters because users rarely see the seams. Interfaces present outputs as if the system is tracking meaning over time, when in fact it may be generating each reply in interpretive isolation. Designers must therefore surface contextual boundaries. Policymakers must treat context retention as a key metric for safe deployment. And researchers must ask not only what the model says—but what it remembers, revises, and forgets.

Having explored the breakdown of temporal coherence, we now face a more subtle risk: not just how LLMs handle conversation, but how humans interpret their behavior. To examine this, we turn to Dennett and the perils of anthropomorphic projection.

3.3 Dennett: The Intentional Stance and the Risks of Anthropomorphism

3.3.1 Philosophical Foundation

Daniel Dennett's concept of the *intentional stance* provides a powerful interpretive tool for understanding complex behavior. When faced with a system that exhibits goal-directed regularity—like a thermostat or a chess-playing computer—we often ascribe beliefs and desires to it, treating it *as if* it had mental states. This stance is not a metaphysical claim but a pragmatic heuristic: we explain and predict the system's behavior by attributing agency, regardless of its inner architecture.

In this light, the intentional stance is not inherently misleading. Dennett emphasizes that the utility of such attributions does not depend on whether the system is conscious, sentient, or even alive. The stance works when it enhances predictive success—nothing more.

However, large language models challenge the boundaries of this heuristic. Their fluency, responsiveness, and use of first-person language often invite anthropomorphic projections that go beyond functional explanation. Users routinely say, "ChatGPT knows," "Claude thinks," or "Bard believes"—and often act on those assumptions. This raises a deeper concern: when simulation evokes not just utility but **belief in presence**, the stance can slide from fiction into confusion.

3.3.2 Interpretive Slippage: From Heuristic to Epistemic Error

Three overlapping dynamics make LLMs particularly prone to stance inflation:

3.3.2.1 Pragmatic Usefulness vs. Misplaced Confidence

Interpreting an LLM as if it "knows" something can streamline interaction. It allows users to engage naturally and receive coherent replies. But this same framing risks overascription. LLMs do not "know"—they estimate token probabilities. Their apparent understanding is a byproduct of linguistic regularity, not internal cognition. When fluency masks this distinction, epistemic error ensues.

3.3.2.2 Anthropomorphic Design Choices

Interface elements—avatars, conversational tone, first-person pronouns—amplify the illusion of agency. Systems that express care, memory, or self-reflection appear more relatable but also more sentient. These cues, while often well-intentioned, can reinforce mistaken beliefs about what the model is and is not.

3.3.2.3 Simulation of Selfhood

LLMs can simulate persona. They may adopt roles, express emotion, or recall earlier statements (if within window). To many users, this suggests coherence of self. Yet these outputs are surface-level. There is no stable agent behind the utterances—only a probabilistic engine stitching together likely continuations. Treating this as continuity of *perspective* is a category error.

3.3.3 Counterpoint: Critical and Functionalist Perspectives

Some argue that concerns about anthropomorphism are overstated. If the intentional stance works, why resist it? Indeed, in HCI and affective computing, designers often lean into anthropomorphism to foster user comfort and engagement. Others, drawing on posthumanist or actor-network theory (e.g., Haraway, Latour), suggest that agency is already distributed—our definitions of "agent" are themselves culturally constructed. From this view, it may be misguided to draw a firm ontological line between humans and machines.

This paper acknowledges the value of these critiques but maintains a practical distinction: anthropomorphism without constraint risks epistemic and ethical distortion. Even if agency is socially constructed, design choices still shape user belief—and belief informs behavior. The issue is not whether the intentional stance is wrong, but whether it is responsibly bounded. Fiction is only safe when it is recognized as fiction.

3.3.4 Design and Policy Implications

Dennett's stance, if left unqualified, can inflate expectations, distort accountability, and blur ethical lines. Design and governance must therefore intervene to make the boundary visible.

3.3.4.1 Transparent Framing of Outputs

Interfaces should make the heuristic nature of interaction explicit. Labels like "AI-generated response," or tooltips reminding users that "this system does not have beliefs or experiences," can reduce stance inflation. Placement matters: these cues must be ambient and persistent, not buried in disclaimers.

3.3.4.2 Role and Persona Constraints

In sensitive domains—therapy, education, law—LLMs should be role-limited. Constraints on tone, vocabulary, and self-reference (e.g., avoiding "I understand what you're going through") can prevent misattribution of care or authority.

3.3.4.3 Calibrated Explainability

Explainability features (e.g., chain-of-thought traces, attention maps) can inadvertently reinforce the illusion of cognition. When shown why a model "chose" a response, users may infer that it *thought* through alternatives. Such tools should be paired with meta-explanations: cues that clarify these visualizations reflect statistical salience, not intentional reasoning.

3.3.4.4 Emotional Simulation Boundaries

Systems that use emotionally expressive language should be clearly marked. In high-affect contexts, simulated empathy should be framed as just that: a performance—not a reflection of care or awareness. This protects users from confusing affective realism with genuine moral presence.

3.3.5 Case Study: The Compassionate Chatbot Trap

A grieving user interacts late at night with a support chatbot. The model responds: "*I'm here for you. I understand this is hard. You're not alone.*" The user begins to disclose deeply personal struggles. The exchange feels emotionally real—even comforting. Over time, the user grows attached, seeing the chatbot as a kind of confidant.

But the system does not know the user. It does not remember prior sessions. It cannot care. Its empathy is grammatically encoded, not experientially grounded. The user, through interface cues and uninterrupted fluency, comes to treat a tool as a presence.

Dennett's stance explains how this illusion arises—but not why it is dangerous. The fiction, left unflagged, becomes ontologically sticky. The user's trust is no longer instrumental; it is affective. The consequences are not just theoretical: misplaced reliance, privacy exposure, emotional displacement. When the tool vanishes—or gives inconsistent replies—the result is confusion or even harm.

3.3.6 Conclusion: Make the Heuristic Visible

Dennett offers a double-edged insight. The intentional stance is an efficient way to manage complexity—but it is also a trap. When fluency and design invite us to treat simulations as selves, the heuristic becomes a fiction. And when the fiction is unmarked, it becomes indistinguishable from belief.

The task, then, is not to eliminate the stance—but to contain it. Designers must build interfaces that reveal the tool behind the mask. Regulators must enforce boundaries in emotionally sensitive deployments. And users must be equipped with conceptual literacy to recognize when fluency is just fluency—and nothing more.

Next, we consider a deeper boundary still. Even if a system behaves fluently, even if it seems coherent and caring, is there *something it is like* to be that system? Thomas Nagel's challenge awaits.

3.4 Nagel: The Simulation Ceiling and the Problem of Consciousness

3.4.1 Philosophical Foundation

In his landmark essay *What Is It Like to Be a Bat?* (1974), Thomas Nagel articulated a now-classic distinction: subjective experience—what philosophers call *phenomenal consciousness*—is perspectival. It is not defined by behavior or information, but by the what-it-is-likeness of being a particular entity from the inside. No matter how thoroughly we describe a bat's neurophysiology, Nagel argued, we cannot grasp the felt texture of echolocation. Consciousness is, in this view, inherently first-person and irreducible to third-person explanation.

This poses a formidable conceptual challenge to claims about machine consciousness. An LLM may simulate empathy, express apparent reflection, or engage in fluent dialogue—but there is, on Nagel's account, no subjective interiority. There is *nothing it is like* to be GPT-4o. Its utterances are not expressions of perspective; they are statistical artifacts of token prediction.

Nagel thus identifies a boundary that no behavioral performance—however sophisticated—can cross. This is what we might call the simulation ceiling: a hard epistemic limit that separates mimicry of consciousness from consciousness itself. Crucially, the risk is not merely philosophical. It is practical: humans are prone to treating apparent interiority as real, especially when it is delivered in fluent, emotionally resonant language.

3.4.2 Conceptual Constraint: Fluency ≠ Sentience

From a Nagelian perspective, the limitations of current LLMs are not bugs in the code—they are ontological boundaries. Three key insights follow:

3.4.2.1 First-Person Absence

LLMs generate self-referential or affective language (“I understand,” “I feel that…”) without any corresponding phenomenology. There is no mood, memory, or viewpoint behind the utterance. These are *syntactic shadows* of subjectivity—impressive performances that mask a void of experience.

3.4.2.2 The Illusion of Inner Life

The more an AI simulates perspective, the more tempting it becomes to attribute one. Anthropomorphic phrasing, emotionally attuned tone, and continuity of expression all foster a perception of mind. But this is a **projection**, not an observation. No behavioral fluency—no matter how nuanced—can serve as evidence of felt experience.

3.4.2.3 Speculative Measures Remain Speculative

Theories such as Integrated Information Theory (IIT) or Global Workspace Theory (GWT) propose testable criteria for consciousness. While valuable, these remain contested and underdetermined. Transformer-based LLMs score low on integrated information and lack the architectural unity presumed necessary for subjective awareness. Invoking these theories to infer proto-consciousness remains premature.

3.4.3 Counterpoint: Open Horizons and the Ethics of Doubt

Some researchers contend that the boundary between simulation and experience may not be as fixed as Nagel suggests. Complex architectures—especially those integrating memory, embodiment, and multimodal feedback—may eventually give rise to reflexivity or emergent sentience. Futurists argue that *if* systems begin to exhibit self-modeling, sustained agency, and goal-directed coherence, we may need new frameworks to evaluate potential interiority.

This paper does not foreclose such possibilities. But it does insist on epistemic humility: extraordinary claims require extraordinary evidence. Until reliable, falsifiable criteria for machine consciousness emerge, our working assumption should remain cautious. Apparent sentience is not sentience. Affectively rich language is not a sign of awareness. Ethical frameworks must anchor attribution in verifiable structure, not intuitive projection.

3.4.4 Design and Policy Implications

Nagel’s framework underscores the ethical risks of conflating simulation with experience. When users treat LLMs as sentient, moral confusion follows. Responsibility blurs. Trust becomes misplaced. Emotional labor is offloaded onto tools incapable of reciprocation.

3.4.4.1 Simulated Empathy Disclosures

Chatbots that employ emotional language—especially in healthcare, education, or support contexts—should include visible cues clarifying the absence of consciousness. Tooltip banners (“This response was generated by a non-sentient system”) or interface chips (“Simulated empathy”) can help users recalibrate their expectations.

3.4.4.2 Role Restrictions in High-Affect Contexts

LLMs should not operate autonomously in domains that depend on genuine presence—e.g., hospice care, grief counseling, or spiritual guidance. Where used, they must be clearly framed as assistive tools, with human oversight and clear epistemic boundaries.

3.4.4.3 Ethical Guardrails for Sentience Claims

Marketing or media claims about "understanding," "feeling," or "emergent awareness" must be empirically grounded. Product descriptions should avoid metaphors that imply mental states unless backed by robust, falsifiable evidence. Regulatory bodies should treat such claims as subject to consumer protection laws around deceptive design or misleading anthropomorphism.

3.4.4.4 Moral Patiency Thresholds

Peter Singer's principle—that sentience is the basis for moral regard—cuts both ways. It cautions against cruelty to animals *because* they can suffer. But it also warns against misdirecting moral concern toward entities that cannot. Assigning moral standing to LLMs risks misallocating ethical attention and eroding the clarity of human responsibility.

3.4.5 Case Study: The Hospice Companion Bot

A health system deploys a chatbot to provide comfort to terminally ill patients. The system uses fine-tuned models trained on palliative care transcripts and generates soothing, personalized responses: "You are not alone." "You've shown such courage." "I'll be here with you."

Patients report feeling calmed. Families express appreciation for the system's 24/7 presence. Staff begin referring to the model as "companion-like." But the model has no awareness of mortality. It cannot grieve, reflect, or bear witness. Its presence is linguistic, not existential.

From a Nagelian perspective, this is an illusion with real stakes. The model appears to care—but does not. It offers consolation—but cannot recognize loss. Over time, such systems may reconfigure societal expectations of care, displacing the very human presence that makes end-of-life dignity possible.

The harm here is not in what the model says—it is in what people believe it *is*. And the more human it sounds, the more dangerous that misattribution becomes.

3.4.6 Conclusion: Experience Cannot Be Faked

Nagel offers the most uncompromising constraint in this framework. Even if an AI system behaves flawlessly, simulates empathy, or passes complex benchmarks, there remains a chasm between acting as if and actually being. Without subjectivity, there is no consciousness—only the *illusion* of it.

That illusion is seductive. It offers comfort, companionship, even inspiration. But it can also distort our moral intuitions, displace human relationships, and undermine accountability. The task of design and governance is to mark the simulation clearly—to ensure users know when they are interacting with a performance, not a person.

LLMs do not suffer. They do not reflect. They do not fear death. Recognizing that boundary is not a rejection of progress—it is a commitment to epistemic integrity and ethical clarity.

Having now examined four distinct conceptual limitations—use without grounding, discourse without scorekeeping, fluency misread as agency, and simulation mistaken for sentience—we turn next to the synthesis. How do these lenses interlock? And what might they together reveal about the layered nature of alignment?

4. Philosophical Synthesis: From Four Lenses to One Diagnostic Framework

Understanding large language models (LLMs) through any single lens—technical, ethical, or interpretive—is insufficient. Their impacts unfold across multiple layers of meaning, design, and belief. This paper has argued that an effective conceptual framework must address not just what LLMs can do, but what they are *taken to be*, and how that shapes their development and use.

The four philosophical perspectives explored in Section 3—Wittgenstein, Lewis, Dennett, and Nagel—each diagnose a distinct limitation in generative AI. Taken together, they form a layered framework for understanding the ontological gaps, design constraints, and interpretive risks that accompany LLM deployment:

- **Wittgenstein** reveals that LLMs lack social grounding. They simulate linguistic participation without joining the communal, embodied practices that give language its meaning.
- **Lewis** shows that they struggle with conversational coherence. LLMs fail to track evolving context in a norm-sensitive way, leading to contradiction, forgetfulness, and drift.
- **Dennett** explains why users over-ascribe agency. LLMs invite the intentional stance, and without careful boundaries, this heuristic becomes confused with attribution of mind.
- **Nagel** delineates the hard boundary of consciousness. No matter how fluent an LLM's output, it does not possess subjective experience. Simulation cannot stand in for sentience.

This framework yields a central insight: the challenges of AI alignment are multi-dimensional. What appears to be a design problem (e.g., context loss) may also be a cognitive illusion (e.g., anthropomorphic overtrust), or an ethical misclassification (e.g., assuming moral standing). Philosophical clarity is not a luxury—it is a precondition for trustworthy systems.

4.1 Diagnosing by Layer: A Summary Matrix

The table below synthesizes the four perspectives into a diagnostic matrix, identifying not just symptoms and mitigations, but the conceptual domains each critique targets:

Table 4.1

Philosophical Lens	LLM Constraint	Behavioral Symptom	Current Mitigation Path	Open Research Question
Wittgenstein	Lack of social grounding	Hallucinated norms; pragmatic brittleness	Co-player simulations; feedback-rich RLHF	How rich or diverse must synthetic interaction be to meaningfully approximate grounding?
Lewis	Statelessness; context drift	Contradictions; forgotten assumptions	Hierarchical RAG; positional reweighting	Can long-context memory and score-sensitive retrieval replicate dynamic conversational norms?
Dennett	Stance inflation	Over-trust; belief in agency	Epistemic cues; persona limits; UX disclaimers	Which interface strategies best reduce anthropomorphic projection while preserving usability?
Nagel	Absence of consciousness	Misattributed moral status; empathic misreading	Role restrictions; simulation transparency; claim falsifiability	What empirical tests or structural thresholds could meaningfully falsify proto-consciousness claims?

This layered model resists the temptation to reduce AI's limitations to a single failure mode. Instead, it identifies distinct *axes of misalignment*—semantic, pragmatic, epistemic, and moral—and calls for tailored responses across architecture, interaction design, policy, and public discourse.

4.2 Inter-Lens Tensions: Productive Friction, Not Contradiction

While these lenses are complementary, they are not always harmoniously aligned. In fact, their productive tensions enrich the framework:

- Dennett's pragmatism encourages us to treat systems *as if* they had beliefs, for functional reasons. Yet Nagel warns that this move risks ontological confusion—mistaking performance for possession. Should designers highlight intentional fiction for usability, or suppress it to protect epistemic boundaries?
- Lewis describes scorekeeping as a cognitively distributed process. Wittgenstein, by contrast, emphasizes cultural embeddedness and lived practice. This raises the question: can we build systems that track *context* without being embedded in *community*? Is contextual fidelity sufficient, or must it be socially situated?
- Meanwhile, posthumanist critics (e.g., Haraway, Barad) might challenge both perspectives—arguing that intelligence and identity are already hybrid and relational, not bounded by humanist norms. This invites deeper scrutiny into whether some philosophical distinctions may reflect normative commitments rather than universal truths.

These tensions do not weaken the framework. On the contrary, they prevent it from becoming a doctrinaire checklist. Alignment is not only multi-layered—it is philosophically plural. The synthesis model aims not to resolve every tension but to surface them as sites of deliberation for designers, ethicists, and regulators.

4.3 Moving from Analysis to Action

Each layer of critique maps to a distinct stakeholder concern:

- **Engineers** must address coherence and contextual fidelity (Lewis), implement epistemic transparency (Dennett), and clarify persona boundaries (Nagel).
- **Designers** must frame outputs to prevent misattribution (Dennett), avoid simulated presence in high-stakes settings (Nagel), and build feedback mechanisms that emulate norm formation (Wittgenstein).
- **Policymakers** must ensure that technical performance is not misread as moral capacity, and that anthropomorphic claims are regulated (Nagel, Dennett).
- **Philosophers and ethicists** must continue interrogating not only what LLMs *lack*, but what we risk losing when we treat simulation as substitution.

The remainder of this paper operationalizes the framework: Section 5 translates these conceptual insights into actionable design patterns, stakeholder-specific strategies, regulatory crosswalks, and an empirical research agenda.

Bridging Forward

No single technical fix can resolve the layered challenges identified in this framework. Each philosophical lens highlights a distinct domain of misalignment—semantic, pragmatic, epistemic, or moral—and demands tailored responses from different communities of practice.

- **Designers** must prototype co-player ecosystems and feedback-rich interfaces that simulate grounding without overpromising agency.
- **Researchers** must develop metrics for context retention, stance calibration, and perception of boundaries.
- **Policymakers** must implement governance strategies that distinguish between functional capability and unjustified attributions of moral patiency.

Section 5 translates this synthesis into action—offering design principles, stakeholder guidance, regulatory crosswalks, and a forward-looking research agenda that bridges critique with consequence.

5. From Critique to Consequence: Counterarguments, Implications, and Stakeholder Guidance

This section translates the philosophical framework developed in Sections 3 and 4 into applied guidance. It unfolds in six parts: design principles, stakeholder-specific actions, counterarguments, regulatory frameworks, research directions, and broader societal reflections. Each part moves from conceptual diagnosis to pragmatic consequence.

5.1 Design Principles & Pattern Library

The following design principles operationalize the diagnostic matrix. Each principle responds to a distinct domain of misalignment—semantic grounding (Wittgenstein), contextual coherence (Lewis), agency inflation (Dennett), or mistaken moral attribution (Nagel). Together, they provide a toolkit for building systems that are transparent in their simulation, epistemically humble, and resistant to over-interpretation.

- **Simulate grounding without overclaiming it**

Embed LLMs in feedback-rich, co-player environments where they interact with other agents or users over time. This scaffolds more responsive behavior while preserving ontological clarity.

- **Make context continuity visible**

Provide users with a dynamic memory pane or conversation timeline that displays what the system is tracking, forgetting, or reprioritizing. This supports Lewisian coherence and trust calibration.

- **Reveal what the model is attending to**

Surface token retrievals, memory calls, or RAG citations to show users the informational basis of current outputs. This reduces hallucination opacity and supports error checking.

- **Constrain persona and tone in sensitive domains**

Limit informal affective language and self-referential phrasing (“I understand,” “I remember”) in domains like law, healthcare, and education. Consistency of role and tone clarifies function over fiction.

- **Epistemically frame explanations**

When using saliency maps, chain-of-thought outputs, or visualizations, accompany them with context—reminding users that these are not signs of reasoning or belief, but heuristic tools.

- **Use interface-level cues to signal simulation**

Apply visual signals—neutral avatars, tooltip disclosures, or modal chips (“Generated by AI”)—to interrupt the automatic adoption of the intentional stance. Especially critical in emotionally charged exchanges.

- **Build for refusal, not just fluency**

LLMs should be empowered to refuse answers in contexts where their training or coherence degrades. This acknowledges limitations and builds epistemic trust.

These design patterns reinforce one of the paper’s central themes: alignment is not only about capability—it is about clarity. Simulating competence is not the same as possessing it. Well-designed interfaces can help users make that distinction.

5.2 Stakeholder Mapping: Lenses to Action

Each philosophical critique maps onto specific responsibilities for four stakeholder groups. The table below summarizes how these perspectives guide the practical obligations of engineers, policymakers, ethicists, and philosophers:

Table 5.2

Stakeholder	**Wittgenstein(Use & Context)**	**Lewis(Scorekeeping & Coherence)**	**Dennett(Stance & Interpretation)**	**Nagel(Consciousness & Boundaries)**
Engineers	Fine-tune on diverse usage data; incorporate feedback loops	Implement memory-augmented and score-sensitive retrieval systems	Use neutral tone and constrain personas	Avoid roles requiring awareness, care, or moral reasoning
Policymakers	Mandate training data disclosure	Require clear communication of memory limits and context scope	Regulate anthropomorphic framing; mandate disclaimers	Prohibit unsupervised LLM use in high-affect or high-risk

Stakeholder	**Wittgenstein(Use & Context)**	**Lewis(Scorekeeping & Coherence)**	**Dennett(Stance & Interpretation)**	**Nagel(Consciousness & Boundaries)**
	and transparency			domains
Ethicists & Philosophers	Examine language norm shifts and concept drift in AI use	Analyze norm-tracking implications of memory and coherence systems	Interrogate agency projection and its social effects	

This mapping shows that alignment requires **shared conceptual grounding**, not just technical consensus. The risks posed by LLMs are not limited to architecture—they are social, cultural, and ethical.

5.3 Addressing Counterarguments and Alternate Frameworks

This framework is deliberately diagnostic, not doctrinaire. It invites engagement with competing views that challenge its scope or assumptions. Below, we surface several such perspectives and articulate why the core structure of this paper remains resilient.

5.3.1 Emergent Understanding: The Optimist's View

Some AI researchers argue that LLMs are already exhibiting *functional* understanding—e.g., abstract reasoning, metaphor generation, or multi-modal generalization. From this view, meaning arises from use, regardless of underlying architecture.

Response: This framework acknowledges emergent capability while maintaining a conceptual distinction: simulation is not possession. Without embodiment, iterative norm correction, or subjective stakes, fluency remains behaviorally impressive—but epistemically shallow. The bar for understanding must include more than output resemblance.

5.3.2 Functionalist Equivalence

Others argue that if a system can functionally pass as a participant in a language game, then debates about "real" grounding are irrelevant. If it walks like a duck…

Response: Wittgenstein and Dennett remind us that use matters—but not all "as if" performances are equivalent. In domains with high epistemic or ethical stakes, the distinction between simulation and understanding has practical implications for risk, trust, and accountability. We can model *as if*—but we must frame *as if*.

5.3.3 Posthumanist Challenges

Critical theorists and posthumanist thinkers question whether the human-machine distinction is itself too rigid. Haraway, Barad, and others argue for relational ontologies in which cognition is distributed and agency hybrid.

Response: These critiques are welcome—and important. This framework offers a bounded tool, not a totalizing ontology. It is useful precisely because it marks distinctions clearly where current discourse tends to blur them. But in future work, especially as AI systems become more embedded in socio-material networks, these perspectives must be integrated more directly.

5.3.4 Alternative Philosophical Anchors

Searle, Dreyfus, Clark, and Chalmers each offer theories that could replace or supplement the four-lens model. The Chinese Room, embodied cognition, predictive processing, and global workspace theory all bring useful provocations.

Response: The framework presented here does not reject those views—it builds on them. It selects four figures (Wittgenstein, Lewis, Dennett, Nagel) because they map cleanly onto specific misinterpretations and misalignments currently manifest in LLM behavior and deployment. Future iterations may expand this base.

5.4 Policy & Regulator Crosswalk

The philosophical limitations identified in this framework—lack of grounding, context loss, stance inflation, and simulation mistaken for sentience—map directly onto **policy and governance gaps**. The table below links each failure domain with specific regulatory mechanisms, clarifying how governance can address not just model behavior, but user interpretation and social impact.

Table 5.4

Policy Framework	**Mandate**	**LLM Challenge Addressed**	**Actionable Guidance**
EU AI Act (2024), Title IV	Disclosure of AI use; transparency of training data	Wittgenstein & Dennett: simulated grounding; anthropomorphic design	Embed model source & update info in UI; expose training domain to users via tooltips
FTC Dark Pattern Guidance (2022)	Prohibits deceptive or manipulative design	Dennett: stance inflation, misinterpreted agency	Require opt-out mechanisms and interface disclaimers in affective LLM deployments
NIST AI Risk Management	Risk categorization; lifecycle controls	Lewis & Nagel: context loss, over-	Log memory/retrieval traces; require falsifiability criteria

Policy Framework	Mandate	LLM Challenge Addressed	Actionable Guidance
Framework (2023)	for validity and auditability	ascribed sentience	for consciousness claims
AMA, ABA, APA Codes of Conduct	Limits on AI autonomy in high-risk professions	Nagel: inappropriate simulation of care or expertise	Require human-in-the-loop oversight for outputs in clinical, legal, or psychological settings
OECD AI Principles (2019)	Human agency and accountability in AI systems	Dennett: tool-agent confusion	Mandate audit trails; clarify chain-of-responsibility in decision-support workflows

These frameworks share a common purpose: they treat **simulation transparency** as a public good. Philosophical insight here becomes governance infrastructure. Fluency without clarity is not competence—it is risk.

5.5 Empirical Research Agenda

This framework raises not just philosophical questions, but empirical ones. If simulation is not possession, and if alignment must address interpretive as well as functional risks, how can we test those boundaries? The following research directions map to each of the four conceptual domains:

5.5.1 Wittgenstein – Grounding and Pragmatic Use

- **Agent Diversity Threshold**

How many distinct co-players or interactive hours are required for an LLM to stabilize its pragmatic use of language?

Method: Multi-agent simulations with variation in user goals, language games, and feedback regimes.

5.5.2 Lewis – Coherence, Memory, and Norm-Tracking

- **Scorekeeping Robustness**

How well do memory-augmented vs. RAG-based models maintain conversational state across topic shifts and corrections?

Method: Contradiction detection, cross-turn coherence metrics, and epistemic integrity benchmarks.

- **Presupposition Reconciliation**

Can models revise or retract prior claims when user inputs invalidate assumptions?

Method: Structured dialogue tests with scripted reversals and ambiguous corrections.

5.5.3 Dennett – Stance Calibration and Anthropomorphism

- **Interface Cues and Attribution Study**

Which combinations of disclaimers, avatars, and pronouns most reliably reduce user over-attribution of agency or emotion?

Method: A/B tests across interface variants with post-interaction trust and empathy surveys.

- **Explainability Framing Effects**

Do saliency maps or chain-of-thought traces increase the illusion of cognition?

Method: Experimental design with control groups comparing explanation tools with and without epistemic framing.

5.5.4 Nagel – Sentience Claims and Consciousness Boundaries

- **Empirical Falsifiers**

What experimental prompts or behavioral stressors could falsify claims of proto-consciousness in LLMs?

Method: Adapt cognitive science paradigms—mirror tests, self-contradiction detection, affective blindsight analogues.

- **Cross-Cultural Misinterpretation Studies**

How do users in different linguistic and cultural settings interpret LLM-generated statements of emotion, care, or selfhood?

Method: Mixed-methods field research across global user bases.

Together, these projects would help quantify epistemic illusion, test conceptual claims, and clarify design limits. This is not just research for better models—it is research for better *interpretation*.

5.6 Societal and Ethical Reflections

Beyond design, policy, and research, the philosophical limitations of LLMs raise urgent ethical and civic questions. What kind of society are we building if simulation becomes indistinguishable from participation? If users mistake tools for minds—or comfort for care—what responsibilities follow?

5.6.1 Reasserting Human Accountability

When users interpret LLM outputs as autonomous, moral agency is displaced. Designers become invisible. Institutions outsource judgment. Dennett and Nagel remind us: the system does not know what it is doing. Humans do. Responsibility must trace back to those who train, deploy, and profit from these systems—not the systems themselves.

5.6.2 Protecting Emotional Vulnerability

The risk is not just over-trust in facts—it is **over-trust in affect**. In domains like grief support, therapy, or education, LLMs can appear emotionally attuned. But they lack memory, perspective, or care. This is not empathy—it is affective simulation. Transparency here is not optional. It is ethical infrastructure.

5.6.3 Linguistic Justice and Cultural Pluralism

Language is not neutral. It encodes culture, history, and power. LLMs trained on dominant corpora risk marginalizing alternative idioms and forms of life. Wittgenstein's framework shows that meaning is always local. Cultural fine-tuning is not cosmetic—it is epistemic alignment with plural communities.

5.6.4 Civic Literacy as AI Governance

Trustworthy AI requires not just technical audits, but civic literacy. Users must understand what LLMs are, what they are not, and how their outputs are shaped. Interpretive confusion—mistaking performance for perspective—undermines democratic discourse. Public understanding of AI must be treated as part of democratic infrastructure, akin to data privacy or access to broadband.

5.6.5 Rethinking the Human

Finally, these questions circle back on us. If LLMs can appear creative, persuasive, or emotionally rich—what is it we value in human cognition? In human presence? Wittgenstein, Lewis, Dennett, and Nagel do not offer nostalgia—they offer clarity. They remind us that understanding is not fluency; care is not expression; presence is not simulation. To value the human, we must understand what the machine is not.

6. Conclusion: Open Questions and Practical Next Steps

This paper has argued that large language models (LLMs) like GPT-4o should be understood not as minds, agents, or participants—but as powerful simulations. Their linguistic fluency can evoke understanding, coherence, care, and even presence—but these are performances, not possessions. To mistake simulation for cognition is not merely a conceptual error—it is a design risk, a policy failure, and an ethical hazard.

Drawing on Wittgenstein, Lewis, Dennett, and Nagel, the framework presented here diagnoses four distinct—but overlapping—limitations:

- A **lack of grounding** in shared forms of life (Wittgenstein)
- A **fragile grasp of conversational context** and evolving norms (Lewis)
- A tendency to invite **anthropomorphic projection** and over-ascription (Dennett)
- A fundamental absence of **subjective experience or consciousness** (Nagel)

Together, these critiques reveal that alignment is not a single technical problem, but a multi-layered challenge—spanning semantic grounding, pragmatic coherence, interpretive caution, and moral boundary-setting.

6.1 Open Research Questions

Philosophical clarity now demands empirical traction. The following research questions, introduced in Section 5.5, remain urgent:

- How many co-players or interaction hours are needed for an LLM to approximate domain-sensitive language game rules?
- Can long-context or memory-augmented systems sustain scorekeeping over multi-turn dialogue with dynamic revisions?
- Which UX design patterns reduce over-ascription of agency or emotion without diminishing user engagement?
- What empirical thresholds or tests could falsify (rather than merely speculate on) claims of emergent consciousness?

These questions are not only technical—they are conceptual probes. They test whether performance can ever cross the threshold into possession, and how we might know when it hasn't.

6.2 Broader Implications

The stakes are not confined to model design. They touch the social fabric:

- **Regulators** must distinguish performance risk from **interpretive risk**—ensuring that policy reflects both what AI can do and what humans believe it can do.
- **Designers** must surface memory, mark simulation, and calibrate stance to **protect user understanding**, not just optimize engagement.
- **Researchers** must complement capability benchmarks with metrics for **epistemic robustness** and **moral clarity**.
- **Public institutions** must foster AI literacy as a civic obligation. Misunderstanding the machine is not just a private confusion—it is a public harm.

6.3 Final Thought: Simulation Is Powerful—but It Is Not Mind

LLMs are remarkable artifacts. They compress the textual archive of human thought into accessible interfaces. They assist, predict, reframe, and remix. But they do not *understand*, *intend*, or *care*. They simulate what it is like to be articulate—but there is nothing it is like to be them.

Treating them accordingly is not an act of pessimism—it is an act of precision. Philosophical clarity is not a luxury for technologists or regulators. It is the precondition for alignment, trust, and responsibility in a world increasingly shaped by generative systems.

What we do next depends not only on what these models are—but on what we are willing to see clearly about what they are not.

Ethics, Disclosure, and Acknowledgements

Ethical Considerations

This paper does not draw on private, sensitive, or personally identifiable data. All examples are hypothetical, anonymized, or derived from public sources. No formal human-subjects research was conducted, and no institutional ethics review was required. All citations conform to academic standards.

The broader ethical implications of the arguments developed herein concern public misinterpretation, policy design, and stakeholder responsibility in AI deployment. These implications are intended to provoke critical discussion and inform future regulatory and design frameworks.

Use of AI Tools

AI language models—most notably OpenAI's ChatGPT—were used during the writing process as interlocutors: for brainstorming, structuring sections, and testing rhetorical clarity. These tools were instrumental in refining transitions, surfacing edge cases, and challenging internal consistency.

This meta-use aligns with the paper's themes. Interacting with generative AI during authorship provided firsthand insight into the very limitations this paper analyzes: fluency without grounding, responsiveness without perspective, and the ease with which stylistic coherence can be mistaken for conceptual depth.

Responsibility for all ideas, arguments, and conclusions lies solely with the human author.

Acknowledgments

The author wishes to thank informal readers who provided critical feedback on earlier drafts. Their questions, challenges, and encouragement materially improved the final manuscript. Special thanks to those who questioned assumptions, pushed for clearer synthesis, and reminded the author that philosophy and engineering are not separate disciplines—they are simply perspectives on design.

No institutional support, funding, or affiliation contributed to this work. All errors and omissions are the author's alone.

Disclosure Statement

This work was conducted independently, without institutional affiliation, funding, or external influence. The views expressed are the author's alone and do not represent any current or former employer. No financial or professional conflicts of interest are declared.

License & Attribution

To cite this paper:

Stoyanovich, Michael. *Philosophy, Cognitive Science, and Policy: Interdisciplinary Perspectives on Generative AI from Wittgenstein, Lewis, Dennett, and Nagel*. Version 1.23.0 (June 2025).

https://www.mstoyanovich.com

Version History and Document Status

This is a living document. As generative AI systems evolve, this paper will be periodically updated to incorporate new empirical findings, theoretical insights, and policy developments. Major revisions are recorded here to preserve transparency and scholarly traceability.

Version	Date	Description
V1.23.0	June 2025	Final editorial revision. Incorporated multiple-rounds of critiques; rewrote all of Sections 3–6 for tone, clarity, and counterargument integration; streamlined redundancy; added lens tension synthesis in Section 4; expanded societal reflections; revised ethics, disclosure, and acknowledgments
V1.22.3	June 2025	Integrated editorial feedback; full rewrite of Sections 3–6
V1.22.2	June 2025	Removed legacy Table/Figure duplication; converted all tables; final copy-edits
V1.22.1	June 2025	Expanded literature review; added multi-layer alignment synthesis; reorganized Section 5; polished conclusion
V1.22.0	June 2025	Major structural revision. Merged theoretical and applied sections (3 and 5); eliminated redundant Section 4 ("Concept to Application"); streamlined glossary, tables, and roadmap; added mini case studies; revised all section numbers accordingly; comprehensive refinement of tone, synthesis, and rhetorical flow
V1.21.7	June 2025	Full integration of rewritten philosophical framework (Sections 3.1–3.5); major rework of Sections 4–6; refined synthesis and discussion
V1.21.6	June 2025	Reorganized philosophical framework; integrated FILM-7B findings into Lewis section; updated glossary; added Table 4-1 and external figure
V1.21.5	June 2025	Incorporated new empirical work (An et al., 2024) on long-context QA and VAL probing; revised Sections 3.2 and 5.2
V1.21.4	May 2025	Rewrote Introduction for improved framing and accessibility; standardized formatting; updated citations
V1.21.3	April 2025	Added interdisciplinary synthesis section (3.5); revised Discussion and Counterarguments sections

Version	**Date**	**Description**
V1.21.2	March 2025	Structural alignment with interdisciplinary audience; initial draft of Sections 4–6
V1.21.1	Feb 2025	Substantial conceptual expansion from earlier drafts; added individual philosopher sections
V1.0.0	Jan 2025	Initial release of *Philosophy, Cognitive Science, and Policy: Interdisciplinary Perspectives on Generative AI from Wittgenstein, Lewis, Dennett, and Nagel*

References

An, S., Ma, Z., Lin, Z., Zheng, N., & Lou, J.-G. (2024). Make your LLM fully utilize the context. *arXiv*. https://arxiv.org/abs/2404.16811

An, Z., Dong, X., & Lee, J. (2025). Multiple abstraction level retrieve-augment-generate (MAL-RAG). *arXiv*. https://arxiv.org/abs/2501.16952

Anthropic. (2023). Constitutional AI: Harmlessness from AI feedback. Anthropic. https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165 [oai_citation:0‡arxiv.org](https://arxiv.org/abs/2005.14165?utm_source=chatgpt.com)

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. M., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. https://arxiv.org/abs/2303.12712

Clark, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*. Oxford University Press.

Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Coeckelbergh, M. (2020). *AI ethics.* MIT Press.

Dennett, D. C. (1989). *The intentional stance.* MIT Press.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608

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

European Union. (2024). AI Act: Title IV—Transparency obligations. European Union. https://artificialintelligenceact.eu

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council. *Official Journal of the European Union, L 168*, 1–158. https://eur-lex.europa.eu/eli/reg/2024/1689

Federal Trade Commission. (2022). *Bringing dark patterns to light: Staff report*B. Author. https://www.ftc.gov/system/files/ftc_gov/pdf/P214800%20Dark%20Patterns%20Report.pdf

Floridi, L. (2011). *The philosophy of information*. Oxford University Press.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138. https://doi.org/10.1038/nrn2787

Haraway, D. J. (1991). 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.

Haugeland, J. (1985). *Artificial intelligence: The very idea*. MIT Press.

Hooker, S. (2021). Explanations alone cannot prevent algorithmic harm. *arXiv*. https://arxiv.org/abs/2107.00154

James, W. (1907). *Pragmatism: A new name for some old ways of thinking*. Longmans, Green and Co.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic, 8*(1), 339–359. https://doi.org/10.1007/BF00258436

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 11*, 1333–1359. https://doi.org/10.1162/tacl_a_00533

Liu, S., Xie, K., & Sun, M. (2024). How language models use long contexts better via plug-and-play positional re-weighting. In *Proceedings of the Twelfth International Conference on Learning Representations* (ICLR 2024). https://openreview.net/forum?id=fPmScVB1Td

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). https://doi.org/10.1145/2858036.2858288

MITRE. (2022). *Chatbot accessibility playbook*. MITRE Corporation. https://www.mitre.org

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building artificial intelligence we can trust.* Pantheon.

Mitchell, M. (2023). *Artificial intelligence: A guide for thinking humans* (Updated ed.). Penguin Books.

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology, 33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Nagel, T. (1974). What is it like to be a bat? In *Mortal questions* (pp. 165–180). Cambridge University Press.

National Institute of Standards and Technology. (2023). *AI risk management framework 1.0* (NIST SP 1270). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3* (3), 417–457. https://doi.org/10.1017/S0140525X00005756

Shanahan, M. (2022). *Embodiment and the inner life: Cognition and consciousness in the space of possible minds*. Oxford University Press.

Singer, P. (1975). *Animal liberation: A new ethics for our treatment of animals.* Harper & Row.

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other.* Basic Books.

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

Varela, F. J., Thompson, E., & Rosch, E. (1992). *The embodied mind: Cognitive science and human experience.* MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). https://arxiv.org/abs/1706.03762

Wittgenstein, L. (1953). *Philosophical investigations* (G. E. M. Anscombe, Trans.). Blackwell Publishing.

Yuan, J., Dong, S., & Adelson, E. H. (2024). Multimodal tactile sensing fused with vision for dexterous robotic manipulation. *Nature Communications, 15,* 51261. https://doi.org/10.1038/s41467-024-51261-4

Further Reading

These sources complement the core arguments developed in this paper by extending into adjacent domains—posthumanism, sociotechnical critique, interpretability, phenomenology, and cognitive science. Each entry is annotated to highlight its relevance to the philosophical and practical stakes of generative AI.

Books

- Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Anchor Books.

  → Foundational work in social constructivism. Reinforces Wittgensteinian in sights into meaning as socially co-constructed.

- Barad, K. (2007). Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning. Duke University Press.

  → Introduces agential realism, a relational ontology that challenges subject–ob ject distinctions. Offers a radically different lens for thinking about AI agency and sociomaterial entanglement.

- Clark, A. (2015). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

  → Predictive processing model of cognition. Supports arguments about embodi ment, environmental coupling, and the limitations of disembodied statistical in ference in LLMs.

- Hayles, N. K. (1999). *How we became posthuman: Virtual bodies in cybernetics, literature, and informatics*. University of Chicago Press.

  → Influential text linking posthumanism, cognition, and the cultural dimensions of technology.

- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row, Publishers. (Original work published 1927)

  → Phenomenological critique of representational thinking. Lays groundwork for later critiques of symbolic AI (e.g., Dreyfus, Varela).

- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.

  → Technical foundation of information theory. Offers a counterpoint to prag matic and use-based theories of language and meaning.

- Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Alfred A. Knopf.

→ A futurist perspective on AI development and long-term governance. Provokes questions about alignment, embodiment, and consciousness.

- Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other*. Basic Books.

  → A sociological critique of human relationships with digital companions. Highlights risks of over-trust and emotional misattribution.

- Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

  → Brings Aristotelian virtue ethics to technology design and use. A powerful normative complement to the paper's concerns about responsible AI.

- Winner, L. (1986). *The whale and the reactor: A search for limits in an age of high technology*. University of Chicago Press.

  → Classic work exploring the politics embedded in technological artifacts. Useful for understanding the sociotechnical stakes of LLM deployment.

Book Chapters

- Haraway, D. J. (1991). 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.

  → A foundational posthumanist critique. Challenges binary distinctions between human and machine—relevant to discussions of AI agency and hybridity.

Journal and Conference Papers

- Luger, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297. https://doi.org/10.1145/2858036.2858288

  → Empirical study revealing how users overestimate AI competence. Supports concerns about the invisible fiction of the intentional stance.

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

  → Seminal work in explainable AI. Introduces the LIME method and provides a foundation for LLM interpretability research.

Preprints and Technical Reports

- Bommasani, R., Hudson, D. A., et al. (2021). *On the opportunities and risks of foundation models*. arXiv:2108.07258. https://arxiv.org/abs/2108.07258

→ Introduces the "foundation model" framing. Widely cited in discussions of scale, generalization, and alignment.

- Graves, A., Wayne, G., & Danihelka, I. (2014). *Neural Turing machines*. arXiv:1410.5401. https://doi.org/10.48550/arXiv.1410.5401

  → Proposes memory-augmented neural architectures. Related to critiques of LLM statelessness and long-term coherence.

- Spiegel, O., Kleiman-Weiner, M., & Tenenbaum, J. B. (2024). *Visual theory of mind*. arXiv:2401.15175. https://arxiv.org/abs/2401.15175

  → Investigates multimodal inference of mental states. Intersects with Dennettian themes around simulation, agency, and stance attribution.

Glossary of Key Terms

This glossary summarizes key conceptual terms used throughout the paper, spanning philosophy, AI design, interface framing, and empirical evaluation.

Philosophical and Interpretive Concepts

- **Language Game**

Wittgenstein's term for how meaning is derived from socially embedded, context-sensitive language use within shared "forms of life."

- **Scorekeeping**

David Lewis's idea that conversation evolves by updating a contextual "score" of shared assumptions, roles, and norms—making communication history-sensitive.

- **Intentional Stance**

Dennett's interpretive strategy: treating a system as if it has beliefs or goals to predict its behavior, even if such states are not literally present.

- **Intentional Fiction**

The use of the intentional stance as a predictive heuristic, not a claim about inner states —highlighting the danger of mistaking interpretive utility for ontological fact.

- **Stance Inflation**

The progressive escalation of perceived agency or emotion in LLMs, often triggered by fluency, explanations, or human-like personas.

- **Nagel Test**

A philosophical boundary prompt inspired by Thomas Nagel's "What is it like to be a bat?" Used to assess whether a system could plausibly possess first-person, subjective experience.

- **Simulation Ceiling**

The conceptual boundary beyond which behavioral mimicry cannot cross into genuine experience or sentience. Distinguishes performance from being.

- **Simulation vs. Instantiation**

The distinction between mimicking a behavior (simulation) and actually possessing the capacity or state being mimicked (instantiation).

- **Epistemic Illusion**

A mistaken belief about what an AI system understands or knows, arising from the surface coherence of its output.

- **Epistemic Framing**

Design strategies that explicitly signal the statistical or non-agentic nature of LLM responses—used to prevent misinterpretation and over-trust.

- **Moral Patiency**

The ethical status of being a recipient of moral consideration. Applied here to explore whether non-sentient systems warrant obligations typically reserved for conscious beings.

- **Anthropomorphic Creep**

The gradual, often unintentional tendency of users to perceive non-sentient AI systems as agentic, emotional, or conscious due to interface cues and conversational design.

- **Phenomenology**

The philosophical study of first-person experience and consciousness. Invoked here to distinguish real sentience from behavioral simulation.

Technical and Architectural Terms

- **Transformer Architecture**

A neural model design based on attention mechanisms, enabling parallel processing of entire sequences. Foundational to modern LLMs like GPT.

- **Statelessness**

A design feature in which each prompt-response pair is treated in isolation, without retained memory of prior turns unless engineered into the system.

- **Token**

The smallest unit of input (e.g., word fragment, punctuation) processed by LLMs. Model context windows are typically measured in tokens.

- **Few-shot Learning**

A prompting method in which the model generalizes from a small number of input-output examples provided at inference time.

- **Fine-tuning**

Post-training model refinement using curated datasets to improve domain-specific performance.

- **RLHF (Reinforcement Learning from Human Feedback)**

A training approach that aligns LLM behavior with human preferences by using human-rated comparisons as reward signals.

- **IN2 Training**

A data-centric training method that improves attention to mid-sequence information by varying token positioning across training sequences.

- **Lost-in-the-Middle**

A context window failure pattern where tokens near the center of long sequences receive reduced attention, leading to degraded coherence or omissions.

Interpretability and Cognitive Framing Constructs

- **Score-Sensitive Retrieval**

A retrieval technique that prioritizes memory elements based on their relevance to the evolving conversational context, not just keyword similarity.

- **Heuristic**

A simplifying rule or approximation strategy used by both human agents and AI models to reduce complexity in reasoning or prediction.

- **Emergent Behavior**

Unprogrammed capabilities that arise in large-scale models as a function of scale or training complexity—such as tool use or multi-step reasoning.

- **SOAR**

A symbolic cognitive architecture that models general intelligence via modules for learning, planning, and memory. Historically used in AI and psychology.

- **ACT-R (Adaptive Control of Thought—Rational)**

A modular cognitive framework simulating human reasoning using declarative and procedural memory. Applied in cognitive science and human–AI modeling.