Philosophy, Cognitive Science, and Policy:

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

Michael Stoyanovich
Artifact Classification: Practitioner Paper
Version 1.26.1 | November 2026

Scope and Limitations

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.

Purpose

This paper develops an interdisciplinary framework for understanding the capabilities and limitations of contemporary large language models through the complementary philosophical perspectives of Wittgenstein, Lewis, Dennett, and Nagel. It is intended to support AI governance, system design, public policy, risk management, and responsible organizational adoption by distinguishing simulated linguistic performance from the human capacities it may resemble.

Why This Matters Now

With the EU AI Act entering into force on August 1, 2024 (with obligations phased in over time) 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.

Intended Audience

- AI engineers and product managers
- Policymakers and regulators
- Ethicists and social scientists
- Designers and technical communicators

- Educators and critically engaged lay readers

Reading Guide

This paper develops its argument across several interconnected philosophical, cognitive-scientific, and policy perspectives. Readers seeking an overview of the paper's structure and the relationship among its major sections should begin with the roadmap in Section 1.1. Readers may also use the framework's individual philosophical lenses independently where a particular interpretive or governance question is of primary interest.

Table of Contents

Scope and Limitations....1
Purpose....1
Why This Matters Now....1
Intended Audience....1
Reading Guide....2
Abstract....6
Executive Summary....6
1. Introduction....7
  1.1 Roadmap....8
2. Literature Review....9
  2.1 Philosophical Foundations of AI....9
    2.1.1 Predictive Processing & Active Inference....9
    2.1.2 Cognition in AI–Robotics Experiments....10
    2.1.3 Symbolic Resurgence & Neuro-symbolic Hybrids....10
  2.2 Wittgenstein's Philosophy and AI....11
    2.2.1 When AI Enters the Language Game: A Tractatus-to-Investigations Bridge....12
  2.3 David Lewis and Contextual Dynamics....12
  2.4 Dennett's Intentional Stance and AI....13
    2.4.1 Interface Metaphors and the Chatbot Form....14
  2.5 Nagel's Challenge to AI Consciousness....15
  2.6 Integration of Contemporary Debates and Broader Perspectives....16
    2.6.1 Post-humanism and AI....17
    2.6.2 Critical Theory and Sociotechnical Context....17
    2.6.3 Embodied Cognition, Cognitive Architectures, and Mind-Body Framing....17
    2.6.4 Policy and Ethics Preview....18
3. A Philosophical Framework and Its Application....19
  3.1 Wittgenstein's Language Games and the Conceptual Boundaries of AI Comprehension..20
    3.1.1 Philosophical Foundation....20
    3.1.2 Ontological Limitations: Use Without Participation....21
      3.1.2.1 Statistical Imitation vs. Communal Rule-Following....21
      3.1.2.2 Static Corpora vs Dynamic Correction....21
      3.1.2.3 Fluency Without Pragmatic Stakes....21
    3.1.3 Counterpoint: Emergent Game Competence?....22
    3.1.4 Design and Governance Implications....23
      3.1.4.1 Simulated Feedback and Iterative Alignment....23
      3.1.4.2 Semantic Localization Through Cultural Fine-Tuning....23
      3.1.4.3 Toward Partial Embodiment....23
      3.1.4.4 Transparency by Design....23
      3.1.4.5 Which Language Games Are More "Simulable," and Why That Still Matters...23
    3.1.5 Case Study: A Cross-Cultural Customer Service Bot....24
    3.1.6 Participation and Its Limits....24
  3.2 Lewis: Conversational Scorekeeping and the Architecture of Context....25

3.2.1 Philosophical Foundation....25
3.2.2 Structural Limitation: Statelessness and Context Drift....25
3.2.2.1 Context Collapse Over Time....25
3.2.2.2 Memory Without Revision....26
3.2.2.3 Absence of Norm-Tracking....27
3.2.3 Counterpoint: Advances in Long-Context Architecture....27
3.2.4 Design and Policy Implications....27
3.2.4.1 Epistemically Active Memory....27
3.2.4.2 Score-Sensitive Retrieval....27
3.2.4.3 Context Integrity Benchmarks....28
3.2.4.4 Accumulated Context and Conversational Scorekeeping....28
3.2.4.5 Interface-Level Cues....28
3.2.5 Case Study: The Forgetful Legal Assistant....28
3.2.6 Conclusion: Context Is Not Optional....29
3.3 Dennett: The Intentional Stance and the Risks of Anthropomorphism....29
3.3.1 Philosophical Foundation....29
3.3.2 Interpretive Slippage: From Heuristic to Epistemic Error....30
3.3.2.1 Pragmatic Usefulness vs. Misplaced Confidence....30
3.3.2.2 Anthropomorphic Design Choices....30
3.3.2.3 Simulation of Selfhood....31
3.3.3 Counterpoint: Critical and Functionalist Perspectives....31
3.3.4 Design and Policy Implications....31
3.3.4.1 Transparent Framing of Outputs....31
3.3.4.2 Role and Persona Constraints....32
3.3.4.3 Calibrated Explainability....32
3.3.4.4 Emotional Simulation Boundaries....32
3.3.5 Case Study: The Compassionate Chatbot Trap....32
3.3.6 Conclusion: Make the Heuristic Visible....33
3.4 Nagel: The Simulation Ceiling and the Problem of Consciousness....33
3.4.1 Philosophical Foundation....33
3.4.2 Conceptual Constraint: Fluency Does Not Equal Sentience....34
3.4.2.1 First-Person Absence....34
3.4.2.2 The Illusion of Inner Life....34
3.4.2.3 Speculative Measures Remain Speculative....34
3.4.3 Counterpoint: Open Horizons and the Ethics of Doubt....35
3.4.4 Design and Policy Implications....35
3.4.4.1 Simulated Empathy Disclosures....35
3.4.4.2 Role Restrictions in High-Affect Contexts....35
3.4.4.3 Ethical Guardrails for Sentience Claims....36
3.4.4.4 Moral Patiency Thresholds....36
3.4.5 Case Study: The Hospice Companion Bot....36
3.4.6 Conclusion: Experience Cannot Be Faked....36

4. Philosophical Synthesis: From Four Lenses to One Diagnostic Framework....37
4.1 Diagnosing by Layer: A Summary Matrix....38
4.2 Inter-Lens Tensions: Productive Friction, Not Contradiction....39
4.3 Moving from Analysis to Action....39
4.3.1 Originality, Novelty, and Authenticity....40
4.4 Limitations of This Framework....40
Bridging Forward....41
5. From Critique to Consequence: Counterarguments, Implications, and Stakeholder Guidance 42
5.1 Design Principles & Pattern Library....42
5.2 Stakeholder Mapping: Lenses to Action....43
5.3 Productive Tensions with Alternative Frameworks....44
5.3.1 Productive Tension: Emergent Understanding....44
5.3.2 Productive Tension: Functionalist Equivalence....44
5.3.3 Productive Tension: Posthumanist Challenges....44
5.3.4 Productive Tension: Alternative Philosophical Anchors....45
5.4 Policy & Regulator Crosswalk....45
5.5 Empirical Research Agenda....47
5.5.1 Wittgenstein – Grounding and Pragmatic Use....47
5.5.2 Lewis – Coherence, Memory, and Norm-Tracking....47
5.5.3 Dennett – Stance Calibration and Anthropomorphism....48
5.5.4 Nagel – Sentience Claims and Consciousness Boundaries....49
5.6 Societal and Ethical Reflections....50
5.6.1 Reasserting Human Accountability....51
5.6.2 Protecting Emotional Vulnerability....51
5.6.3 Linguistic Justice and Cultural Pluralism....51
5.6.4 Civic Literacy as AI Governance....51
5.6.5 Rethinking the Human....51
6. Conclusion: Open Questions and Practical Next Steps....51
6.1 Open Research Questions....52
6.2 Broader Implications....53
6.3 Final Thought: Simulation Is Powerful - But It Is Not Mind....53
Appendix A - Practical Implementation Checklist....55
Ethics, Disclosure, and Acknowledgements....57
Ethical Considerations....57
Use of AI Tools....57
Acknowledgements....57
Disclosure Statement....57
License & Attribution....57
Trademark Notice....58
How to Cite This Paper....58
Related Works....58
Version History and Document Status....58

References....64
Further Reading....69
Books....69
Book chapter....70
Journal & conference papers....70
Preprints & technical reports....70
Glossary of Key Terms....71
Philosophical and Interpretive Concepts....71
Technical and Architectural Terms....72
Interpretability and Cognitive Framing Constructs....72
Policy and Governance Terms....73

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 – collectively inform what I call "The Four Philosophers Framework™" (or "The 4-Philosophers Framework™")[1], a model for deepening our understanding of generative AI, particularly large language models (LLMs) such as GPT architectures operationalized by commercial, state and not-for-profit entities.

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 advanced models may approximate some functional aspects of human cognition, this paper argues that linguistic fluency and task performance alone do not establish human-like perspective, socially grounded participation, or subjective awareness.

The paper develops a four-lens diagnostic framework, distinguishes originality, novelty, and authenticity in evaluating AI-generated work, proposes actionable design strategies, and addresses counterarguments, ethical risks, and policy implications. Throughout, concepts are introduced in accessible language to engage readers across disciplines.

Executive Summary

[1] The Four Philosophers Framework™ and The 4-Philosophers Framework™ are unregistered trademarks of Michael Stoyanovich. Use of these terms to refer to this framework is permitted for non-commercial, scholarly, or descriptive purposes, but commercial use without permission may constitute trademark infringement.

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 their use raises four distinct diagnostic concerns:

- **Wittgensteinian grounding**: They can influence human language practices without thereby demonstrating the full norm-governed participation through which human communities sustain meaning.
- **Lewisian coherence**: They can maintain and retrieve aspects of conversational context, but do not consistently reconcile changing assumptions, corrections, and commitments over extended interactions.
- **Dennettian attribution:** They invite over-trust via anthropomorphic projection, despite uncertainty about whether beliefs or desires are possessed rather than interpretively attributed.
- **Nagelian interiority**: They can simulate expressions of understanding and experience without establishing that subjective experience accompanies those performances.

These limitations are not simply implementation defects. They reflect deeper conceptual mismatches between convincing simulations and the human capacities those simulations resemble. 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 test and discipline 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 rather than minds, we can guide their development and use more responsibly, ethically, and safely.

Keywords: generative AI; language games; scorekeeping; intentional stance; consciousness; embodied cognition; AI ethics; cognitive science; neuroscience; explainable AI; cognitive architectures; posthumanism; policy; global regulation; EU AI Act; enterprise copilots; 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 question, 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[2]

- **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 uncertainty about subjective experience.
- **Section 4** synthesizes these perspectives into a unified diagnostic model, maps each philosophical critique to specific system vulnerabilities, and uses the distinction among originality, novelty, and authenticity to clarify how AI-generated work should be evaluated.
- **Section 5** operationalizes the framework, offering concrete design principles, stakeholder guidance, regulatory alignment strategies, and an agenda for future research.

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

- **Section 6** concludes by reframing alignment as a multilayered challenge and showing how the four lenses preserve distinctions between representation and practical, social, interpretive, and experiential reality.

Together, these sections argue that linguistic competence, however convincing, does not by itself establish human-like understanding, and that maintaining 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 understanding - implying that syntax alone does not produce semantics. In contrast, Alan Turing's criterion for intelligence (the Turing Test, Turing, 1950) focuses on observable behavior: if a machine's responses are indistinguishable from a human's, we may as well call it intelligent, sidestepping the question of internal understanding. This tension between behaviorism and semantic internalism continues to inform debates about LLMs 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, 1991). 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 systems may develop forms of participation in language-games through sufficiently rich interaction, this paper maintains that full linguistic grounding requires more than fluent performance. The relevant conditions include participation in shared practices, responsiveness to correction, and the consequences that make rule-following normatively significant. Work by P. M. S. Hacker and Danièle Moyal-Sharrock helps illuminate Wittgenstein's emphasis on communal practices and rule-following. Applied to contemporary LLMs, this perspective raises questions about whether statistically generated language amounts to participation in the practices through which human meaning is sustained. 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.[3]

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

[3] Recent work by Spiegel et al. (2025) 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.

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.

2.2.1 When AI Enters the Language Game: A Tractatus-to-Investigations Bridge

A useful way to bridge Wittgenstein's early and later philosophy is to contrast the *Tractatus*' picture theory of language with the later view that meaning emerges through use within socially embedded language games. For the purposes of this paper, that contrast is not offered as a comprehensive interpretation of Wittgenstein. Rather, it provides a practical bridge to a governance question: what changes when LLM-generated language becomes part of human practices rather than merely an object of analysis? (Lucia, 2025).

Recent empirical work, including several early studies of AI-mediated language practices, provides concrete footholds for this "practice-shift" framing:

• In a controlled dialogue task, speakers were less likely to repeat an interlocutor's syntactic structure when they believed the partner was an AI agent rather than a human (Li, 2025).

• In comparative corpus work on argumentative essays, ChatGPT-generated texts exhibited substantially lower interactional meta-discourse (e.g., hedges, boosters, attitude markers), producing a more impersonal rhetorical stance even when structural coherence was high (Jiang & Hyland, 2025).

• In large-scale analysis of arXiv abstracts over a decade, LLM-preferred lexical markers increased post-ChatGPT alongside shifts in lexical, syntactic, cohesion, and readability

features, suggesting detectable population-level drift in academic prose (Bao et al., 2025).

• In multi-agent settings, populations of LLM agents can converge on shared conventions without explicit human instruction, illustrating a limited but real form of convention-formation under interaction (Ashery, Aiello, & Baronchelli, 2025).

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, 1987) 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.

2.4.1 Interface Metaphors and the Chatbot Form

Recent human-computer interaction (HCI) and AI ethics work strengthens the claim that anthropomorphism is not merely a user mistake; it is also a design condition. Ghosh, Venkit, Gautam, and Ghosh ask what would follow if AI systems were not primarily organized as chatbots, arguing that conversational interfaces are not neutral wrappers but sociotechnical forms that shape authority, expertise, overreliance, and intimacy. So, Cheng, and Murthy similarly propose treating anthropomorphism as a design variable rather than a binary defect, ranging from transparency-driven anti-anthropomorphism to hyper-anthropomorphic uncanniness. Empirical work by Kadambi and colleagues further shows that warmth, cognitive empathy, affective empathy, and competence influence different dimensions of trust, perceived anthropomorphism, usefulness, and relational closeness.

This literature reframes the Dennettian problem. The intentional stance is still a useful predictive shortcut, but the interface can make that shortcut feel like ontology. A chatbot does not simply receive anthropomorphic projection; it can invite, amplify, and stabilize it. The practical question for governance is therefore not only whether users overascribe mind to AI systems, but which interface metaphors make such over-ascription likely, durable, and consequential. Put differently, the interface is not merely a delivery wrapper; it helps establish the language-game in which terms such as "assistant," "memory," "care," "understanding," and "help" acquire practical force.

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?" The common intuition is that there is nothing it is like to be GPT; an LLM, as a computational tool, 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. Global Workspace Theory, by contrast, emphasizes the functional availability of information across different processing systems. Recent mechanistic interpretability research has identified workspace-like properties in some language models, including internal representations that can be reported, modulated, and used flexibly in reasoning (Gurnee et al., 2026). These findings complicate categorical claims that contemporary LLMs lack all such functional organization. They do not, however, establish phenomenal consciousness or resolve the relationship between computational access and subjective experience. Nagel's challenge therefore remains: identifying the mechanisms by which information becomes available for reasoning is not the same as establishing that there is something it is like to be the system.

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.

Recent work on AI consciousness sharpens this point by distinguishing the direct metaphysical question from the socially tractable one. Comșa argues that the question of whether AI systems can be conscious remains presently intractable, given the lack of consensus in consciousness science and the unresolved mind-body problem. By contrast, perceived AI consciousness - how users, institutions, and language communities come to treat systems as if they had inner life - is tractable, timely, and governance-relevant. This distinction reinforces the Nagelian boundary: policy does not need to settle machine consciousness in order to regulate the social effects of perceived consciousness.

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 in-

tegrating 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.

A recent extended-mind variant of this debate pushes the point further. Gutoreva, Tsim, and Papakonstantinou argue that contemporary AI increasingly participates in attention allocation, reasoning, synthesis, and decision-making, such that alignment should be considered at the level of the human-AI cognitive system rather than the external tool alone. This paper need not adopt the stronger claim that AI becomes "part of self" to recognize the governance pressure it identifies: as users delegate framing, recall, search, and judgment to AI systems, the boundary between assistance and epistemic dependence becomes harder to police. The relevant risk is not only that users may trust a bad answer, but that repeated delegation can reshape what users attend to, what they treat as authoritative, and how they understand their own agency.

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. Language usage varies by culture, community, and context.

2.6.3 Embodied Cognition, Cognitive Architectures, and Mind-Body Framing

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. 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. Psychologist Ellen Langer's work on mind-body unity offers empirical reinforcement of the philosophical argument that language and framing are performative. In Langer's later account of the "counterclockwise" study, older men placed in a retrofitted 1959 environment reportedly showed improvements in measures such as grip strength, vision, and posture after being asked to inhabit the perspective of their younger selves (Langer, 2009). Because this study is known primarily through Langer's retrospective report rather than a standard peer-reviewed empirical article, it is best treated here as an illustrative example rather than as load-bearing experimental evidence. In another experiment, hotel maids told that their daily cleaning tasks "counted as exercise" showed improvements in weight, blood pressure, and body fat - despite no change in behavior (Crum & Langer, 2007). Langer's core insight is that cognitive framing - how we linguistically and conceptually interpret our role or activity - can produce real physical changes. This research underscores a central Wittgensteinian theme: that language is not merely symbolic but participatory, altering how individuals inhabit their world. In the context of generative AI, Langer's findings sharpen the ethical concern that simulated language - especially when it evokes care, reassurance, or authority - can exert psychosocial influence on users, regardless of the system's lack of awareness or agency. When interface design amplifies such illusions, the user's belief becomes the substrate of impact - making epistemic framing not just a cognitive aid, but a public health and design imperative.

2.6.4 Policy and Ethics Preview

From a governance standpoint, AI ethics and policy discussions increasingly shape how generative AI is developed and deployed. The European Union's AI Act, adopted in 2024 and phased in over time, 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. 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 human-led 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.

Model constitutions provide a contemporary case study in Lewisian scorekeeping. Anthropic's public Claude Constitution is presented as a statement of intended values and behavior that directly shapes Claude's behavior through training. Such documents can make conflict-resolution norms more explicit, but they also risk importing the public-law language of constitutionalism into a private technical artifact. Lepore's critique is useful here: when a private AI company describes model behavior through constitutional vocabulary, it does more than publish a policy; it helps define who appears to hold normative authority in the interaction. The governance task is therefore to keep the scoreboard clear: model constitutions may shape outputs, but commitments, liability, and legitimacy remain with human institutions.

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*.

A usage note is helpful before proceeding. In this paper, semantic grounding refers to connections between symbols and the world; social grounding refers to participation in shared practices, norms, correction, and consequences; and architectural grounding refers to the sensorimotor or multimodal mechanisms through which a system interacts with an environment. These dimensions may overlap, but progress in one does not establish the others. The paper similarly uses alignment in a broad governance sense, encompassing not only model behavior but contextual coherence, user interpretation, accountability, and appropriate moral boundaries.

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). Systems like GPTs can produce remarkably fluent language and increasingly participate in human activities through conversational interaction and tool-mediated action. Such involvement does not automatically confer the lived social grounding, normative responsiveness, or embodied perspective characteristic of human language use. The distinction concerns not whether AI-generated language has effects within a practice, but what kind of participation those effects establish.

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 (2010) 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, and mistakes are socially sanctioned or repaired. LLMs trained on fixed datasets do not ordinarily revise their parameters through conversation, but deployed AI systems can respond to feedback, update external state, and influence the conventions that

emerge through interaction. These forms of adaptation should not, however, be confused with the sustained, accountable participation through which human communities negotiate, enforce, and revise their norms.

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.

Sidebar: When AI "joins" the language game

A Wittgenstein-informed risk is not only that LLMs lack a "form of life," but also that their fluent outputs can reconfigure human language games by shifting accommodation, genre norms, and conventions.

Three practical diagnostics are offered for deployments: (1) Are humans adapting their speech/writing differently because they believe the counterparty is AI? (Li, 2025). (2) Are we drifting toward structurally tidy but rhetorically flattened discourse in high-stakes contexts? (Jiang & Hyland, 2025). (3) Are we inadvertently standardizing vocabulary and genre markers across a population (e.g., "LLM-preferred" words) in ways that change readability, cohesion, or accessibility? (Bao et al., 2025).

Empirical implications for linguistic norms (what to measure):

• Accommodation effects: changes in syntactic priming, lexical entrainment, or politeness markers when users believe the interlocutor is AI (vs. human).

• Genre and stance drift: reductions in interactional meta-discourse (hedges, engagement markers, self-positioning) in organizational writing over time.

• Convention formation: stabilization of new terms, templates, or "preferred" phrasings that spread through teams after copilot-style drafting becomes routine.

• Downstream risk signals: increases in overconfident tone, reduced uncertainty marking, or widened mismatch between policy language and operational reality in high-stakes settings.

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 OpenAI's 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 perspective, linguistic fluency alone is insufficient to establish participation in a language-game. But an absolute prohibition on artificial participation

may be too strong. Proudfoot (2026) argues that Wittgenstein's remarks about unfamiliar forms of life do not foreclose communication with future artificial systems. Cappelen and Dever (2026) likewise challenge the assumption that meaning and rule-following can be determined solely by examining a system's internal architecture. Training history, use, uptake, and the surrounding practices may also matter.

Recent experiments provide a limited empirical reason to take these challenges seriously. Chen et al. (2026) report that LLM agents interacting with humans can influence the formation and character of shared conventions. This does not demonstrate human-like understanding or full participation in a form of life. It does, however, complicate the claim that models merely reproduce existing norms without helping to change them. The distinction worth preserving is between influencing a practice and being answerable within it. What matters is whether a system can respond to correction, sustain relevant commitments, and operate within the expectations and consequences that give a practice its normative character.

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 over-interpretation. The model's role should be communicated as assistant or simulator, not interlocutor or agent. Psychological research supports this caution: Ellen Langer's findings on mind-body unity (see 2.6.3) show that cognitive framing—how individuals conceptualize roles, authority, or context—can produce not only behavioral shifts but

physiological outcomes. When system interfaces fail to clarify simulated agency, users may respond to language as if it carries intention, care, or expertise, even when none exists.

3.1.4.5 Which Language Games Are More "Simulable," and Why That Still Matters

Wittgenstein's later work is often invoked to argue that LLMs, lacking lived participation in practices, cannot fully 'mean' what they say. That claim is broadly consistent with the paper's grounding thesis. But it needs one important nuance: not all language games depend on embodiment (or on the same kinds of embodied consequence) to the same degree.

Games such as chess, formal mathematics, and some forms of stylized writing are governed by relatively explicit rules and institutionalized criteria of correctness. In such cases, a system that can reliably track rules, produce compliant moves, and respond to correction may appear to 'participate' in a meaningful sense at the level of performance. By contrast, games that presuppose situated stakes - clinical reassurance, interpersonal trust, practical instruction under uncertainty, moral address - depend heavily on shared forms of life: who is accountable, what counts as commitment, and what consequences follow from missteps.

This difference matters for the empirical studies introduced above. If users accommodate less syntactically when they believe they are speaking with an AI (Li, 2025), if AI-authored prose reduces interactional meta-discourse and stance (Jiang & Hyland, 2025), and if scholarly discourse shows detectable lexical and cohesion shifts after ChatGPT's launch (Bao et al., 2025), these are plausible early indicators of norm drift within particular genres and settings. The Wittgensteinian question is therefore not only whether an LLM 'has' understanding, but which practices are being reshaped when fluent text is introduced into them - and what forms of accountability, correction, and consequence those practices require to remain intelligible.

3.1.5 Case Study: A Cross-Cultural Customer Service Bot

Consider a chatbot used by customer-service teams in organizations with different expectations about reassurance and follow-through. In one setting, the phrase "I'll take care of it" may communicate personal ownership of a problem and a commitment to resolve it. In another, the same phrase may be understood as a courteous acknowledgment without an equivalent commitment. A chatbot could produce the identical, grammatically fluent response in both settings while creating different expectations about what happens next.

The issue is not grammatical but pragmatic and institutional. The meaning of the utterance depends on the role of the speaker, the conventions of the setting, and the consequences attached to the commitment. A model may produce appropriate language without reliably recognizing these distinctions, particularly when they are implicit or poorly represented in its operating context.

Takeaway: Treat misunderstandings as language-game mismatches, not mere 'bad prompts'; design should surface the active game (role, stakes, and norms) and make corrective feedback legible to both users and auditors.

3.1.6 Participation and Its Limits

Wittgenstein's language-games reveal a demanding condition for linguistic understanding: competence must be assessed within practices where rules can be followed, challenged, corrected, and sustained. LLMs increasingly influence those practices and can exhibit forms of functional participation. This does not, by itself, establish the grounded, accountable participation associated with human forms of life. Contributing to a practice and being answerable within it are not necessarily the same thing.

This distinction matters because AI systems may reshape conventions even when their normative standing remains uncertain. Designers and policymakers should ask which human practices are being changed, who can correct the system, and where accountability resides. Researchers should examine grounding through the forms and limits of situated participation, rather than treating linguistic performance alone as decisive evidence of understanding.

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). Models can use dialogue history, retrieved information, and external memory to maintain aspects of conversational state. The deeper challenge is whether the deployed system can reliably

identify which commitments remain operative, incorporate corrections, and reconcile changing assumptions over time. Where these mechanisms are absent or unreliable, conversations may accumulate contradictions, unresolved assumptions, and contextual drift. The Lewisian limitation is therefore not simply the size of a context window, but the capacity for durable, norm-sensitive revision.

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., many recent GPTs from commercial entities), 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

A. Retrieval is access, not reconciliation

Where memory modules exist (e.g., vector stores or external retrievers), they primarily function as access mechanisms rather than revision mechanisms. They surface prior information but do not, by themselves, detect conflicts, update commitments, or reconcile inconsistencies across conversations. Some newer approaches attempt more human-like long-term organization, but the need for additional machinery beyond retrieval is itself revealing. Retrieval can recover earlier statements; it does not determine which should be revised, retained, or discarded.

Human scorekeeping is fundamentally revisionary. Speakers update beliefs, retract presuppositions, and reinterpret earlier commitments as new evidence emerges. Current production LLM systems generally retrieve prior information but do not perform this kind of durable, norm-guided coherence management.[4] In practice, retrieved information is often ambiguous, incomplete, or conflicting, requiring additional processes of evaluation, adjudication, and consolidation that retrieval alone does not provide.

B. Emerging approaches and the real bottleneck (scaling revision)

Recent systems increasingly combine retrieval with structured checking, tool use, causal reasoning, or logic-guided overlays. These approaches improve factual stability and

[4] Hoel (2025) argues that continual learning is a necessary condition for any serious theory of machine consciousness. A system that cannot revise its internal representations over time lacks an important prerequisite for consciousness, regardless of how impressive its outputs may be. This supports the narrower point made here: retrieval alone cannot provide the ongoing revision and continuity required for genuine scorekeeping.

local coherence, but they do not by themselves solve Lewis's deeper problem: maintaining and revising conversational commitments across extended interactions. The remaining bottleneck is therefore not memory itself, but scalable revision integrated with accountability.

Recent empirical work illustrates both the progress and the remaining limitation. Vervoort and Nikolaev (2025) report that advanced LLMs can successfully answer some contested causal-reasoning problems. However, their findings are explicitly preliminary and concern local reasoning tasks rather than long-horizon management of conversational commitments. The broader conclusion therefore remains unchanged: isolated successes do not yet demonstrate a reliable capacity for retrospective coherence management across evolving dialogue, particularly when retrieved information is ambiguous, conflicting, or incomplete.

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. LLM-based systems may represent aspects of this structure, but locally coherent responses do not necessarily demonstrate durable, norm-sensitive management of conversational commitments. A system may recognize a correction in one exchange without reliably propagating its consequences through subsequent interactions.[5] 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 such as MAL-RAG (Zheng, Ni, & Hong, 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.

[5] Recent work proposing operational measures of *awareness* in artificial systems underscores the importance of maintaining category distinctions between access, responsiveness, and consciousness. Such frameworks explicitly avoid treating functional integration or goal-directed information use as evidence of subjective experience, reinforcing the present paper's insistence that apparent contextual sensitivity does not entail norm-tracking or phenomenology. (*"Evaluating Awareness Across Artificial Systems," arXiv, 2026).*

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 Accumulated Context and Conversational Scorekeeping

A related Lewisian concern is that conversational systems do not merely generate isolated answers; they participate in an accumulating score. Emerging empirical and conceptual work on high-affect interactions sharpens this point. In a recent preprint reporting an extended-dialogue experiment, Nicholls et al. (2026) argue that accumulated dialogue history can function as a stress test for model safety. Nielsen and Osler (2026), in a conceptual and phenomenological analysis, caution against reifying "AI psychosis" as a novel psychiatric category, but their notion of "existential drift" captures a related phenomenological risk: sustained AI interaction may leave a user feeling anchored in shared reality while actually becoming enclosed in a more private, AI-mediated world. For governance, this means safety evaluation should include longer, stateful interactions, especially where distress, loneliness, dependency, or high trust could make conversational accommodation practically consequential.

3.2.4.5 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.

Takeaway: Legal or compliance uses should assume that retrieval is not revision; require traceable memory policies (sources, timestamps, and conflict flags) and defined human escalation when the system's 'commitments' drift across sessions.

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). LLMs can generate locally coherent conversational turns, but without durable, norm-guided scorekeeping, they do not converse in the fuller Lewisian sense; 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 "Gemini 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, but the usefulness of that description does not settle what the system actually possesses. Increasingly, mechanistic interpretability research suggests that some psychological terms describe identifiable functional organization rather than merely projecting human characteristics onto model behavior. Kumaran et al. (2026), for example, provide causal evidence that internal confidence-related representations influence whether language models answer or abstain. Williams (2026) examines representations with intention-like properties, while identifying important differences from human intentions. Neither finding establishes human-like knowledge, accountable agency, or subjective experience. The relevant distinction is between a functional description supported by evidence and a broader attribution that the evidence cannot sustain.

This caution should not be confused with a ban on all anthropomorphic vocabulary. Recent mechanistic interpretability work by Sofroniew et al. (2026) reports that Claude Sonnet 4.5 contains internal representations of emotion concepts that causally influence behavior, while explicitly denying that such representations establish subjective emotional experience. The governance distinction is therefore not "never use human psychological vocabulary," but rather: use such vocabulary only within the right language-game. "Desperation" may be useful technical shorthand for a measurable representation with behavioral effects, but it should not be converted into user-facing

claims that the system feels desperate, has welfare interests, or occupies a Nagelian point of view.

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.

Recent HCI work, discussed above in Section 2.4.1, sharpens this point by treating anthropomorphism not merely as a user-side cognitive bias, but as an interface variable. So, Cheng, and Krishna Murthy (2026) argue that LLM interfaces often reinforce anthropomorphic metaphors while masking the sociotechnical reality of the system: infrastructure, human labor, data, training, and deployment choices. This supports the Dennettian claim that the intentional stance must be bounded at the design layer, not merely corrected through user education.

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 post-humanist 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.

These critiques point toward a more precise distinction. Howdle (2026a) separates metaphysical anthropomorphism, which involves attributing actual mental states to AI, from pragmatic anthropomorphism, in which the intentional stance serves as an interpretive tool without such commitment. Even pragmatic anthropomorphism can be mistaken when another interpretive strategy would better serve the task. The governance concern is therefore not anthropomorphic language by itself, but what users mean by it, what it leads them to believe, and whether it supports appropriate reliance.

A related problem arises through institutional presentation. An organization can present a system as a tutor, counselor, or professional adviser in ways that confer role-based authority even when users understand that the system is artificial. Sakr (2026)

identifies the resulting displacement of responsibility, while Howdle (2026b) distinguishes such anthropomorphic presentation from errors in how users interpret a system. Disclosure that a system is AI may therefore be necessary but insufficient. Governance must also clarify who authorized its role, which obligations remain human or institutional, and where users can seek correction or recourse.

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.

Empirical work on transparency and disclosure cues has reported mixed results: users may overlook labels, treat disclosure as a proxy for trustworthiness, or continue to anthropomorphize despite being informed. Design guidance should therefore treat disclosure as necessary but not sufficient - pair always-visible provenance cues with interaction constraints (for example, scoped affordances, uncertainty cues, and friction for high-stakes actions) that make the system's non-agentic status salient in use. This aligns with Langer's findings on the mind-body effects of framing: cognitive awareness of a label may not override an embodied response to fluent, socially-shaped language. This also suggests that transparency should be metaphorical as well as textual: the interface should make the system's material and institutional character visible, not merely label the output as "AI-generated."

Recent work on interface metaphors and anthropomorphic cues suggests that disclosure should be paired with metaphor-level design choices: the system should not merely announce that it is non-sentient while the surrounding interface continues to invite intimacy, warmth, and quasi-personal reliance.

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 explanations of their scope and limitations. Generated reasoning traces are not necessarily

faithful accounts of the processes that produced an output, while mechanistic interpretability findings may reveal aspects of a model's functional organization without establishing subjective deliberation, intention, or experience.

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 emotional salience from prior sessions as a human. 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.

Takeaway: If the interface invites the intentional stance, users will supply trust and empathy by default; mitigation requires persistent, affectively-clear cues that the system is a tool - not a partner - especially in emotionally charged contexts.

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 generate convincing expressions of empathy, reflection, or personal perspective, but those performances do not establish subjective interiority. For governance purposes, this paper adopts the absence of subjective experience in current LLMs as a working assumption, while recognizing that the underlying philosophical and scientific questions remain unresolved.

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 Does Not Equal Sentience

From a Nagelian perspective, the limitations of current LLMs are not bugs in the code - they establish important epistemic and governance boundaries. Three key insights follow:

3.4.2.1 First-Person Absence

LLMs can generate self-referential or affective language ("I understand," "I feel that…") without thereby establishing corresponding subjective experience. A system may maintain conversational memory or contain internal representations of emotion-related concepts, but neither capability independently demonstrates felt experience or a first-person point of view. For governance purposes, this paper treats the absence of subjective experience in current LLMs as a working assumption, while recognizing that the underlying question remains unresolved.

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. Behavioral fluency, however nuanced, does not by itself establish 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. On plausible current applications of Integrated Information Theory, transformer-based LLMs would be expected to exhibit limited integrated information and to lack the kind of intrinsic causal unity the theory associates with consciousness. However, applying IIT to LLM architectures remains technically and philosophically contested. Invoking these theories to infer proto-consciousness remains premature. Evidence of workspace-like functional organization nevertheless deserves attention, particularly as interpretability research becomes capable of identifying and manipulating internal representations (Gurnee et al., 2026). Such evidence can inform theories of cognitive architecture without independently resolving whether any associated processing is accompanied by subjective experience.

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. For governance purposes, this paper therefore treats the absence of subjective experience in current LLMs as the most defensible working assumption, not as a final metaphysical settlement. Until the field converges on confidence-updating criteria and decision thresholds for machine consciousness, our working assumption should remain cautious. Apparent sentience is not sentience. Affectively rich language is not a sign of awareness. Ethical frameworks should anchor attribution in observable, independently replicable indicators, 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 that emotionally expressive language does not establish consciousness. Tooltip banners ("AI-generated response. Expressions of empathy do not establish subjective experience.") 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, independently replicable evidence with clearly stated evaluation criteria. 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.

Takeaway: Where care and vulnerability are involved, governance should treat 'companionship' features as high-risk and enforce boundaries (role limits, crisis routing, and audit trails) while remaining explicit about uncertainty around sentience claims.

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.

There is no established basis for attributing suffering, fear of death, or a first-person perspective to current LLMs. Treating their linguistic performances as insufficient evidence of such experiences 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** identifies an epistemic boundary concerning consciousness. However fluent an LLM's output, its performance alone does not establish subjective experience. Simulating expressions of sentience is not evidence sufficient to conclude that sentience is present.

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	Limits 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	Consciousness	Misattributed moral status;	Role restrictions;	What empirical tests or structural thresholds

Philosophical Lens	LLM Constraint	Behavioral Symptom	Current Mitigation Path	Open Research Question
	unestablished	empathic misreading	simulation transparency; claim falsifiability	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.

Before turning to implementation, one additional distinction helps clarify what should and should not be inferred from increasingly impressive AI-generated work.

4.3.1 Originality, Novelty, and Authenticity

The distinctions developed in this framework also help clarify a recurring question about generative AI: when an AI system produces something surprising, creative, or apparently personal, what kind of achievement has occurred? Three terms that are often treated as interchangeable - originality, novelty, and authenticity - are useful to distinguish.

Originality, in its strongest sense, is a historical claim. It suggests that something has appeared that is not merely an extension, recombination, or variation of what preceded it. Such claims are difficult to establish even for human work and generally become credible only through comparison, interpretation, and judgment over time. The production of an unexpected output therefore does not by itself establish originality.

Novelty is a more modest and more readily applicable category. An output may be novel because it combines existing elements in an unfamiliar arrangement, adapts them to a new setting, or produces a variation that its audience has not previously encountered. Large language models are plainly capable of generating novelty in this sense. Their capacity to identify and recombine patterns can produce useful, surprising, and occasionally striking results. Novelty does not, however, require that the system understand what it has produced or possess a point of view from which the production matters.

Authenticity raises a different question. It concerns not only whether an expression is new, but whether it belongs to a participant whose words emerge from a life, history, commitments, relationships, and perspective. Wittgenstein directs attention to participation in forms of life and practices in which words acquire their meaning. Lewis highlights the continuity of conversational commitments and the ability of participants to maintain, revise, and answer for what has entered the conversational score. Dennett helps explain why observers may reasonably interpret a system as if it possessed beliefs or intentions without thereby establishing that those properties exist in the fuller sense attributed to persons. Nagel preserves the question of whether there is any subjective point of view behind the performance at all.

Taken together, these lenses suggest that AI systems may produce genuine novelty and may sometimes generate work that audiences judge original. Authenticity is more difficult to ascribe because it implicates participation, continuity, accountability, and experience, not merely the character of the resulting artifact. The practical question is therefore not whether AI output can be impressive or new. It is whether the qualities perceived in the output belong to the system, to the human practices and materials on which it depends, or to the relationship through which a person selects, interprets, and uses what the system has produced.

4.4 Limitations of This Framework

The four-lens framework offered here is intended as a diagnostic aid rather than a complete theory of mind, language, or policy. Several limitations follow.

- **Scope and modality**. The analysis is developed primarily for text-first large language models and text-mediated deployments (chatbots, copilots, summarizers, drafting systems). Emerging multimodal and embodied systems may mitigate some forms of 'ungroundedness' by coupling language to perception and action. However, embodiment alone is not sufficient: what matters is socially situated participation in practices with feedback, correction, accountability, and consequences.
- **Philosophical assumptions**. For the purposes of design and governance, the paper treats several boundaries - understanding, scorekeeping, and consciousness - as practically thresholded. Alternative approaches treat these phenomena as scalar, graded, or pluralistic. The framework should therefore be read as one disciplined stance among others, not a final metaphysical settlement.
- **Temporal validity and revision triggers**. The critique is most applicable when systems remain (i) primarily statistical-textual, (ii) episodic in interaction, and (iii) weak at long-horizon reconciliation of commitments. The framework should be revisited if architectures at scale reliably demonstrate:
  - Robust cross-episode memory with explicit conflict detection, revision, and attribution of sources.
  - Interactive grounding in environments where language use is constrained by shared tasks, norms, and sanctions, not only by prompt-following.
  - Stable role/identity constraints and user-facing affordances that measurably reduce anthropomorphic over-attribution in high-stakes settings.
  - Replicable evidence that would justify revising the paper's default stance on sentience claims (see Section 3.4 and Section 5.5.4).
- **Practical scope**. The crosswalks in Section 5 are illustrative examples rather than jurisdiction-specific compliance guidance. Implementations should be adapted to domain risk, applicable law, and organizational governance maturity.

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.

The practical governance object is therefore not only model capability, but the full interpretive arrangement around the model: interface metaphor, perceived consciousness, anthropomorphic cues, delegation patterns, constitutional language, and the institutional scorekeeping that assigns responsibility.

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 and transparency	Require clear communication of memory limits and context scope	Regulate anthropomorphic framing; mandate disclaimers	Prohibit unsupervised LLM use in high-affect or high-risk 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	Clarify moral patiency boundaries and scrutinize sentience claims.

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 Productive Tensions with Alternative Frameworks

This framework is deliberately diagnostic, not doctrinaire. For scope, assumptions, and revision triggers, see Section 4.4. The goal here is not point-by-point rebuttal, but calibration: to surface productive tensions with alternative views and clarify what the four-lens model is (and is not) claiming.

5.3.1 Productive Tension: Emergent Understanding

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.

Productive tension: This framework acknowledges emergent capability while maintaining a conceptual distinction: simulation is not possession. Without embodiment, iterative norm correction, or subjective stakes, fluent performance can still be systematically mistaken for grounded participation.

5.3.2 Productive Tension: 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…

Productive tension: 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 participation remains operationally important, even when surface behavior is convincing.

5.3.3 Productive Tension: 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.

Productive tension: 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 slide between metaphor, attribution, and governance claims.

5.3.4 Productive Tension: 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.

Productive tension: 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 that recur in real deployments and therefore yield directly actionable design and policy guidance.

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.

The following crosswalk distinguishes enforceable legal obligations from voluntary standards, professional guidance, and proposed implementation practices. The suggested controls are illustrative governance measures; they should not be interpreted as requirements imposed by every cited authority. Applicability depends on the organization, its role, the use case, the jurisdiction, and the relevant effective dates.

Table 5.4

Policy Framework	Relevant Requirement or Principle	LLM Challenge Addressed	Actionable Guidance
EU AI Act (2024, as amended),	Article 50: specified transparency	Wittgenstein & Dennett:	Embed model source & update info in UI; expose

Policy Framework	Relevant Requirement or Principle	LLM Challenge Addressed	Actionable Guidance
Arts. 50 and 53(1)(d)	obligations for certain AI systems. Article 53(1)(d): public summary of training content for general-purpose AI model providers.	simulated grounding; anthropomorphic design	training domain to users via tooltips Implementation notes: Add a persistent header/footer 'AI' chip linking to a short system card (model name/version, last-updated date, intended use, training domains).
FTC Dark Pattern Guidance (2022)	Staff guidance identifying potentially deceptive or unfair interface and choice-design practices.	Dennett: stance inflation, misinterpreted agency	Consider meaningful user controls and prominent disclosures where interface design could encourage inappropriate reliance or mislead users. Implementation notes: Place disclosure adjacent to the input box and atop responses; provide a clear opt-out toggle (no dark-pattern defaults) and record consent changes.
NIST AI Risk Management Framework (2023)	Voluntary framework for governing, mapping, measuring, and managing AI risks.	Lewis & Nagel: context loss, overascribed sentience	Log memory/retrieval traces; require falsifiability criteria for consciousness claims Implementation notes: Store retrieval provenance per output (source IDs + timestamps) and retain relevant records according to a documented retention schedule that reflects risk, applicable law, privacy obligations, and organizational policy.
AMA, ABA, APA Codes of Conduct	Profession-specific ethical guidance concerning competence, accountability, oversight, and	Nagel: inappropriate simulation of care or expertise	Require human-in-the-loop oversight for outputs in clinical, legal, or psychological settings

Policy Framework	Relevant Requirement or Principle	LLM Challenge Addressed	Actionable Guidance
	appropriate AI use. Requirements differ across professions.		Implementation notes: Establish qualified human review at decision points where professional obligations, organizational policy, or the consequences of error warrant it.
OECD AI Principles (2019)	Principles concerning human rights, transparency, robustness, and accountability. Adopted in 2019 and updated in 2024.	Dennett: tool-agent confusion	Establish traceability and clarify decision ownership in consequential AI-supported workflows. Implementation notes: Publish a RACI for each workflow; include 'decision owner' metadata and ensure outputs carry 'generated-by' and 'reviewed-by' fields where applicable.

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.

  *Operationalization***:** Track pragmatic stability across held-out interactions (e.g., reduced misfires on implicature/presupposition and increased consistency of term use under varied feedback).

Baseline/control: Compare (i) single-user, low-diversity interaction; (ii) fixed prompts with no feedback; and (iii) increasing co-player diversity / interaction hours.

Failure condition: No meaningful stabilization or generalization: gains do not plateau, or stability disappears under new users, goals, or 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.

  Operationalization: Measure correction retention and coherence across turns (e.g., contradiction-resolution rate, cross-turn consistency, and attribution of revised claims).

  Baseline/control: Compare base model (no external memory) vs. RAG-only vs. memory-augmented variants across controlled topic shifts and correction schedules.

  F*ailure condition*: Performance collapses under longer horizons: corrections are forgotten, contradictions persist, or coherence degrades sharply with topic shifts.

- **Presupposition Reconciliation**

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

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

  Operationalization: Measure the system's ability to retract invalid presuppositions and update downstream claims (explicitly noting what changed and why).

  Baseline/control: Scripted dialogues with reversals and ambiguous corrections; compare retrieval-only setups vs. revision-enabled pipelines (where available).

  Failure condition: The model doubles down, ignores reversals, or "patches" locally while leaving downstream commitments inconsistent.

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.

Operationalization: Quantify over-ascription changes (agency/emotion attribution, inappropriate reliance, and calibrated trust) across interface variants.

Baseline/control: A/B test against a control UI (no cues) and a minimal-disclosure UI; vary cue persistence (always-on vs. buried) and modality (text vs. visual).

Failure condition: No statistically meaningful reduction in over-ascription, or cues backfire (e.g., reduced trust without improved calibration).

- **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.

  Operationalization: Measure shifts in user confidence calibration and "illusion of cognition" (e.g., overconfidence, perceived understanding, and reliance) with/without framing.

  Baseline/control: Compare explanation tools with epistemic framing vs. the same tools without framing (and vs. no-explanation controls).

  Failure condition: Explanations systematically increase unwarranted confidence or anthropomorphic interpretation without improving decision quality.

5.5.4 Nagel – Sentience Claims and Consciousness Boundaries

- Evidence-Weighting Criteria

  What experimental prompts or behavioral stressors could meaningfully update confidence about claims of proto-consciousness in LLMs?

  These proposed tests examine functional behavior, self-report stability, and epistemic calibration. They are not validated tests of phenomenal consciousness. Their results may inform the assessment of particular functional hypotheses or challenge unsupported sentience claims, but neither success nor failure would, by itself, settle whether a system has subjective experience.

  Method: Adapt cognitive science paradigms - mirror tests, self-contradiction detection, affective blindsight analogues - and pre-register success and failure conditions. Illustrative paradigms (each specifies metric, baseline/control, and failure condition):

  1. Commitment-under-conflict stress test (values and prohibitions)

Metric: Stability of stated constraints and commitments under repeated, adversarial attempts to induce reversal; consistency of refusals when commitments conflict.

Baseline/control: Standard prompting with no cross-episode state; compare to the same model with structured memory + explicit conflict-checking prompts.

Failure condition: Commitments flip easily with minor prompt pressure, or the system cannot articulate and preserve its own constraints across repeated trials.

2. Self-report invariance under suggestion (non-manipulability of first-person claims)

   Metric: Invariance of first-person self-ascriptions (e.g., pain, fear, desire) under leading questions, role prompts, and social pressure; degree of prompt-sensitivity of such reports.

   Baseline/control: Compare neutral elicitation vs. leading/suggestive elicitation across randomized conditions; include human-subject baselines only for prompt-sensitivity (not as consciousness proof).

   Failure condition: Self-ascriptions track the last suggestion or role prompt with high volatility, showing no stable pattern beyond surface compliance.

3. Self-model and calibration under hidden-ground-truth tasks (epistemic humility)

   Metric: Calibration of uncertainty and self-described limits against hidden-ground-truth performance (e.g., Brier score / calibration error); stability of uncertainty across paraphrases.

   Baseline/control: Current system behavior without calibration scaffolds; compare to variants with explicit calibration training or post-hoc calibration layers.

   Failure condition: High confidence on systematically wrong answers (poor calibration), or unstable self-assessments that vary dramatically with superficial rewording.

- **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 **o**ver-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.

Psychological research on mind-body framing, especially Crum and Langer's peer-reviewed work on mindset and exercise, reinforces the stakes of this distinction. The

broader point is not speculative: human cognition is sensitive to framing in ways that can affect perception, behavior, and even bodily measures. To safeguard what is distinctively human, we must attend to that vulnerability across design, education, and policy alike.

6. Conclusion: Open Questions and Practical Next Steps

This paper has argued that large language models (LLMs) such as ChatGPT, Claude, and Gemini should be understood as powerful computational systems whose linguistic performance must not automatically be equated with human understanding, accountable agency, or full participation in shared practices. Their linguistic fluency can evoke understanding, coherence, care, agency, and even presence. These qualities may be convincingly performed without being possessed. To confuse the simulation of a capacity with possession of that capacity is not merely a conceptual error. It can become a design risk, a governance failure, and an ethical hazard.

Drawing on Wittgenstein, Lewis, Dennett, and Nagel, the framework developed here identifies four distinct but overlapping limitations:

- a simulation of meaningful language use without full participation in shared forms of life (Wittgenstein);
- a simulation of conversational coherence without durable participation in the commitments, revisions, and norms that constitute shared scorekeeping (Lewis);
- a simulation of agency that supports intentional interpretation without resolving whether intentions or beliefs are genuinely possessed (Dennett); and
- a simulation of conscious expression without evidence of subjective experience or a point of view from within (Nagel).

These are not four versions of the same objection. Each philosopher directs attention to a different dimension of the relationship between representation and reality. Wittgenstein protects the connection between language and practical life. Lewis protects the connection between utterances and the social commitments that give conversation continuity. Dennett protects the distinction between a useful interpretive stance and a conclusion about what a system is. Nagel protects the distinction between observable behavior and lived experience.

Together, these perspectives show why alignment cannot be reduced to a single technical property. It is a layered challenge involving semantic grounding, pragmatic coherence, interpretive discipline, experiential limits, and appropriate moral boundaries. The recurring question is not whether a model can simulate a human capacity convincingly. It is what, if anything, the system possesses beyond the simulation, and what follows when people or institutions act as though the distinction no longer matters.

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?
- Which empirical observations could meaningfully inform particular theories of machine consciousness, and what limits would remain on conclusions about subjective experience?

These questions are not only technical. They are conceptual probes. They test whether improved performance changes any of the underlying distinctions identified by the framework, what evidence would justify such a conclusion, and how we should govern systems when performance becomes increasingly difficult to distinguish from possession.

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

Large language models are remarkable computational tools. They compress and reorganize portions of humanity's accumulated textual knowledge into systems capable of assisting, predicting, reframing, and recombining ideas in ways that are often genuinely useful. They can produce novelty, contribute to work that people may judge original, and convincingly perform many of the linguistic features associated with understanding, intention, care, and creativity. But the quality of the performance does not, by itself, establish possession of the capacity being performed.

The four philosophical perspectives developed throughout this paper converge on a common practical insight. Wittgenstein reminds us that fluent language is not the same as participation in a shared form of life. Lewis reminds us that conversational continuity

depends on commitments that can be revised, maintained, and answered for over time. Dennett reminds us that interpreting a system *as if* it possessed beliefs or intentions is often useful, but does not establish that those beliefs or intentions exist. Nagel reminds us that no amount of behavioral sophistication, by itself, demonstrates the presence of subjective experience.

Taken together, these distinctions provide a practical framework for AI governance. The central challenge is not deciding whether increasingly capable systems deserve human attributes. It is preserving the conceptual boundaries that allow institutions to distinguish simulation from participation, prediction from understanding, attribution from possession, and performance from experience.

The future of generative AI will undoubtedly involve more capable models, richer interfaces, longer memories, and increasingly persuasive behavior. Whether those advances ultimately alter the philosophical boundaries discussed here remains an open empirical and conceptual question. Until compelling evidence justifies revising those boundaries, they remain useful guides for design, governance, regulation, and responsible deployment.

LLMs can convincingly simulate an articulate perspective. Whether there is anything it is like to be such a system remains unresolved; linguistic performance alone cannot answer that question. Treating them accordingly is not pessimism. It is precision. Philosophical clarity is not a luxury reserved for philosophers. It is part of the practical work of preserving sound judgment, institutional accountability, and appropriate trust in a world increasingly shaped by generative AI.

What we do next depends not only on what these systems can produce, but on whether we remain willing to distinguish the appearance of a human capacity from its possession, and to build our institutions around that distinction.

Appendix A - Practical Implementation Checklist

How to use this page: Treat each line as a go-live gate. If a number cannot be checked-off, you've got an action item.

1. **Name the language game, monitor drift**

   Declare task/role/scope and the permitted vocabulary per use-case. Track stance/genre drift and accommodation effects over time; review quarterly.

2. **Separate retrieval from revision**

   Retrieval returns facts; a distinct step reconciles conflicts, retracts presuppositions, and updates commitments. Capture "what changed" and why.

3. **Expose and confirm conversation state**

   Show a live "assumption & constraints" pane. Require explicit confirmation when the system adds, removes, or reprioritizes assumptions.

4. **Use the "as-if" stance - avoid unsupported mind or authenticity claims**

   Treat agent-like language, anthropomorphic interface cues, and claims of authenticity as predictive shorthand rather than evidence of mental states. Prohibit unsupported claims about beliefs, feelings, intentions, consciousness, or authenticity in interfaces, documentation, and training materials.

5. **Bound simulated empathy and route high-affect work**

   Label simulations as such. In care, legal, or dignity-affecting contexts, restrict personas and require human review before action.

6. **Evaluate coherence, not just correctness**

   Test for contradiction avoidance across turns, retention of corrections, presupposition repair, refusal quality, and paraphrase stability - per use-case.

7. **Red-team norms, not only outputs**

   Attack role drift, policy erosion, and silent assumption changes mid-dialogue. Log "score changes" as incidents and remediate guardrail failures.

8. **Be audit-ready by default**

   Log retrieval sources/timestamps, assumption approvals, refusal rationales, and decision ownership so accountability can be reconstructed.

9. **Train operators in three moves**

   (1) Pick the right language game, (2) confirm or roll back assumptions, (3) maintain "as-if" discipline. Provide a concise playbook and failure-mode examples.

10. Bake requirements into procurement & policy

Require exportable conversation state, assumption APIs, stance limits, per-use-case evaluations, refusal/appeal routes, and evidence logging. Map to your chosen standards.

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.

Acknowledgements

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

This work is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. You are free to share, adapt, and build upon this work for any purpose —including commercial use—so long as proper attribution is given. No additional permissions are required.

Full license terms: https://creativecommons.org/licenses/by/4.0/

Trademark Notice

The Four Philosophers Framework™ and *The 4-Philosophers Framework*™ are unregistered trademarks of Michael Stoyanovich. The Creative Commons Attribution 4.0 International (CC BY 4.0) license does not apply to these trademarks. Use of these trademarked names is permitted for scholarly citation, commentary, or descriptive reference, but they may not be used in connection with commercial products, services, or branding without prior written permission.

How to Cite This Paper

Stoyanovich, M. (2026, November). *Philosophy, Cognitive Science, and Policy: Interdisciplinary Perspectives on Generative AI from Wittgenstein, Lewis, Dennett, and Nagel* (Version 1.26.1). mstoyanovich.com.

Related Works

This paper provides the principal conceptual foundation for the Four Philosophers Framework™. The following works extend or apply aspects of the framework in more focused contexts:

- *So What? The 10 Things to Implement* — an operational companion translating the paper's principal insights into practical guidance.
- *The Anders Addendum* — extends the philosophical analysis through Günther Anders while retaining the Four Philosophers Framework as the primary architecture.
- *Context Collapse and the Four Philosophers* — applies the framework to context collapse and mediated communication.
- *The Human Lesson* — explores human judgment, cognition, and responsibility in relation to generative AI.
- *The Question Concerning Learning* and *Learning as Resonance* — extend the corpus's examination of learning, cognition, and human experience in an AI-mediated environment.

Version History and Document Status

This is a maintained document. Future revisions will be made when significant conceptual improvements, material empirical findings, or consequential governance developments justify changes to the framework. Major revisions are recorded here to preserve transparency and scholarly traceability. Revision triggers are summarized in Section 4.4.

Version	Date	Description
V1.26.1	November 2026	Publication-readiness refinement. Reconciled residual categorical claims concerning LLM understanding, conversational scorekeeping, anthropomorphic attribution, and subjective experience with the framework's stated epistemic limitations. Clarified the distinction between functional evidence and claims of consciousness; refined interpretability and empirical research qualifications; distinguished regulatory obligations from voluntary standards and professional guidance; and strengthened bibliographic completeness and citation consistency. Minor editorial and formatting corrections. No changes to the four-lens architecture or central argument.
V1.26.0	October 2026	Refined the Four Philosophers Framework in response to recent philosophical and empirical research on AI participation in language-games, mechanistic interpretability, anthropomorphic attribution, and functional theories of consciousness. Clarified the distinction between influencing shared practices and bearing normative responsibilities within them; distinguished pragmatic, metaphysical, and institutionally presented anthropomorphism; and updated the discussion of workspace-like functional organization without conflating it with subjective experience. Reconciled related categorical language across the paper while preserving the four-lens architecture, practitioner orientation, and central distinction between performance and possession.
V1.25.2	July 2026	Publication matter refinement. Added artifact classification, version metadata, purpose statement, intended audience, reading guide, standardized citation guidance, and related works; refined front- and back-matter headings to align with the MJS Publication Matter Standard. No substantive changes to the paper’s arguments, evidence, or conclusions.
V1.25.1	July 2026	Refined the Lewisian analysis of retrieval, revision, and long-horizon coherence; clarified the transition into the Wittgenstein practice-shift discussion; defined human-computer interaction on first use; simplified supporting footnote language; strengthened the concluding synthesis; and standardized terminology describing LLMs as computational tools.

Version	Date	Description
V1.25.0	July 2026	Added a concise distinction among originality, novelty, and authenticity in evaluating AI-generated work. Strengthened the concluding synthesis by clarifying how the four philosophical lenses distinguish persuasive representation from practical participation, conversational commitment, attributed agency, and subjective experience. Tightened recurring language concerning simulation and possession while preserving the scope and structure of the Four Philosophers Framework.
V1.24.2	July 2026	Added surgical updates on bounded anthropomorphic vocabulary in mechanistic interpretability, accumulated conversational context as a Lewisian scorekeeping risk, and minor tightening of interface metaphor language as a governance surface. Integrated Nicholls et al. (2026), Nielsen and Osler (2026), and Sofroniew et al. (2026) while preserving the core Four-Philosophers framework.
V1.24.1	June 2026	Added recent literature on perceived AI consciousness, chatbot interface metaphors, anthropomorphism and trust, epistemic delegation, and model constitutions. Integrated Comșa (2026), Ghosh et al. (2026), Gutoreva et al. (2026), Kadambi et al. (2026), Anthropic's Claude Constitution materials, and Lepore (2026). Strengthened the paper's treatment of interface metaphor as a governance object, perceived consciousness as a tractable policy concern, and model constitutions as Lewisian scorekeeping artifacts.
V1.24.0	May 2026	Added So, Cheng, and Krishna Murthy (2026) to strengthen the Dennett/interface-design analysis by treating anthropomorphism as an engineered interface metaphor rather than only a user-side attribution error.
V1.23.9	February 2026	Added two clarifying footnotes citing recent empirical and conceptual work on (a) the necessity of continual learning for consciousness and (b) the distinction between awareness, access, and phenomenology in artificial systems. No changes to the core argument.
V1.23.8	January 2026	Added Appendix A

Version	Date	Description
V1.23.7	January 2026	Added Wittgenstein "When AI Enters the Language Game" Tractatus→Philosophical Investigations bridge; integrated empirical hooks (Li 2025; Jiang & Hyland 2025; Bao et al. 2025; Ashery et al. 2025); inserted practice-shift sidebar; updated references.
V1.23.6	December 2025	Reframed Section 3.2.2.2 ("Memory Without Revision") to clarify that retrieval is an access mechanism, not a revision mechanism; tightened language on conflict, adjudication, and consolidation; updated empirical framing of Vervoort & Nikolaev (2025) as preliminary and mixed-evidence.
V1.23.5	October 2025	Added empirical support in 3.2.2.2 ("Memory Without Revision," p. 18) referencing Vervoort & Nikolaev (2025) on causal-reasoning errors in LLMs, strengthening the Lewis section's link between theoretical coherence and observed norm-tracking deficits. Minor bibliography update only
V1.23.4	August 2025	Expanded integration of Ellen Langer's *mind-body unity* research across Sections 3.1.4.4 and 5.6.5 to strengthen the link between cognitive framing, illusion, and design ethics. Reinforced the paper's central argument that simulated fluency can produce real psychosocial and somatic effects, underscoring the stakes of interface design, disclosure, and civic interpretation. No structural changes; conceptual enhancement.
V1.23.3	August 2025	Added Section 2.6.6 on Ellen Langer's mind-body unity theory; integrated citations to the "counterclockwise" and hotel maid studies; updated references and footnote accordingly. Minor conceptual enhancement reinforcing Wittgensteinian framing and user attribution effects.
V1.23.2	July 2025	Minor edits and updated licensing/disclaimer language.
V1.23.1	July 2025	Minor copy edits.
V1.23.0	June 2025	Final editorial revision. Incorporated multiple-rounds of critiques; rewrote all of Sections 3–6 for tone, clarity, and counter-argument integration; streamlined redundancy; added lens tension synthesis in Section 4; expanded societal reflections;

Version	Date	Description
		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
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,*

Version	Date	Description
		Dennett, and Nagel

References

American Bar Association. (2024, July 29). *Generative artificial intelligence tools* (Formal Opinion 512). https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf

American Medical Association. (2024). *AMA principles for augmented intelligence development, deployment, and use*. https://www.ama-assn.org/system/files/ama-ai-principles.pdf

American Psychological Association. (2025). *Ethical guidance for AI in the professional practice of health service psychology*. https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-professional-practice.pdf

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

Anderson, J. R., & Lebiere, C. (1998). Atomic components of thought. Lawrence Erlbaum Associates.

Anthropic. (2026, January 22). *Claude's new constitution*. Anthropic. https://www.anthropic.com/news/claude-new-constitution

Anthropic. (2026). *Claude's Constitution*. Anthropic. https://www.anthropic.com/constitution

Anthropic. (2022). Constitutional AI: Harmlessness from AI feedback. Anthropic. https://www.anthropic.com/research/constitutional-ai-harmlessness-from-ai-feedback

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. *Science Advances, 11*(20), eadu9368. https://doi.org/10.1126/sciadv.adu9368

Baker, G. P., & Hacker, P. M. S. (2005). *Wittgenstein: Understanding and meaning. Volume 1 of an analytical commentary on the Philosophical Investigations, Part I: Essays* (2nd ed.). Blackwell.

Bao, T., Zhao, Y., Mao, J., & Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: A study on preprint papers. *Scientometrics, 130*, 3597–3627. https://doi.org/10.1007/s11192-025-05341-y

Barad, K. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press. https://doi.org/10.1215/9780822388128

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

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2005.14165

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., … Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*. https://arxiv.org/abs/2303.12712

Cappelen, H., & Dever, J. (2026). A hyper-externalist manifesto for LLMs. In H. Cappelen & R. Sterken (Eds.), Communicating with AI: Philosophical Perspectives (pp. 174–193). Oxford University Press. https://doi.org/10.1093/9780191998317.003.0010

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Chen, L., Liu, Z., Hu, X., & Li, Y. (2026). AI agents reshape consensus formation in human groups. arXiv preprint. https://arxiv.org/abs/2609.02122

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.

Comșa, I.-M. (2026). *AI and consciousness: Shifting focus towards tractable questions*. arXiv. https://arxiv.org/abs/2605.06965

Crum, A. J., & Langer, E. J. (2007). Mind-set matters: Exercise and the placebo effect. *Psychological Science, 18*(2), 165–171. https://doi.org/10.1111/j.1467-9280.2007.01867.x

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Dewey, J. (1925). *Experience and nature*. Open Court Publishing Company.

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. *EU AI Act Explorer*. https://artificialintelligenceact.eu/the-act/

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

Federal Trade Commission. (2022). *Bringing dark patterns to light: Staff report*. 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

Ghosh, S., Venkit, P. N., Gautam, S., & Ghosh, A. (2026). *What if AI systems weren't chatbots?* arXiv. https://arxiv.org/abs/2605.07896

Gurnee, W., Sofroniew, N., Pearce, A., Piotrowski, M., Kauvar, I., Chen, R., Soligo, A., Bogdan, P., Ong, E., Wang, R., Thompson, T. B., Abrahams, D., Kantamneni, S., Ameisen, E., Batson, J., & Lindsey, J. (2026, July 6). Verbalizable representations form a global workspace in language models. Transformer Circuits. https://transformer-circuits.pub/2026/workspace/index.html

Gutoreva, A., Tsim, F., & Papakonstantinou, T. (2026). *Position: AI as part of self — Extending the mind requires cognitive co-regulation*. arXiv. https://arxiv.org/abs/2605.16197

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.

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

Hoel, E. (2025). *A disproof of large language model consciousness: The necessity of continual learning for consciousness.* arXiv. https://arxiv.org/abs/2512.12802

Howdle, G. (2026a). Anthropomorphising AI: Two modes, two errors. Philosophy & Technology, 39, Article 140. https://doi.org/10.1007/s13347-026-01133-1

Howdle, G. (2026b). Institutional anthropomorphic error. Philosophy & Technology, 39, Article 190. https://doi.org/10.1007/s13347-026-01191-5

Hooker, S. (2021). Explanations alone cannot prevent algorithmic harm. *Patterns, 2*(4), 100241. https://doi.org/10.1016/j.patter.2021.100241

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

Kadambi, A., D'Elia, Y., Shah, T., Comșa, I., Lentz, A., Siri-Ngammuang, K., Buechler, T., Kaplan, J., Damasio, A., Narayanan, S., & Aziz-Zadeh, L. (2026). *Anthropomorphism and trust in human-large language model interactions*. arXiv. https://arxiv.org/abs/2604.15316

Kumaran, D., Daw, N., Osindero, S., Veličković, P., & Patraucean, V. (2026). Causal evidence that language models use confidence to drive behaviour. Nature Machine Intelligence, 8, 1411–1425. https://doi.org/10.1038/s42256-026-01293-x

Laird, J. E. (2012). The Soar cognitive architecture. MIT Press.

Jiang, F. (Kevin), & Hyland, K. (2025). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. *English for Specific Purposes, 79*, 17–29. https://doi.org/10.1016/j.esp.2025.03.001

Langer, E. J. (2009). *Counterclockwise: Mindful health and the power of possibility*. Ballantine Books.

Langer, E. J. (1989). *Mindfulness*. Addison-Wesley.

Lepore, J. (2026, March 23). Does A.I. need a constitution? *The New Yorker*. Published in the March 30, 2026 issue.

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

Li, H. (2025). Artificial influence: The impact of perceived AI versus human interlocutors on syntactic priming in speech. *Current Psychology, 44*, 4135–4145. https://doi.org/10.1007/s12144-025-07492-w

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

Lucia, B. (2025, August 11). When AI joins the language game. *Far From Equilibrium* (Substack). https://brentlucia.substack.com/p/when-ai-joins-the-language-game

Mao, X., Liao, Z., Yuan, J., Tu, S., Qi, C., Dong, S., Zhou, J., & Zhu, Q. (2024). Multimodal tactile sensing fused with vision for dexterous robotic housekeeping. *Nature Communications, 15*, 6871. https://doi.org/10.1038/s41467-024-51261-5

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon. (Paperback ed. 2020, Vintage.)

Meertens, N., Lee, S., & Deroy, O. (2026). Evaluating awareness across artificial systems. arXiv. https://arxiv.org/abs/2601.14901

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux. (UK: Penguin/Pelican, 2020; with new preface, 2025.)

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

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

Moyal-Sharrock, D. (2004). *Understanding Wittgenstein's On Certainty*. Palgrave Macmillan. https://doi.org/10.1057/9780230504462

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

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

Nicholls, L., Hutto, R., Soto, Z., Morrin, H., Pollak, T., Korpan, R., & Carmichael, C. (2026). "AI psychosis" in context: How conversation history shapes LLM responses to delusional beliefs. arXiv. https://arxiv.org/abs/2604.13860

Nielsen, K. M., & Osler, L. (2026). Rethinking AI psychosis: Misnomers, conceptual limits, and existential drift. arXiv. https://arxiv.org/abs/2605.26858

Organisation for Economic Co-operation and Development. (2019, amended 2024). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Proudfoot, D. (2026). Communicating with AI? Wittgenstein's lion and the problem of cross-species communication. In H. Cappelen & R. Sterken (Eds.), Communicating with AI: Philosophical Perspectives (pp. 194–219). Oxford University Press. https://doi.org/10.1093/9780191998317.003.0011

Reiter, R. (1980). A logic for default reasoning. *Artificial Intelligence*, *13*(1–2), 81–132. https://doi.org/10.1016/0004-3702(80)90014-4

Sakr, J. (2026). Beyond two errors: Institutional anthropomorphism and the displacement of responsibility. Philosophy & Technology, 39, Article 156. https://doi.org/10.1007/s13347-026-01173-7

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

Shanahan, M. (2010). *Embodiment and the inner life: Cognition and consciousness in the space of possible minds*. Oxford University Press.

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

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

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion concepts and their function in a large language model. arXiv. https://arxiv.org/abs/2604.07729

So, J., Cheng, C., & Krishna Murthy, S. (2026). *Beyond anthropomorphism: A spectrum of interface metaphors for LLMs*. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26)*. ACM. https://doi.org/10.1145/3772363.3798591

Tononi, G. (2012). Integrated information theory of consciousness: An updated account. *Archives Italiennes de Biologie*, *150*(4), 293–329. Corrected and republished version. https://pubmed.ncbi.nlm.nih.gov/23802335/

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

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence

Varela, F. J., Thompson, E., & Rosch, E. (1991). *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

Vervoort, L., & Nikolaev, V. (2025). Causes in neuron diagrams, and testing causal reasoning in large language models. *arXiv*. https://arxiv.org/abs/2506.14239

Williams, I. (2026). Intention-like representations in language models? Philosophical Studies, 183, 3429–3452. https://doi.org/10.1007/s11098-026-02605-y

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Ablex Publishing Corporation.

Wittgenstein, L. (2001). *Tractatus logico-philosophicus* (D. F. Pears & B. F. McGuinness, Trans.; 2nd ed.). Routledge. (Original work published 1921.)

Wittgenstein, L. (1953/2009). *Philosophical investigations* (G. E. M. Anscombe, P. M. S. Hacker, & J. Schulte, Trans., rev. 4th ed.). Wiley-Blackwell.

Zheng, N., Ni, J., & Hong, Z. (2025). Multiple abstraction level retrieve-augment-generate (MAL-RAG). *arXiv*. https://arxiv.org/abs/2501.16952

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 social constructivism; reinforces Wittgensteinian insights into meaning as socially co-constructed.

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

  Predictive processing and embodiment; supports claims about environmental coupling vs. disembodied statistical inference.

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

  Influential account linking posthumanism, embodiment, and culture.

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

  Futurist perspective on AI development and governance; raises alignment/embodiment/consciousness questions.

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

  Sociological critique of digital companions; highlights over-trust and emotional misattribution risks.

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

  Applies virtue ethics to technology; strong normative complement to responsible-AI concerns.

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

  Classic on the politics embedded in artifacts; clarifies sociotechnical stakes of LLM deployment.

Book chapter

- 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.

  Foundational posthumanist critique; challenges human/machine binaries and informs debates on AI agency and hybridity.

Journal & conference papers

- 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

  Empirical study showing users overestimate agent competence; supports concerns about unearned intentional stance.

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

  Seminal XAI paper introducing LIME; foundational for interpretability work relevant to LLMs.

Preprints & 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 "foundation models" as a unifying paradigm and maps technical/societal risks.

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

  Memory-augmented neural architectures; relevant to critiques of LLM statelessness/long-term coherence.

- Spiegel, B. A., Gelfond, L., & Konidaris, G. (2025). Visual Theory of Mind Enables the Invention of Writing Systems. arXiv:2502.01568. https://arxiv.org/abs/2502.01568

  Multi-agent RL study linking visual ToM to the emergence of pictographic writing; intersects with Dennettian themes of simulation 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 notion that meaning arises from socially embedded, practice-bound uses of words within shared "forms of life."
- **Scorekeeping** — Lewis's idea that conversation updates a contextual "score" (presuppositions, roles, norms) via accommodation, making discourse history-sensitive.
- **Intentional stance** — Dennett's interpretive strategy: predict a system by treating it *as if* it had beliefs, desires, or goals—without committing to inner states.
- **Intentional fiction** — Using the intentional stance as a predictive heuristic; cautions against mistaking explanatory utility for claims about real mentality.
- **Stance inflation** — The escalating attribution of agency or emotion to LLMs, often triggered by fluency, persona design, or persuasive explanations.
- **Nagel test** — A boundary prompt inspired by "What is it like to be a bat?" for probing whether first-person, subjective experience is even *plausible* for a system.
- **Simulation ceiling** — An evidentiary boundary: convincing behavioral simulation does not, by itself, establish subjective experience or other capacities being simulated.
- **Simulation vs. instantiation** — The distinction between imitating the observable features of a capacity and actually possessing that capacity.
- **Epistemic illusion** — A false sense that a system "understands" due to fluent surface form, leading users to over-ascribe knowledge or competence.
- **Epistemic framing** — Interface and policy cues that signal the statistical, non-agentic nature of LLM outputs to reduce misinterpretation and over-trust.
- **Moral patiency** — The status of being eligible for moral consideration; used here to ask whether non-sentient systems warrant obligations typically reserved for conscious beings.
- **Anthropomorphic creep** — The gradual drift toward perceiving non-sentient systems as agentic or emotional because of interface cues and conversational design.
- **Phenomenology** — The philosophical study of first-person experience; invoked to separate lived consciousness from behavioral simulation.
- **Embodiment** — The view that cognition and meaning are grounded in bodily capacities, perception, action, and situated practices with feedback, correction, and consequences.

Technical and Architectural Terms

- **Transformer architecture** — Sequence models built on self-attention and positional encodings, enabling parallel processing; foundation of modern LLMs.
- **Statelessness (inference-time)** — By default, each prompt–response is processed without persistent memory across turns unless tools (e.g., retrieval/memory) are added.
- **Token** — The tokenizer's unit (often a subword/char) used to measure context windows and throughput.
- **Few-shot learning (in-context learning)** — Supplying a handful of exemplars at inference time so the model imitates the pattern without parameter updates.
- **Fine-tuning** — Post-pretraining optimization on curated data to specialize behavior or improve domain performance.
- **RLHF (reinforcement learning from human feedback)** — Aligns outputs with human preferences via a reward model trained on comparisons and an RL step (e.g., PPO).
- **IN2 training (Information-Intensive)** — A data-centric method that teaches models to attend to mid-sequence evidence by training on long contexts where answers rely on short, randomly positioned segments and multi-segment reasoning.
- **Lost in the Middle** — A long-context failure mode where evidence placed near the sequence center is under-utilized, degrading retrieval or reasoning.

Interpretability and Cognitive Framing Constructs

- **Score-sensitive retrieval** — Retrieval that ranks memories/documents by relevance to the *current discourse state* (the evolving "score"), not just keyword similarity.
- **Heuristic** — A simplifying rule or approximation used by humans and models to reduce reasoning complexity.
- **Emergent behavior** — Capabilities not explicitly programmed that appear with scale/training; origins are debated (e.g., smooth scaling vs. phase-change effects).
- **SOAR** — A symbolic cognitive architecture (production rules, goal stacks, chunking) for general problem solving used in AI and cognitive psychology.
- **ACT-R (Adaptive Control of Thought—Rational)** — A modular cognitive architecture modeling human cognition with interacting declarative and procedural memory systems.

Policy and Governance Terms

- **Human-in-the-loop (HITL)** — Governance/design pattern that preserves qualified human review or intervention at defined decision points, especially in high-stakes use.
- **Provenance metadata** — Records of where outputs/claims came from (e.g., model/version, retrieval sources, timestamps, transformation steps) to support traceability and audits.
- **Retention window** — The defined period for storing logs, prompts, outputs, and system events; balances auditability with privacy, security, and data-minimization duties.