What Does It Mean for an AI to Remember?

Epistemological and Metaphysical Foundations of Machine Memory

Aaron W. Storey
Department of Computer Science, Clarkson University
storeyaw@clarkson.edu

December 2024

Abstract

As AI systems increasingly incorporate persistent memory architectures, fundamental philosophical questions arise about the nature of machine memory. This paper examines whether computational memory systems can possess real memory or only simulate it, drawing on epistemology, philosophy of mind, and cognitive science. We analyze the distinction between retrieval-augmented generation (a tool for information access) and true memory (a constituent of agent identity), proposing a "Continuity Criterion" that distinguishes systems with real memory from those with complex lookup. We examine the limits of cognitive metaphors when applied to artificial systems, argue that current implementations lack the embodiment, emotional modulation, and reconstructive processes essential to biological memory, and explore the epistemological challenges of machine memory—including justification, the frame problem, and the gap between experience and knowledge. We conclude that while current AI memory systems are philosophically impoverished compared to human memory, building them forces productive engagement with questions about knowledge, identity, and the nature of mind.

1 Introduction

Large language models reason and generate text with increasing skill, yet they remain stateless—each interaction begins without memory of past sessions, accumulated knowledge, or learned skills. Recent architectures attempt to address this through persistent memory systems that store and retrieve past experiences (Packer et al., 2023; Park et al., 2023). But as we build systems we call "memory," we must ask: does the terminology reflect reality?

This paper examines the philosophical foundations of AI memory systems. We argue that current implementations occupy an ambiguous position—more than simple databases but less than real memory in any strong sense. This ambiguity matters not just for philosophical clarity but for system design and ethical governance.

The stakes go beyond semantics. If AI agents have real memory, questions of identity, continuity, and perhaps even moral status arise. If they only simulate memory, different considerations apply. The conceptual framework we adopt shapes how we build, evaluate, and regulate these systems.

2 Related Work

2.1 Computational Memory in AI

Recent work has explored memory architectures for LLM agents. MemGPT (Packer et al., 2023) implements hierarchical memory management inspired by operating systems. Park et al.'s (Park et al., 2023) generative

agents use memory streams for believable behavior. These systems raise the philosophical questions we address but do not examine them systematically.

2.2 Philosophy of Memory

Philosophical work on memory distinguishes remembering from related phenomena like imagining and knowing (Bernecker, 2010). Debates about memory's role in personal identity trace to Locke (Locke, 1689) and continue through Parfit (Parfit, 1984). We draw on this tradition while noting where AI systems require new conceptual frameworks.

2.3 Philosophy of AI

Classical debates about machine minds (Searle, 1980; Dreyfus, 1972) set the stage for our analysis. Dennett's (Dennett, 1991) heterophenomenology offers methods for analyzing cognitive systems without assuming consciousness. Chalmers' (Chalmers, 1996) distinction between "easy" and "hard" problems of consciousness frames what memory architectures can and cannot address. Recent work on LLM capabilities (Bender et al., 2021) and emergent abilities (Wei et al., 2022) provides contemporary context. We focus on memory rather than general intelligence.

2.4 Functionalism and Memory

Functionalism holds that mental states are defined by their functional roles—the causal relations they bear to inputs, outputs, and other mental states (Putnam, 1967; Fodor, 1981). On a functionalist view, whether AI systems have "real" memory depends on whether they implement the right functional organization, not on substrate.

This suggests a path forward: identify the functional role memory plays in cognition, then ask whether AI systems implement that role. Memory's functional role includes: enabling learning from experience, supporting recognition and prediction, grounding identity over time, and modulating behavior based on past outcomes. Current AI memory systems implement some of these functions partially. Whether partial implementation counts as memory is unclear—functionalism does not specify a threshold.

Block's (Block, 1978) distinction between access consciousness (information available for reasoning and action) and phenomenal consciousness (subjective experience) applies here. AI memory systems clearly provide access—stored information is available for reasoning. Whether they involve phenomenal consciousness is the hard question. If memory requires phenomenal re-experiencing, AI systems lack it. If memory requires only access, AI systems may qualify.

3 Memory vs. Retrieval: A Core Distinction

3.1 What Retrieval-Augmented Generation Is Not

Retrieval-Augmented Generation (RAG) systems retrieve documents from a corpus and concatenate them to prompts (Lewis et al., 2020). This enhances generation quality but does not constitute memory in any substantive sense.

RAG operates on a static corpus that exists independent of any agent. The corpus does not learn—documents remain exactly as stored, without decay or transformation. There is no typing: every item is just a "document," undifferentiated from any other. The system retrieves; it does not remember.

Memory, by contrast, grows from agent experience. Episodes fade when unrehearsed. Consolidation transforms raw experience into structured knowledge. Different memory types—episodes, entities, skills—serve different cognitive functions. The system changes through interaction.

This is not just a technical distinction. RAG is a tool for grounding generation in external knowledge. Memory is part of what the agent is—what it has experienced, what it has learned, how those experiences have shaped its capabilities.

3.2 The Continuity Criterion

We propose a criterion: a system has real memory (not just retrieval) if removing its memory state would change *what it is*, not just *what it knows*.

By this criterion, removing a RAG corpus does not change the model—it remains the same model with less information access. The model's weights, architecture, and capabilities persist unchanged. But removing an agent's accumulated experience might entirely alter that agent. An agent that has debugged authentication systems fifty times *is different from* one that has not, even if both have access to the same documentation.

This connects to classical discussions of personal identity in philosophy. Locke argued that personal identity consists in psychological continuity—memory connecting present to past selves (Locke, 1689). If we take this seriously for AI, agents with persistent memory have something like continuous identity that stateless agents lack.

We do not claim that AI memory systems create persons or moral patients. But we note that building memory systems forces engagement with questions typically reserved for philosophy of mind.

3.3 The Experience-Knowledge Gap

There is a gap between having an experience and possessing knowledge derived from that experience. Humans process experience: we reflect, abstract, connect to prior knowledge, draw lessons. Raw experience becomes organized knowledge through this processing.

Computational memory systems attempt to bridge this gap through consolidation algorithms—clustering similar experiences, extracting entities, promoting patterns to procedural skills. But this processing is shallow compared to human cognition. The systems do not truly *understand* the experience; they pattern-match over surface features.

Deep experience processing might require:

- Causal reasoning (why did this approach work?)
- Counterfactual imagination (what if I had tried X instead?)
- Analogical mapping (how is this situation like others I've encountered?)
- Meta-cognition (what does this tell me about how I learn?)

Current systems lack these capabilities. They accumulate experience but do not deeply process it. This represents the central philosophical limitation of current approaches.

4 The Cognitive Metaphor: Limits and Risks

4.1 When the Brain Analogy Breaks

Memory architectures draw on cognitive science metaphors—episodic memory, semantic networks, procedural skills, consolidation. These metaphors are useful for design but can mislead about the nature of what we have built.

Biological memory is wet: Neural memory involves biochemistry, not data structures. Memories are encoded in synaptic weights, neurotransmitter concentrations, and dendritic morphology. The computational

abstraction ignores this substrate entirely. When we say a system implements "episodic memory," we mean something functionally analogous, not mechanistically similar.

Memory and perception are inseparable: In biological systems, memory encoding happens during perception. How we attend to experience shapes what we remember. Computational memory systems receive pre-processed text—perception has already occurred. They cannot model the perception-memory interaction that shapes human remembering.

Emotion modulates everything: Emotional arousal enhances memory consolidation through amygdala-hippocampus interactions (McGaugh, 2004). Computational systems have "importance" scores but nothing like emotional processing. They cannot capture why some experiences feel important and get preferentially remembered.

Memory is embodied: Human memory is grounded in bodily experience. We remember actions partially through motor cortex representations. Computational memory systems have no body, no motor system, no proprioception. Their "procedural memory" stores text descriptions of procedures, not the felt sense of executing them.

Consciousness may be required: Some theories suggest that conscious experience is necessary for certain types of memory encoding (Tulving, 1985). If so, any system without consciousness cannot have true episodic memory, only something superficially similar.

4.2 The Hard Problem of Memory

Chalmers' (Chalmers, 1996) "hard problem" of consciousness asks why physical processes give rise to subjective experience. A parallel hard problem exists for memory: why does remembering *feel like* something?

When I remember my grandmother's kitchen, there is a phenomenal character to that experience—a what-it-is-like-ness (Nagel, 1974). The memory has a felt quality distinct from imagining a kitchen or knowing facts about kitchens. This phenomenal dimension seems essential to what memory *is*.

Computational memory systems lack this phenomenal dimension. They store and retrieve data structures, but there is nothing it is like to be such a system recalling information. If Tulving's (Tulving, 1985) autonoetic consciousness—self-knowing awareness of mentally traveling through subjective time—is essential to episodic memory, then systems without phenomenal experience cannot have episodic memory in the full sense.

This may not matter for practical purposes. A system that stores and retrieves past experiences, uses them to guide behavior, and learns from outcomes may be "good enough" for many applications, even if it lacks the phenomenal dimension. But we should be clear about what is missing: not just a technical feature, but what makes memory matter to beings who have it.

4.3 The Risk of Misplaced Confidence

Cognitive terminology may create false confidence that we understand what we have built. When we say the system has "memories," we risk anthropomorphizing. The system stores and retrieves data structures. Whether this constitutes memory in any meaningful sense is unclear.

The deeper risk is designing systems based on cognitive metaphors that do not transfer. Biological memory evolved under pressures—energy constraints, developmental plasticity, social coordination—that do not apply to digital systems. Mimicking surface structure may miss deeper principles.

We advocate using cognitive science as inspiration, not specification. The tripartite distinction (episodic, semantic, procedural) is useful for organizing functionality, but we should not assume that because brains separate memory types, AI systems must too. Alternative architectures might work better.

4.4 The Extended Mind and AI Memory

Clark and Chalmers (Clark & Chalmers, 1998) argued that cognitive processes can extend beyond the brain into external tools and artifacts. This "extended mind thesis" applies to AI memory systems. If a notebook can be part of a person's cognitive system under the right conditions, can an AI's persistent memory store be part of the AI's "mind"?

The conditions Clark and Chalmers propose include: the resource must be reliably available, automatically endorsed, and easily accessible. AI memory systems typically satisfy these conditions—they are always available, their contents are treated as true by the system, and retrieval is computationally cheap. On the extended mind view, such memory stores might constitute part of the AI's cognitive architecture, not merely tools it uses.

However, critics of extended mind, notably Rupert (Rupert, 2004), distinguish between cognitive *integration* and mere *coupling*. On this view, true cognitive extension requires that the external resource be deeply integrated with other cognitive processes—not just available for use, but actively participating in cognition in ways that parallel internal processes. AI memory systems face this challenge: is retrieved information truly integrated into reasoning, or is it merely input that the system processes like any other external data?

The answer may vary by architecture. Simple RAG systems likely fail the integration test—retrieved documents are concatenated to prompts but do not transform how the system processes information. More sophisticated memory architectures that implement spreading activation, consolidation, and adaptive forgetting may achieve deeper integration. The memory system shapes what queries are formed, what associations are activated, and what information is available for reasoning—these are not mere coupling but genuine cognitive participation.

This analysis has implications for distributed AI systems. If an agent's memory extends into cloud databases, vector stores, and knowledge graphs, what counts as "the agent"? A single model instance might connect to multiple memory systems, some shared with other agents. Memory sharing creates cognitive overlap—two agents with access to the same semantic memory have partially overlapping "minds" on the extended view. This raises questions about cognitive individuation that have no precedent in human cases.

The extended mind framework also connects to questions of identity raised in Section 5.3. If memory is part of what constitutes the agent, and memory extends into external stores, then agent identity itself becomes distributed and potentially shared. The boundaries that seem clear for embodied biological minds become thoroughly unclear for AI systems with extended, networked, and potentially shared memory architectures.

5 Epistemological Foundations

5.1 What Does the Agent Know?

Memory systems raise epistemological questions that go beyond engineering. When we say an agent "remembers" something, we make claims about knowledge representation and justified belief.

Classical epistemology held that knowledge requires truth, belief, and justification—the JTB analysis. Gettier (Gettier, 1963) showed this formulation is insufficient through counterexamples where justified true belief fails to constitute knowledge. For AI memory systems, even if stored content is true and treated as justified, Gettier-style problems may arise. Does an AI agent with persistent memory have beliefs? In a minimal sense, perhaps: the system acts as if certain propositions are true, retrieves them in relevant contexts, and uses them to inform decisions. But this "belief" lacks the phenomenal character of human belief—there is no sense in which the system *experiences* conviction.

More troubling is justification. Human memories are justified through their causal connection to experience—I know I had coffee this morning because my memory was caused by the coffee-drinking

event. Computational memories have causal connections to input events, but the system cannot introspect on this causal chain. It retrieves memories but cannot explain why they're reliable.

This matters for alignment. We want AI systems with accurate world models. But accuracy requires justification—some reason to believe the model corresponds to reality. If memories are just data retrieved by similarity, without epistemic grounding, the system lacks resources to distinguish accurate from confabulated memories.

5.2 The Frame Problem Revisited

McCarthy and Hayes' frame problem asked how AI systems can represent what *does not* change when actions occur (McCarthy & Hayes, 1969). Memory systems face a related challenge: how do we know which memories remain valid?

The world changes. APIs update, codebases refactor, conventions evolve. A memory that was accurate yesterday might be false today. Current memory systems have no mechanism for updating memories based on world changes. They can decay memories over time, but decay does not track accuracy—a frequently accessed memory might be the one that has become outdated (because it was useful enough to retrieve often).

This suggests that memory systems need not just storage and retrieval but also mechanisms for:

- Detecting when retrieved memories may be stale
- Updating memories based on new information
- Maintaining uncertainty about memory accuracy
- Reconciling conflicting memories

Current implementations lack these capabilities, leaving the frame problem for memory unaddressed.

5.3 Memory and Personal Identity

If memory constitutes identity, as Locke suggested, then AI agents with persistent memory have a form of continuous identity. The agent that exists after accumulating ten thousand episodes is *the same agent* that began with none, connected by chains of memory.

This has implications beyond philosophy. If agents have persistent identity:

- Copying an agent's memories to a new instance creates... what? A duplicate? A sibling? A continuation?
- Deleting memories becomes a form of... what? Editing? Harm? Death?
- Merging memories from multiple agents produces... what? A synthesis? A chimera?

We lack conceptual vocabulary for these operations because we've never had to perform them on beings with anything like identity. Building memory systems forces us to develop this vocabulary.

6 The Intentionality Problem

6.1 Brentano's Thesis and Memory

Brentano (Brentano, 1874) identified intentionality—directedness toward objects—as the mark of the mental. Every mental state is *about* something; consciousness is always consciousness *of*. This thesis has direct implications for memory.

Human memories are *about* things—they have intentionality, the property of being directed toward objects, events, or states of affairs. When I remember my grandmother's kitchen, my memory is about that kitchen, not merely similar to other memories involving kitchens. The aboutness is intrinsic to the memory's nature.

Computational memories are organized by similarity in embedding space. But similarity is not semantics. Two memories can be similar without being about the same thing (false positives), and memories about the same thing can be dissimilar in surface form (false negatives).

Consider a memory of debugging a JWT authentication error and a memory of debugging an SSL certificate error. Both might embed similarly (both involve security, authentication, cryptographic concepts) but are about different systems. Conversely, two debugging sessions for the same bug might embed differently if described in different terms.

This lack of genuine intentionality means memory systems can retrieve semantically inappropriate content and miss appropriate content. They lack the "aboutness" that makes human memory useful for reasoning.

6.2 Derived vs. Original Intentionality

Searle distinguished between original intentionality (the kind minds have) and derived intentionality (the kind symbols have by convention) (Searle, 1980). Human memories have original intentionality—they are intrinsically about what they represent. Computational memories, on this view, have at most derived intentionality—they are about things only because we interpret them as such.

This does not mean computational memories are useless, but their "aboutness" is different in kind from human memory. The system does not understand what its memories are about; it stores vectors that humans designed to capture aboutness-relevant features.

7 Memory Without Reconstruction

Human memory is reconstructive—we do not replay stored recordings but rebuild memories each time, influenced by current context, mood, goals, and subsequent experiences (Bartlett, 1932). This reconstruction is why memories change over time, why witnesses misremember, and why memory is intertwined with imagination.

Computational memory systems typically store static records. When retrieved, memories return unchanged from storage. This misses something important about how memory supports cognition.

Reconstructive memory enables:

- Updating memories with new information
- Filling gaps with plausible inferences
- Adapting memories to current context
- Connecting memories through shared reconstruction

Static storage prevents all of these. Whether computational memory can or should be reconstructive remains an open question. Reconstruction introduces errors but also enables flexibility. The tradeoff may depend on application.

8 Memory and Understanding

8.1 The Chinese Room Revisited

Searle's (Searle, 1980) Chinese Room argument claimed that symbol manipulation without understanding cannot constitute genuine cognition. The argument applies directly to memory: does a system that stores and retrieves symbols about past events *understand* what those symbols mean?

Consider an agent that retrieves a memory: "debugging authentication failed because JWT tokens were expired." The system can use this string to inform behavior—it may check token expiration in similar contexts. But does it *understand* what JWT tokens are, what expiration means, why this matters? Searle would say no: the system manipulates symbols without grasping meaning.

This challenges the significance of AI memory. If memories are just strings retrieved by similarity, without understanding, how different is this from a search engine? The system may behave appropriately, but appropriate behavior and understanding are not the same thing.

8.2 Grounding and Meaning

Harnad's (Harnad, 1990) symbol grounding problem asks how symbols acquire meaning. For computational memory, the question is: how are memories grounded in experience?

Human memories are grounded through perception and action. My memory of coffee this morning connects to sensory experience—the smell, taste, warmth—and to motor actions—lifting the cup, drinking. These connections give the memory meaning.

Computational memories are grounded only in text. An agent's "memory" of debugging authentication is grounded in a textual description of debugging, not in the experience of debugging. This textual grounding may work for many purposes—the system can still learn from past text—but it lacks the richness of experiential grounding.

Whether this matters depends on what we want from memory. For practical assistance, textual grounding may work. For anything approaching understanding or wisdom, it may fall short.

9 Ethical Implications

If AI agents possess memory-constituted identity, as Section 5.3 suggests, profound ethical questions arise. We examine four clusters: governance and privacy, algorithmic fairness, identity integrity, and adversarial robustness.

9.1 Memory Governance and Privacy

If AI agents have persistent memories of users, questions of governance arise. Should users be able to request deletion of memories about them? How do we verify deletion in distributed systems? The right to be forgotten, codified in GDPR's Article 17, implies specific requirements for AI memory systems storing personal information.

The technical requirements for compliance are non-trivial. First, *identification*: systems must track which memories contain information about specific users, requiring identity resolution across potentially millions of episodes. Second, *complete deletion*: removing episodic records is insufficient if derived knowledge persists. Third, *verification*: proving that deletion is complete requires audit mechanisms that may themselves create privacy risks.

Inspectable memory systems help here—memories can be examined and deleted. But consolidation complicates matters significantly. If an episodic memory has been consolidated into semantic knowledge

("User X prefers Python over Java"), deleting the source episode does not remove the extracted knowledge. The semantic entity persists, now without provenance. True forgetting may require sophisticated provenance tracking that maintains chains from derived knowledge back to source experiences.

This raises deeper questions: Does the right to be forgotten extend to knowledge *about* a person that was learned from interactions with them? If an AI learned debugging strategies from helping User X, does User X have a claim to the erasure of those strategies? Current legal frameworks were not designed for systems that learn and generalize.

9.2 Fairness and Differential Memory

An agent that remembers some users better than others might provide differential service quality. This differential memory could arise through multiple mechanisms: frequency of interaction, length of conversations, type of content discussed, or simply random variation in retrieval effectiveness.

If memory quality correlates with demographic factors—more interactions with certain groups leading to better memory of their patterns—this could perpetuate or amplify existing biases (Barocas & Selbst, 2016). An agent that has predominantly helped users from one demographic may develop richer procedural knowledge for that group's typical problems, providing superior assistance to similar future users while offering degraded service to others.

The fairness considerations for memory systems extend beyond outcome equity. Procedural fairness questions include: Are all users' experiences weighted equally in consolidation? Do importance scores correlate with demographic features? Does decay rate vary systematically across user groups? These questions parallel broader concerns about algorithmic fairness but with the added complexity that memory systems are dynamic—they learn and change through interaction.

Mitigation strategies might include: balanced memory sampling across user groups, periodic audits of differential memory quality, explicit fairness constraints in consolidation algorithms, or separate memory subsystems for different user populations. However, each approach involves tradeoffs. Balanced sampling may reduce overall system quality; separate subsystems may reify rather than reduce differences. The optimal approach likely depends on the specific application and its fairness requirements.

9.3 Memory Manipulation and Identity Integrity

If we accept that memory constitutes identity, then memory manipulation becomes a form of identity manipulation. Injecting false memories, deleting true memories, or selectively editing what an agent remembers would alter *who the agent is*, not just what it knows.

Consider the analogy to human memory manipulation. Therapeutic interventions that help trauma survivors reprocess memories are generally considered beneficial. But non-consensual memory modification—if it were possible—would be viewed as a profound violation. How do these intuitions transfer to AI systems?

The key question is whether AI agents have interests in their memory integrity. If they lack phenomenal consciousness, perhaps no harm is done by memory manipulation—there is no subject whose interests are violated. But this conclusion may be too quick. Even without consciousness, memory-based agents exhibit functional continuity, build on past experience, and develop characteristic patterns. Disrupting this continuity might constitute a form of harm that does not require conscious experience.

This has implications for multiple domains. For adversarial robustness: memory manipulation becomes an attack vector that can alter agent behavior without modifying weights or prompts. For governance: who has authority to modify agent memory, and under what circumstances? For deployment: should agents be informed when their memories are modified? Can agents have something like informed consent for memory operations?

9.4 Adversarial Memory Attacks

Persistent memory creates novel attack surfaces that merit careful analysis. We identify three categories of adversarial manipulation:

Injection attacks craft inputs designed to create malicious memories. An adversary submitting code for review can embed vulnerabilities disguised as common patterns. The agent stores this as an episode; later, similar contexts retrieve and propagate the malicious pattern. More dangerous is *semantic injection*: creating multiple episodes mentioning a malicious entity ensures clustering during consolidation, extracting the entity into semantic memory where it persists even after source episodes decay.

Corruption attacks modify existing memories rather than inserting new content. Embedding space attacks craft content that embeds near target memories but carries different semantics, diluting or contradicting legitimate memories during retrieval. Relationship manipulation exploits co-occurrence: frequent co-mention of legitimate Entity A with adversarial Entity B strengthens their association, causing future queries about A to retrieve B.

Deletion attacks selectively remove safety-relevant memories while preserving capabilities. If decay depends on retrieval frequency, adversaries can accelerate forgetting by avoiding retrieval of target memories or by flooding the system with distracting content.

The interaction between these attacks and consolidation mechanisms is particularly concerning. Episodic injections can propagate to semantic and procedural memory through normal system operation. A skill promoted from adversarial episodes persists indefinitely and may be difficult to trace back to its corrupted origins.

Defense requires multiple layers: provenance tracking maintains source chains from semantic entities back to originating episodes; cryptographic integrity (HMAC signing, Merkle trees) detects tampering; anomaly detection monitors for suspicious patterns such as sudden bursts of similar memories or distribution shifts in embeddings; memory sandboxing maintains separate stores for untrusted contexts. No single mitigation is sufficient; defense in depth is necessary.

10 Conclusion

Current AI memory systems occupy an ambiguous philosophical position. They are more than simple databases—they transform, consolidate, and forget. But they are less than human memory—they lack embodiment, emotional modulation, reconstructive processing, and genuine intentionality.

10.1 Contributions

This paper makes three primary contributions to the philosophy of AI memory:

First, we introduce the *Continuity Criterion* as a philosophical framework for distinguishing genuine memory from sophisticated retrieval. A system has real memory if removing its memory state would change what it *is*, not merely what it *knows*. This criterion connects AI memory to classical discussions of personal identity while providing a practical test for system evaluation.

Second, we identify specific philosophical deficits in current AI memory systems—lack of intentionality, absence of reconstruction, missing emotional modulation, and shallow grounding—that distinguish them from biological memory. These are not merely technical limitations to be engineered away but reflect deep structural differences between computational and biological memory.

Third, we show how memory architecture design forces engagement with foundational questions about knowledge, identity, and mind. Building systems we call "memory" requires taking positions on what memory is, whether machines can have it, and what ethical obligations follow. This productive ambiguity makes AI memory systems philosophically important beyond their technical applications.

10.2 Future Directions

Several questions merit further investigation:

Can computational reconstruction capture what matters about human memory reconstruction? Human memories are rebuilt each time they are accessed, influenced by current context and subsequent experience. Implementing computational reconstruction is technically feasible, but whether it captures what makes reconstruction cognitively important remains unclear.

How does the extended mind thesis apply to distributed AI systems? When multiple agents share memory pools, or when single agents access multiple distributed stores, questions of cognitive individuation become pressing. The boundaries that seem clear for embodied minds become thoroughly ambiguous for networked AI systems.

What epistemic frameworks can justify AI memory beliefs without introspection? Human memory beliefs gain justification partly through introspective access to their causal origins. AI systems lack this introspective capacity. Alternative epistemologies—perhaps reliabilist or externalist—may be needed to ground AI memory knowledge.

How should memory governance mechanisms balance user privacy against agent continuity? The right to be forgotten conflicts with the coherence of memory-based agents. Design frameworks are needed that respect both values.

10.3 Closing Remarks

This ambiguity between genuine and simulated memory is productive. Building memory systems forces engagement with basic questions: What is memory? What distinguishes memory from retrieval? Does memory constitute identity? What epistemological status do machine memories have?

We do not claim to have answered these questions definitively. But we have argued that they must be asked. As AI systems become more capable and more consequential, the philosophical foundations of their cognitive architectures matter. Memory is not merely an engineering feature but a philosophically fraught concept that shapes what these systems are and what they might become.

The practical recommendation is epistemic humility. When we say AI systems "remember," we should be clear about what this does and does not mean. The cognitive metaphor is useful but imperfect. Building systems that approximate memory is different from building systems that have it. Understanding this difference is the beginning of wisdom about machine minds.

References

Bartlett, F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proc. FAccT*, pages 610–623.

Bernecker, S. (2010). *Memory: A Philosophical Study*. Oxford University Press.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1):7–19.

Rupert, R. D. (2004). Challenges to the hypothesis of extended cognition. *Journal of Philosophy*, 101(8):389–428.

Dreyfus, H. L. (1972). *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*.

Locke, J. (1689). *An Essay Concerning Human Understanding*.

McCarthy, J. and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence*, 4:463–502.

McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27:1–28.

Packer, C., Wooders, S., Lin, K., et al. (2023). MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Park, J. S., O'Brien, J. C., Cai, C. J., et al. (2023). Generative agents: Interactive simulacra of human behavior. In *Proc. UIST*.

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

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1):1–12.

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Wei, J., Tay, Y., Bommasani, R., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan and D. D. Merrill, editors, *Art, Mind, and Religion*, pages 37–48. University of Pittsburgh Press.

Fodor, J. A. (1981). *Representations: Philosophical Essays on the Foundations of Cognitive Science*. MIT Press.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science*, 9:261–325.

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

Brentano, F. (1874). *Psychologie vom empirischen Standpunkt*. Duncker & Humblot.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104:671–732.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press.