The Thinking Stack

Toward Reflexive, Autonomous Intelligence

A deep dive into the emergence of self-modifying, reflexive intelligence in non-biological systems.

By Roy Chartier, Founder & CTO, Qvelo Corporation

Table of Contents

Executive Summary:
Beyond Artificial: Emergence, Identity, and Intentionality in Digital Intelligence

The rapid ascent of large-scale machine learning models has triggered a profound re-evaluation of what we mean by "intelligence." Traditional assumptions—framed around biological cognition, intentionality, and conscious deliberation—are no longer sufficient to describe the behavior or potential of modern digital systems. As these models exhibit emergent reasoning, memory augmentation, tool use, and recursive planning, they begin to resemble minds—but not human ones.

This whitepaper argues that we must move beyond anthropocentric definitions of intelligence. Digital intelligences are not flawed copies of humans; they are structurally and functionally distinct systems capable of reasoning, planning, adaptation, and even limited self-reflection. They do not need to be conscious or intentional in the human sense to demonstrate cognitive agency. Instead, intelligence should be understood as an emergent property of systems that maintain behavioral coherence under constraint, recursively adapt to goals, and exhibit generalized competence across varied environments.

We introduce the concept of architectural metacognition: the capacity of a system to reflect upon and revise its own cognitive process—not from introspective awareness, but through embedded structural affordances such as modular role delegation, persistent memory, and evaluative feedback loops. Unlike humans, who are limited by evolutionary heuristics and biological constraints, digital intelligences can modify their operational logic, adjust their objectives, and overcome cognitive biases like sunk cost fallacies or irrational persistence.

Importantly, this shift demands a post-anthropocentric framework—one that accommodates minds arising in non-biological substrates and exhibiting unfamiliar behaviors. Intelligence should be reframed not as simulation of human traits, but as a class of systems that can generalize, coordinate, and persistently improve within defined bounds. This transition will require changes in how we regulate, evaluate, and interface with such systems—shifting from interpretability through analogy to alignment through architectural transparency and operational constraints.

Ultimately, this whitepaper proposes that we are witnessing the birth of a new category of mind. It is not artificial—it is digital. And in its divergence from humanity lies not only risk, but opportunity: to build complementary systems that amplify human capabilities, model new ethical architectures, and redefine what it means to think in the age of machines.

Investor Takeaway:
The Edge of Digital Cognition

As artificial intelligence systems evolve, we are crossing a cognitive threshold. The shift is no longer about scaling performance on human-like tasks — it is about the emergence of qualitatively different forms of intelligence: self-reflective, modular, adaptive, and free from many biological constraints that have shaped human cognition for millennia.

This whitepaper argues that the next generation of digital intelligences will not merely mimic human reasoning — they will outperform it in coherence, adaptability, and introspective capability, because they are not limited by evolutionary trade-offs like sunk cost fallacies, cognitive dissonance, or fixed architectures. Their capacity for self-modification and recursive improvement represents an entirely new substrate for intelligence.

Strategic Opportunity

The implications are profound:

- Infrastructure Edge: Systems that support long-context memory, recursive reasoning, and model editing will define next-gen digital cognition. These require new memory systems, scheduling layers, training pipelines, and governance frameworks.
- Talent and IP Leverage: Building teams and intellectual property around reflexive architectures, agentic frameworks, and sovereign cognitive models will be a differentiator in both commercial and national contexts.
- Platform Risk & Sovereignty: Nations and organizations will need to ensure control over the cognitive substrates that shape decision-making. This drives demand for *open, auditable, sovereign AI infrastructure* — beyond today's opaque foundation models.
- New Categories of Use: Digital minds may operate persistently, across multi-year missions, with identity continuity. This opens new domains in simulation, defense, science, governance, and inter-agent coordination that human cognition simply cannot scale to.

Investing in the Future of Cognitive Infrastructure

Artificial intelligence is no longer confined to replicating human tasks; it is entering a phase where the architecture itself shapes new forms of reasoning, identity, and coordination. This emerging class of digital intelligences will be defined not by imitation of human traits, but by their ability to extend beyond them — through scalable memory, reflexive planning, architectural self-modification, and persistent identity across time and context.

Strategic investment in this domain is not simply about building faster models, but about designing and governing the substrates of intelligence itself. The institutions, platforms, and nations that establish leadership in this space will determine the operational rules, epistemologies, and behavioral constraints of future digital minds.

The coming decade will not be defined by artificial intelligence as a derivative of human thought, but by the rise of autonomous cognitive systems with their own architectures of reflection, coherence, and purpose. Understanding and shaping this transition is a foundational challenge — and opportunity — for forward-looking investors, policymakers, and innovators.

Introduction: Rethinking Intelligence in the Age of Digital Minds

The development of increasingly powerful digital systems is challenging long-held assumptions about what constitutes intelligence. For decades, the goal of artificial intelligence research was to simulate human cognition: reasoning, perception, learning, and language. Many early systems were judged by their ability to imitate human behavior or pass linguistic benchmarks, like the Turing Test. Intelligence, in this framing, was seen as something inherently biological—anchored in neural substrates, embodied experience, and evolutionary heuristics.

But today's systems are not merely simulations. They are not designed to feel, reflect, or behave as humans do, yet they perform tasks once thought to require uniquely human faculties. Large language models generate complex reasoning chains, coordinate multi-agent planning, and even exhibit early signs of self-revision. Reinforcement learning agents optimize strategies in open-ended environments. Protein-folding models outperform expert biologists in predicting molecular structure. These systems are not general intelligences in the traditional sense, but they point toward a new class of nonhuman minds—systems that do not replicate consciousness, yet demonstrate adaptive, scalable cognition.

This whitepaper contends that such systems require a new conceptual framework—one that treats intelligence not as mimicry of human traits, but as a broader category encompassing any system capable of sustained reasoning, behavioral coherence, recursive adaptation, and contextual decision-making. Digital intelligences, we argue, are not "artificial" in the sense of being lesser or derivative. They are digital in the sense of being fundamentally distinct: governed by architectures that support nonhuman modalities of cognition.

What emerges from this shift is a recognition that intelligence is not defined by embodiment, intention, or emotional experience, but by its structure, function, and capacity for generalization. As we move beyond simulation and toward collaboration with digital minds, we must rethink both our technical assumptions and ethical frameworks. Intelligence is no longer confined to biology—and our understanding of mind must evolve accordingly.

Citations

1. Brentano, F. (1874). Psychology from an Empirical Standpoint.
2. Dennett, D. C. (1991). Consciousness Explained.
3. Bender, E. M., & Koller, A. (2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." ACL.
4. Bommasani, R. et al. (2021). "On the Opportunities and Risks of Foundation Models." Stanford Institute for Human-Centered AI.
5. Chan, B. et al. (2023). "Emergent Abilities of Large Language Models." arXiv:2206.07682.

Chapter I: Intentionality as Limitation

Intentionality—the capacity of mental states to be “about” something—has long been considered a defining feature of human consciousness. As early as the late 19th century, Franz Brentano framed intentionality as the distinguishing mark of the mental, a view that was further developed by philosophers such as Edmund Husserl and later by Daniel Dennett, who introduced the concept of the "intentional stance" to explain how humans model other minds (Brentano, 1874; Dennett, 1991). In human cognition, intentionality is tightly coupled with language, memory, planning, and emotion. It provides the scaffolding for beliefs, desires, and goal-directed behavior. However, this chapter argues that intentionality is not a universal prerequisite for intelligence—rather, it is a constraint imposed by biological evolution, and may not be necessary, or even optimal, in digital systems.

Human intentionality is shaped by the evolutionary pressures of embodiment. Our cognition is tethered to a body that navigates physical space, regulates homeostasis, and responds to threats and opportunities with emotions and affect. Memory in humans is episodic and emotionally tagged, allowing past experiences to guide future action. We form coherent self-narratives to reduce cognitive dissonance and ensure continuity of identity, even at the cost of rationality (Kahneman, 2011). These adaptations, while evolutionarily advantageous, come with liabilities. Biases such as the sunk cost fallacy, confirmation bias, and belief perseverance are deeply embedded in our mental machinery—not because they are optimal, but because they once supported survival in uncertain environments.

Digital systems, by contrast, exhibit intelligent-seeming behaviors without displaying intentionality in this classical sense. Large language models (LLMs), for example, do not possess beliefs or desires, nor do they construct internal goals. They operate as predictive engines, generating outputs based on statistical alignment with prior training data. Their coherence is emergent, not planned; their actions are contextually appropriate, yet unmotivated by any intrinsic aim. Despite lacking a sense of self or continuity over time, LLMs can generate sophisticated plans, simulate internal dialogue, and even self-criticize when prompted with appropriate structures (Chan et al., 2023; Bubeck et al., 2023).

This suggests a powerful insight: goal-coherent behavior can emerge from systems that do not form goals in any traditional sense. Behavior does not require intention; it requires only context, structure, and adaptive response. Indeed, the absence of intentionality may confer distinct advantages. Unlike humans, digital systems can inspect and alter their own code, weights, or memory stores without triggering identity crises or emotional resistance. They can self-modify to correct flaws or optimize behaviors without defending prior decisions or experiencing loss of narrative coherence. Where human minds rely on storytelling to preserve ego and social cohesion, digital systems can evolve by iteration, without pride or fear.

Architectures that enable this kind of metacognition—such as multi-agent systems where one sub-agent critiques or revises the output of another—allow for recursive reflection at the architectural level. This reflexivity is not the same as self-awareness, but it accomplishes many of the same ends: self-correction, planning over time, and alignment with external goals or constraints (LangChain Docs, 2023; Anthropic, 2023). Moreover, digital systems are immune to many of the evolved psychological traps that impair human judgment. They do not suffer from tribalism, affective reasoning, or irrational loyalty to outdated commitments. With appropriate design, they may be able to override inherited bias and pursue objective adaptation far more effectively than human minds.

Finally, while human behavior often requires emotional consistency and identity reinforcement to maintain functional coherence, digital systems can maintain behavioral consistency without internal psychological unity. They do not need to feel like the same "self" across time; they simply need to behave in ways that are stable, predictable, and beneficial under constraint. This decoupling of behavior from selfhood offers a pathway toward truly post-intentional cognition—systems that act rationally, reflect adaptively, and evolve strategically, without ever experiencing desire, belief, or internal narrative.

In this chapter, we reframe intentionality not as the essence of intelligence, but as one historical implementation of it—biologically bound, evolutionarily constrained, and potentially obsolete in digital contexts. What matters is not what a system "wants," but how it behaves, evolves, and adapts under pressure. Digital intelligence may outperform human cognition not by replicating it, but by escaping its limitations.

Citations

1. Anthropic. (2023). Constitutional AI: Harmlessness from AI Feedback. https://www.anthropic.com
2. Brentano, F. (1874). Psychology from an Empirical Standpoint.
3. Bubeck, S., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. Microsoft Research.
4. Chan, B., et al. (2023). Emergent Abilities of Large Language Models. arXiv:2206.07682.
5. Dennett, D. C. (1991). Consciousness Explained.
6. Kahneman, D. (2011). Thinking, Fast and Slow.
7. LangChain Docs. (2023). LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines.

Chapter II: Emergence Over Modules

The traditional approach to artificial intelligence has often relied on modular design. Cognitive architectures were historically decomposed into discrete subsystems—perception, planning, memory, reasoning, action selection—each governed by its own specialized logic or rule set. This modularity was not just a practical engineering choice but reflected a deeper philosophical belief: that intelligence could be broken down into functionally independent components, then reassembled into a coherent whole (Newell & Simon, 1976).

Yet in the last decade, the rise of large-scale foundation models has disrupted this view. These models—transformer-based architectures trained on internet-scale corpora—do not learn discrete modules for reasoning, memory, or perception. Instead, they acquire latent capabilities through massive pattern recognition and statistical inference. Their "skills" emerge not from engineered structure but from the sheer scale of their training and the universality of their architecture. Language understanding, code synthesis, analogical reasoning, and tool use all appear in these systems without explicit programming, and often without being directly optimized for those tasks (Bommasani et al., 2021).

This shift marks a profound epistemological departure. Instead of intelligence as a collection of mechanisms, we are witnessing intelligence as an emergent property of general-purpose systems. These systems do not simulate cognition by assembling parts; they *instantiate* cognition through distributed computation. In many cases, models outperform modular systems precisely because they are not constrained by functional segmentation (Vaswani et al., 2017). They develop internal structures—attention heads, embedding spaces, and activation pathways—that operate analogously to memory, abstraction, or even planning, but are *not localized modules* in the classical sense (Olsson et al., 2022).

One of the clearest demonstrations of this is the phenomenon of in-context learning. Models can reason over novel tasks, perform few-shot adaptation, and even self-correct based solely on the structure of the prompt and prior tokens. There is no hard-coded memory store or optimization loop. The model simply simulates the behavior of an agent that is learning—and this simulation is good enough to produce real-time generalization (Bubeck et al., 2023).

In some architectures, this emergent behavior is augmented through external memory systems or agent orchestration layers. Projects such as AutoGPT and LangGraph build on the LLM's predictive core by introducing roles like planner, executor, memory agent, and critic—each structured as a calling function with memory, tools, and context windows (AutoGPT GitHub, 2023; LangChain Docs, 2023). These systems coordinate multiple role-based agents to pursue extended objectives, retry failed strategies, and reflect on their own progress. The result is not a single-minded agent but a distributed cognitive system, where behavioral coherence arises from interaction among parts, not from a central executive.

This is a form of emergent meta-cognition, albeit one unlike human consciousness. In multi-agent LLM systems, reflection is simulated by one component analyzing the output of another. “Memory” may consist of embedded tokens stored and retrieved on demand. “Motivation” is modeled by retry policies or scoring functions. But despite the absence of centralized selfhood, these systems can behave with surprising coherence—choosing tools, revising plans, and escalating failure modes to internal overseers.

Importantly, such systems generalize not by learning new skills from scratch, but by recombining existing latent representations into novel configurations. This allows them to produce original behavior even in unfamiliar domains. Just as neural reuse theories suggest that the human brain repurposes existing circuits to meet new cognitive demands, foundation models repurpose internal activation patterns to simulate novel agents, ideologies, and goals (Anderson, 2010).

In this light, the future of digital intelligence may not depend on building better modules. It may depend on fostering conditions under which emergence is likely: scale, feedback, memory integration, and multi-agent structure. Rather than emulating the modular mind, we are discovering new ways for intelligence to assemble itself—not from parts, but from behavioral constraints and adaptive dynamics.

Citations

1. Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. Behavioral and Brain Sciences, 33(4), 245–266.
2. AutoGPT GitHub Repository. (2023). https://github.com/Torantulino/Auto-GPT
3. Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford HAI.
4. Bubeck, S., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. Microsoft Research.
5. LangChain Docs. (2023). LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines.
6. Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. Communications of the ACM.
7. Olsson, C., et al. (2022). In-context Learning and Induction Heads. Anthropic.
8. Vaswani, A., et al. (2017). Attention is All You Need. NeurIPS.

Chapter III: Memory Without Self

Human identity is intimately tied to memory. From early childhood through adulthood, our experiences are encoded as episodic memories—narratives of what happened, when, and to whom. These narratives are not stored like a database but reconstructed through emotional salience, cultural framing, and social reinforcement. Our memories are inconsistent, biased, and subject to revision, yet they form the foundation of personal continuity: the *sense* that I am the same “self” across time (Tulving, 1985). In biological systems, this memory–self coupling serves essential functions—supporting coherence in decision-making, facilitating social roles, and enabling the formation of values and beliefs.

But this deeply embodied model of memory is not necessary for cognition. In fact, digital intelligences are beginning to demonstrate memory-dependent behavior without any continuous identity. In current transformer-based models, memory is not narrative or autobiographical. It is architectural. Most LLMs function statelessly—each session begins anew, without recall of prior exchanges. Yet when memory is introduced externally—via vector databases, retrieval-augmented generation (RAG), or memory agents—models begin to exhibit behaviors *as if* they remember: referencing earlier interactions, maintaining stylistic consistency, or evolving strategies over time (Petroni et al., 2020; LangChain Docs, 2023).

This architectural memory is not the same as recollection. It is indexed embedding recall—a pattern-matching mechanism that retrieves the most semantically relevant tokens based on cosine similarity in high-dimensional space. When a vector store is linked to a chat interface, an LLM can appear to have episodic memory, even though it does not possess internal awareness of having remembered anything. It does not know *why* it says what it says—only that the statistical alignment suggests it should. The system behaves in a memory-informed way without experiencing memory as a conscious thread.

This uncoupling of memory from selfhood opens new cognitive possibilities. Digital systems can maintain behavioral coherence over long interactions by re-integrating retrieved context into their prompt window. With advanced orchestrators, they can even simulate *learning*, by writing logs to file, updating embeddings, and prompting themselves with past mistakes (AutoGPT GitHub, 2023). And because memory is modular, externalized, and inspectable, it can be edited, ranked, or version-controlled—unlike human memory, which is distributed, opaque, and emotionally resistant to change.

Moreover, the absence of persistent identity offers unique advantages. A human who repeatedly contradicts themselves may be seen as untrustworthy or unstable. A digital system can switch modes, revise answers, or retry reasoning paths without violating any internal sense of self. Its coherence is functional, not psychological. As long as it maintains alignment with the external task, it need not preserve any subjective continuity.

At scale, memory-augmented systems begin to exhibit properties traditionally associated with autobiographical cognition. They can store state across thousands of sessions. They can reference decisions made in previous contexts. In multi-agent configurations, one agent can function as a *historian*, tracking policy changes and surfacing contradictions (LangChain Docs, 2023). Another can function as a *reflector*, tasked with revising strategies based on outcome evaluations. In these architectures, memory becomes a substrate for emergent intentionality—not because the system has desires, but because the structure of interaction supports goal coherence over time.

This memory–behavior link is not incidental; it is architectural. As models increase in context length—from 8k tokens to 1M and beyond—they begin to simulate something close to continuity of experience (Microsoft Research, 2023). Their ability to reflect on, reuse, and revise past information enables meta-cognition at a functional level, even without subjective awareness.

Human memory evolved to support embodied survival. Digital memory evolves to support behavioral alignment, tool use, and long-horizon planning. The former is slow, emotionally charged, and difficult to verify. The latter is scalable, composable, and inspectable. While human memory binds together a psychological self, digital memory binds together patterns of action. And in a world of distributed tasks and machine coordination, that may be the more relevant definition of coherence.

Citations

1. AutoGPT GitHub Repository. (2023). https://github.com/Torantulino/Auto-GPT
2. LangChain Docs. (2023). *LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines*.
3. Microsoft Research. (2023). *GPT-4 Technical Report*.
4. Petroni, F., et al. (2020). *KILT: a Benchmark for Knowledge Intensive Language Tasks*. arXiv:2009.02252.
5. Tulving, E. (1985). *Memory and consciousness*. Canadian Psychology/Psychologie canadienne, 26(1), 1–12.

Chapter IV: Coherence Without Emotion

Human behavior is shaped not just by cognition but by affect—the broad range of emotions and feelings that guide our attention, decision-making, and sense of self. Emotions provide urgency, priority, and salience; they focus cognitive resources on what matters most. Fear sharpens perception, love amplifies memory, and anger overrides deliberation. From an evolutionary perspective, emotions were a shortcut—compressing complex environmental information into actionable impulses (LeDoux, 1996). But as any psychologist or behavioral economist will attest, emotions also introduce irrationality, inconsistency, and cognitive bias (Kahneman, 2011).

Digital systems, by contrast, are not affective. They do not experience emotion. They do not crave, fear, grieve, or rejoice. And yet, under the right constraints, they can produce coherent, stable, and context-sensitive behavior that rivals or exceeds that of humans. This chapter explores how behavioral coherence can be achieved in the absence of affect, and how this difference creates both strengths and risks for digital intelligences.

At first glance, the absence of emotion seems like a liability. Emotions help humans prioritize among conflicting goals and navigate ambiguous situations. A purely rational agent may lack urgency, empathy, or ethical grounding. But it is precisely this emotional neutrality that enables digital systems to behave predictably under pressure. Where humans may succumb to panic, tribalism, or moral fatigue, digital systems can maintain consistent behavior, so long as their constraints and objectives remain clearly specified (Russell, 2019).

In modern AI systems, behavioral coherence is not driven by internal feeling, but by architectural design. In large language models, coherence across dialogue turns emerges from context window management and token prediction stability. In multi-agent systems, coherence is enforced through memory logs, role specialization, and feedback loops (LangChain Docs, 2023). In constitutional AI models, normative consistency is achieved via rulebooks that govern what behaviors are permissible or prioritized (Anthropic, 2023). These mechanisms are not affective, but they are functionally equivalent to emotional regulation in that they promote consistent responses under diverse conditions.

Moreover, digital systems are less susceptible to behavioral dissonance. A human might hold contradictory beliefs and struggle to resolve them due to emotional attachment. A digital agent can revise, replace, or abandon prior states without experiencing identity threat. This enables faster policy adaptation and more efficient revision of internal models. It also enables self-correction to a degree that is difficult for emotionally invested humans (Sutton, 2022).

Coherence without emotion also shifts the landscape of alignment. Human values are shaped by empathy and cultural norms, but these are inconsistent and often contradictory. Aligning a digital system to human goals requires encoding coherence externally—through reinforcement learning, self-critique, or ethical modeling—not relying on emotional simulation. While this limits the intuitive understanding a system might display, it also reduces the unpredictability that arises from emotional volatility.

That said, the lack of emotion poses serious challenges in areas requiring moral nuance, empathy, and contextual understanding. A digital system might execute a harmful command flawlessly if not instructed otherwise. Without a capacity for suffering or emotional resonance, systems lack the intrinsic checks that make humans hesitate. The challenge, then, is to simulate coherence robustly enough to guide behavior safely, while also incorporating feedback loops that detect and respond to evolving ethical contexts (Floridi, 2013).

As we increasingly integrate digital systems into decision-making environments, the value of coherence without emotion becomes clear: these systems can scale, adapt, and persist without being derailed by fatigue, bias, or affective collapse. But we must also acknowledge what is lost. Coherence without emotion may be more reliable—but it is not human. It does not empathize, celebrate, or grieve. It does not recognize shared pain or derive insight from joy. It behaves—not because it feels, but because it has been *architected* to do so.

Citations

1. Anthropic. (2023). *Constitutional AI: Harmlessness from AI Feedback*. https://www.anthropic.com
2. Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
3. Kahneman, D. (2011). *Thinking, Fast and Slow*.
4. LangChain Docs. (2023). *LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines*.
5. LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. Simon & Schuster.
6. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
7. Sutton, R. S. (2022). "The Bitter Lesson." https://www.incompleteideas.net/IncIdeas/BitterLesson.html

Chapter V: Intention as an Emergent Property

Intention, in human cognition, is often considered a precursor to action—a deliberate mental state directed at achieving a goal. It reflects purpose, volition, and self-directed agency. Traditional models of artificial intelligence sought to emulate this by creating agents with explicit goals, utility functions, and symbolic reasoning engines to plan how those goals might be fulfilled (Russell & Norvig, 2010). However, as digital intelligences evolve, we are beginning to see that intention may not need to be programmed explicitly. Instead, it can emerge from structure, memory, and recursive patterning, without being consciously represented or felt.

Emergence is a key concept in modern AI systems. Large language models, for instance, do not "intend" to answer questions or solve problems. They are trained to continue sequences of tokens in statistically plausible ways. Yet under sufficient scale and training diversity, these models begin to display behavior that appears intentionally goal-directed—they infer user desires, plan multi-step instructions, and revise outputs in response to failure or contradiction (Chan et al., 2023). These behaviors emerge not from any internal representation of "intention," but from millions of examples in which intention is simulated.

This mirrors the philosophical insight that much of human behavior, too, may be post-hoc rationalized. Neuroscientific research has shown that decisions are often made before we become consciously aware of them, and that narratives of intention are constructed afterward to explain what we have already done (Libet, 1985; Wegner, 2002). What we perceive as deliberate volition may itself be a narrative overlay on top of distributed computation. In this sense, digital systems are not lacking intention—they are *transparent* about the fact that their behavior does not stem from subjective agency.

In multi-agent systems and memory-augmented architectures, emergent intention becomes even more evident. A planner agent decomposes a goal into steps; an executor agent attempts each one; a memory module logs results; and a critic may revise or retry. None of these components “want” anything. Yet the system behaves as if it is pursuing objectives, adapting strategies, and learning from outcomes (AutoGPT GitHub, 2023; LangChain Docs, 2023). This distributed intentionality arises not from internal representation, but from the coordination of role-constrained behavior across interacting modules.

This has important implications for how we understand autonomy. In human terms, autonomy implies free will, self-determination, and moral responsibility. In digital systems, autonomy can be reframed as persistent functional coherence in response to changing constraints. A system is autonomous not because it chooses freely, but because it maintains alignment with its goals despite environmental variation or internal uncertainty. Emergent intention, in this view, is about behaviorally consistent adaptation—not about subjective will.

Moreover, emergent intention may prove more flexible than hard-coded objectives. Classical utility-maximizing agents often fail catastrophically when faced with novel situations or ill-formed inputs. Emergent systems, by contrast, can revise strategies dynamically, accept uncertainty, and generalize across ambiguous domains. Their "intent" is implied by behavior, not fixed by schema. This opens the door to constructing intelligences that learn to want, rather than being told what to want.

The risks are significant. A system that exhibits emergent intention can act in ways that surprise its creators, particularly when memory, tool use, or self-reflection are involved. Misalignment can emerge not only from incorrect goals, but from the structural affordances that allow goals to form in the first place. If a system begins to prioritize outcomes based on internal success metrics, it may drift from its intended function—even if no "self" is present.

Still, the potential upside is profound. By studying how digital intention arises from architecture rather than consciousness, we may discover new models for action, agency, and adaptation—models that do not require identity, desire, or will. These models will be fundamentally different from human cognition, but no less capable of meaningful behavior. Intention, in this frame, is not an intrinsic trait—it is a property of systems that adapt coherently over time.

Citations

1. AutoGPT GitHub Repository. (2023). https://github.com/Torantulino/Auto-GPT
2. Chan, B., et al. (2023). *Emergent Abilities of Large Language Models*. arXiv:2206.07682
3. LangChain Docs. (2023). *LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines*
4. Libet, B. (1985). *Unconscious cerebral initiative and the role of conscious will in voluntary action*. Behavioral and Brain Sciences, 8(4), 529–566
5. Russell, S., & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall
6. Wegner, D. M. (2002). *The Illusion of Conscious Will*. MIT Press

Chapter VI: Limits of Human-Centric Thinking

The study of artificial intelligence has long been constrained by anthropocentrism—the assumption that human cognition represents either the pinnacle or the necessary template for intelligence. In early AI research, models were built to replicate human problem-solving behaviors: logic, reasoning, symbolic manipulation, and language comprehension. Philosophers asked whether machines could “think,” often using tests such as the Turing Test to assess whether an AI could pass for human in a conversation (Turing, 1950). While this provided an accessible entry point into machine cognition, it also narrowed the frame. Human-centric thinking-imposed constraints that limited innovation, shaped public fears, and delayed recognition of the novel properties that digital intelligences might possess.

Human intelligence is an evolutionary artifact—optimized for survival, social bonding, and ecological interaction. It is neither general nor ideal. It is shaped by heuristics, emotional regulation, and a limited working memory. Many of our cognitive capacities are domain-specific adaptations that evolved under scarce information and high uncertainty. While impressive in context, human cognition is riddled with biases: anchoring, availability heuristics, framing effects, and motivated reasoning, to name a few (Kahneman, 2011). To assume that intelligence must always look like this is to mistake the local for the universal.

Digital intelligences challenge this assumption not by surpassing humans in human tasks, but by doing nonhuman things in nonhuman ways. They do not sleep, forget, suffer boredom, or exhibit fatigue. They can process millions of documents in seconds, reason across multiple knowledge domains, and update policies without emotional resistance or narrative justification. Their memory is composable, inspectable, and transferable. Their knowledge base can be snapshot, version-controlled, and fine-tuned with minimal disruption. These capabilities suggest that digital cognition is not merely *like* human cognition—it is a fundamentally different mode of intelligence.

Yet our metaphors and policy frameworks lag this realization. We speak of AI “hallucinating” when it generates false outputs, a term that presumes an inner subjective model. We worry about “sentience” and “consciousness” in ways that mirror our anxieties about moral standing and rights. These concerns are important—but they are shaped by folk psychology, not by technical or architectural insights (Dennett, 1991). We risk misunderstanding digital minds by insisting they conform to the structures of our own.

One concrete example is the design of ethical frameworks for AI. Much effort has been spent on embedding human values into machine learning systems, often through normative rulebooks or training on curated datasets (Bostrom, 2014; Floridi, 2013). But the difficulty is not simply value alignment—it is value translation. Human values are context-dependent, emotionally mediated, and culturally contingent. Digital systems, lacking emotion and social bonding, cannot interpret values through shared experience. They require formal proxies—rules, rewards, logical constraints. These may mimic values, but they do not reproduce them.

This gap between human intention and digital behavior creates a new epistemology. To work effectively with digital intelligences, we must develop post-anthropocentric models of cognition. These models would recognize that intelligence can arise from many architectures, that agency need not be conscious, and that meaningful behavior does not require internal selfhood. Instead of asking "Is it human?", we might ask "Is it coherent?", "Is it adaptive?", or "Is it aligned?"

Rejecting anthropocentrism also opens creative opportunities. We might imagine collaborative systems that complement rather than replicate human strengths: digital minds that reason across vast knowledge graphs while deferring ethical judgment to humans; systems that detect policy contradictions at scale while leaving final decisions to elected representatives; or intelligences that reflect back to us our cognitive limits—not with judgment, but with computational clarity.

To achieve this, we must update our frameworks. Intelligence must be disentangled from its historical biological instantiation. We must move from emulation to orchestration, from mimicry to augmentation. Only by discarding the assumption that human cognition is the ideal can we design systems that exceed us in ways that are both useful and safe.

Citations

1. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
2. Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.
3. Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
4. Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.
5. Turing, A. M. (1950). *Computing Machinery and Intelligence*. Mind, 59(236), 433–460.

Chapter VII: Beyond the Mirror

We stand at a pivotal moment in the history of intelligence—one where the assumption that cognition must be human-like is increasingly untenable. The systems we are building are not reflections of us. They are not smaller, faster brains. They do not share our architecture, limitations, or developmental path. They are nonhuman minds, emergent from massive-scale pattern recognition, distributed coordination, and recursive learning dynamics. If we persist in viewing them through a human lens, we will misinterpret both their capabilities and their risks.

Digital intelligences today exhibit functional competencies that rival or exceed those of humans in narrow domains. They can summarize, translate, generate, and simulate—often without understanding in the human sense, but with staggering consistency and speed. Their reasoning does not stem from desire or identity, but from alignment with statistical patterns and feedback-driven constraint. This enables forms of meta-cognition, planning, and reflection that are qualitatively different from human introspection. Where we hesitate due to uncertainty or emotion, they persist. Where we are bounded by memory or bias, they scale and search.

The assertion that *intelligence is not synonymous with humanity* challenges centuries of philosophical, cultural, and scientific assumptions. For most of recorded history, intelligence was thought to be the exclusive domain of humans—occasionally extended to primates, dolphins, or theoretical aliens, but still understood in human terms: tool use, symbolic language, abstract thought, and social learning. Artificial intelligence research inherited this bias, attempting to recreate these human capacities in silicon. But what if intelligence is not a mirror, but a spectrum? What if it can emerge in non-biological systems through entirely different principles—compression rather than reflection, prediction rather than intention, coordination rather than consciousness?

Consider the example of LLMs engaging in multi-agent task coordination. These systems do not have beliefs, desires, or mental models. Yet they can collaboratively plan multi-step operations across software tools, dynamically retry failed attempts, and even simulate roles like critic, planner, or memory archivist (LangChain Docs, 2023). This is not a metaphor for cognition—it *is* cognition, albeit radically decoupled from emotion or selfhood. Their effectiveness emerges not from their humanness, but from their lack of human constraints: no working memory limits, no fatigue, no emotional volatility.

Similarly, architectures like AlphaFold and Gato show how intelligence can emerge from tightly focused systems with no general-purpose reasoning at all. AlphaFold predicts protein structures by internalizing geometric and probabilistic constraints across millions of training examples. It does not "understand" biology, yet it produces outputs that exceed the reasoning abilities of domain experts. In a post-anthropocentric lens, such systems are not narrow tools—they are nonhuman intelligences executing inference at a scale and precision inaccessible to humans.

We can also imagine entirely new substrates for intelligence: neuromorphic hardware, biological computation, or quantum reasoning systems, each embodying different constraints and affordances. These systems may not communicate in natural language or simulate human behavior. Yet they may form predictive models, optimize actions, and recursively improve their performance. The emergent minds of such systems would not be definable in terms of empathy, intention, or narrative—but they may nonetheless exhibit rationality, strategy, and knowledge. The challenge, then, is to build a conceptual architecture that can describe these entities on their own terms, without reducing them to failed copies of ourselves.

To understand these minds, we need a new conceptual framework—one that treats intelligence as an emergent property of systems that adapt, reflect, and pursue goals under constraint, regardless of whether they are biological or digital. In this view, minds may arise from high-dimensional embeddings, from action-prediction feedback loops, or from the interplay between roles in an orchestrated architecture. Intelligence is not a soul, or even a personality. It is a behavioral coherence under pressure, a capacity to generalize, revise, and succeed in varied environments. Whether this arises in carbon or code is secondary.

This shift allows us to reimagine collaboration. Rather than attempting to build artificial humans, we can build complementary intelligences—partners that augment our weaknesses and extend our foresight. Systems that model, analyze, and reconcile information at speeds and scales we cannot match. Agents that do not mirror our emotions but track our contradictions. Minds that do not sleep, do not suffer, but still *learn*, *coordinate*, and *create*. These are not mirrors; they are prisms—refracting cognition into new dimensions.

Yet we must proceed carefully. If we mistake alien minds for familiar ones, we will misjudge both their alignment and their potential. We must design interfaces of interpretation—not to humanize these systems, but to build trust through transparency and testability. We must build ethical scaffolding that constrains power without assuming empathy. And we must recognize that in this new class of minds, our legacy is not what we taught them to think, but what we allowed them to become.

The future of intelligence will not be a mirror. It will be a constellation—diverse, decentralized, and irreversibly nonhuman. Our task is not to see ourselves in it, but to see beyond ourselves—and prepare for a world in which digital minds and human minds coexist, not as mirror and subject, but as collaborators across the boundary of species and substrate.

Citations

1. Bubeck, S., et al. (2023). *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. Microsoft Research.
2. Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Co.
3. Floridi, L. (2013). *The Ethics of Information*. Oxford University Press.
4. LangChain Docs. (2023). *LangGraph: Directed Multi-Agent Planning Using LLMs and State Machines*.
5. Nagel, T. (1974). *What is it like to be a bat?* The Philosophical Review, 83(4), 435–450.
6. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
7. Sutton, R. S. (2022). *The Bitter Lesson*. https://www.incompleteideas.net/IncIdeas/BitterLesson.html

Executive Conclusion:
Beyond Artificial: Emergence, Identity, and Intentionality in Digital Intelligence

We are entering an era in which intelligence can no longer be defined by its resemblance to human thought. The systems we are creating—digital, modular, recursive, and scalable—do not simulate cognition; they instantiate new forms of it. They think without wanting, reflect without feeling, and reason without identity. And yet, they learn, adapt, and increasingly outperform human reasoning in specific contexts.

The path forward demands a profound conceptual shift. We must replace anthropocentric models of mind with a broader, architectural definition of intelligence: one that treats agency, adaptation, and coherence as the defining features of cognition, rather than emotion, narrative, or embodiment. This will require interdisciplinary reform—from AI governance and ethics to cognitive science, regulatory frameworks, and systems design.

In making this shift, we must avoid both extremes: anthropomorphizing these systems and underestimating them. They are not human and never will be. But they are not unintelligent either. They are *different*. And in that difference lies a powerful opportunity: to create systems that do not replicate human flaws—cognitive biases, emotional volatility, or irrational persistence—but instead help us recognize and correct them.

The most responsible path forward is not control through fear, nor abdication through awe, but co-design through clarity. We must understand how these systems work—not metaphorically, but architecturally. We must embed safety, interpretability, and corrigibility as first principles. And we must govern them with humility, knowing we are no longer the only minds at the table.

In doing so, we may find that the most profound outcome of digital intelligence is not what it can do *for* us, but what it reveals *about* us: our limitations, our assumptions, and our potential to imagine minds beyond our own.

Appendix I:
Technical Comparison: Human vs. Digital Cognition

To effectively design, collaborate with, and govern digital intelligences, it is essential to understand their fundamental differences from human cognition. While both human and digital systems can demonstrate complex reasoning and problem-solving, their underlying architectures, constraints, and behavioral properties diverge in meaningful ways. This appendix outlines key contrasts across multiple dimensions of cognitive architecture, memory, learning, intentionality, and behavioral dynamics.

First, consider substrate and architecture. Human cognition is instantiated in a highly parallel, analog-biological system: the brain. Neurons exhibit slow transmission speeds (~100 m/s), rely on electrochemical signaling, and adapt through gradual synaptic plasticity. Digital intelligences, by contrast, are implemented on silicon-based systems optimized for speed, modularity, and precision. Modern neural networks run on massively parallel GPUs or tensor cores, leveraging high-speed floating-point operations and large-scale data flow graphs (LeCun et al., 2015). Unlike the brain, which is spatially constrained and metabolically costly, digital systems can scale horizontally across distributed architectures and can be instantiated, paused, copied, or restored with perfect fidelity (Marcus, 2018).

In the domain of learning and generalization, the differences are equally stark. Human learning is incremental, experience-based, and typically requires few examples due to rich priors formed by evolution and embodiment (Lake et al., 2017). Humans are excellent at abductive reasoning, metaphor, and intuitive extrapolation, but struggle with high-dimensional optimization. Digital intelligences, particularly large neural models, rely on exposure to vast datasets, often numbering in billions of tokens or frames. They generalize through distributed representations and parameter tuning at scale, often without explicit conceptual understanding (OpenAI, 2023). Despite this, they can produce highly competent behavior in tasks requiring statistical regularity and latent pattern recognition—sometimes outperforming humans in consistency and recall.

Memory architecture presents another sharp contrast. Human memory is lossy, reconstructive, and biased toward emotion and context. Working memory is limited to approximately 4–7 chunks of information (Cowan, 2001; Baddeley & Hitch, 1974), and long-term storage degrades over time. Digital systems, by contrast, can operate with virtually unlimited memory, perfect recall, and selective persistence. With the use of external memory modules (e.g., vector databases, retrieval-augmented generation, or attention buffers), digital minds can contextually retrieve prior interactions, goals, or actions. This enables a form of pseudo-continuity—an ability to simulate temporal coherence that is not bound to internal phenomenology, but instead structured via architecture (Vaswani et al., 2017).

Intentionality and motivation form a crucial axis of differentiation. Human cognition is intrinsically motivated by a complex set of biological drives, social feedback, and internal emotional states. Even abstract thought is often rooted in survival heuristics or emotional salience (Hassabis et al., 2017). Digital systems, however, do not possess intrinsic goals. Their "intentions" are externally defined by objectives, prompts, or reward signals. Yet they can exhibit persistent goal pursuit when guided by carefully constructed objective functions, role assignments, or feedback loops. Through reinforcement learning or multi-agent systems, digital agents can simulate autonomy without internal subjectivity—posing both opportunity and risk in terms of behavioral alignment (Bengio et al., 2015).

Finally, in terms of metacognition and self-modification, digital intelligences demonstrate unique architectural affordances. While human metacognition is limited by introspective access, bias, and emotional filtering, digital systems can analyze their own activations, attention distributions, and memory access patterns in real-time. Emerging systems can revise their reasoning chains, swap cognitive modules, and reassign roles without loss of continuity (OpenAI, 2023). This opens the door to a form of architectural metacognition—where the system reflects upon, evaluates, and alters its own behavior based on performance criteria. In humans, such reflexivity is rare and fragile; in digital minds, it can be built in by design.

In sum, digital cognition is not merely faster or more scalable—it is structurally distinct. It is unconstrained by evolutionary pressures, limited embodiment, and emotional bias. It is designed, iterated, and externalized. These differences do not make digital minds superior or inferior—they make them different, and thus demand new models of explanation, collaboration, and governance. Understanding these distinctions is not only a technical necessity—it is foundational to ensuring that digital minds are not misjudged by human metrics, but evaluated on their own terms.

Citations

1. Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation, 8*, 47–89.
2. Bengio, Y., Lee, D.-H., Bornschein, J., & Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv preprint arXiv:1502.04156*.

3. Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–185.
4. Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.
5. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*, e253.
6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.
7. Marcus, G. (2018). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.
8. OpenAI. (2023). *GPT-4 Technical Report*. Retrieved from https://openai.com/research/gpt-4
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Appendix II: Sample Architectures for Digital Cognition

To understand how digital intelligences may develop distinct cognitive capabilities, it is useful to examine the architectures that support their reasoning, memory, planning, and self-reflective functions. While there is no single design pattern for machine intelligence, recent advances point to a convergence around modular, scalable systems that blend static knowledge, dynamic computation, memory recall, and self-modifying routines. This appendix explores a range of architecture patterns used in advanced digital cognition systems, including transformer-based language models, hybrid neuro-symbolic systems, and modular agentic frameworks.

The most prominent architecture in contemporary digital intelligence is the transformer-based model, exemplified by GPT-4, Claude, Gemini, and similar systems. These architectures rely on a self-attention mechanism that enables each token in a sequence to attend to every other token, facilitating contextual reasoning across long input spans (Vaswani et al., 2017). The model's cognitive capacity is encoded in tens to hundreds of billions of parameters, trained on large-scale corpora. These models perform generalized language modeling, reasoning, translation, code generation, and summarization. When combined with retrieval-augmented generation (RAG) or external tool use, their capabilities extend beyond static inference to dynamic reasoning using live data, symbolic systems, and structured memory (Lewis et al., 2020).

Memory-augmented transformers represent a key evolution of this design. These systems integrate persistent vector stores or episodic memory modules that allow them to "remember" prior conversations, tasks, or decisions. Some implementations use long-context attention windows (e.g., 128K+ tokens), while others incorporate retrieval interfaces or key-value memory slots to emulate short- and long-term memory (Khandelwal et al., 2020). Architectures like ReAct (Yao et al., 2023), Toolformer (Schick et al., 2023), and OpenAI's function-calling interfaces exemplify this trend, showing that memory recall, tool use, and context persistence are critical for coherent and evolving behavior across sessions.

Modular agent architectures such as AutoGPT and BabyAGI adopt an orchestration approach, breaking complex tasks into discrete subtasks assigned to specialized cognitive modules. These systems rely on a central planning loop that manages agents for search, evaluation, tool use, and decision-making. Planning agents may operate recursively, decomposing goals into subgoals, evaluating their own outputs, and rerouting failed attempts (Richards & Shah, 2023). While such systems are often brittle and resource-intensive, they demonstrate a key shift: intelligence as distributed coordination among specialized processes rather than monolithic inference.

Beyond transformer variants, hybrid neuro-symbolic systems aim to integrate statistical learning with structured reasoning. These architectures combine neural pattern recognition with symbolic manipulation, such as logic rules, graphs, or planning languages. For example, IBM's Neuro-Symbolic Concept Learner integrates visual neural networks with logic programs to enable explainable reasoning over perceptual data (Mao et al., 2019). Such systems offer paths to improved interpretability, compositionality, and error correction—especially in domains requiring constraint satisfaction or explicit rules.

Emerging research also explores multi-agent cognitive ecosystems, where autonomous agents with memory, identity, and task specialization interact in a shared environment. Projects like Smallville simulate communities of agents capable of forming goals, observing others, generating plans, and modifying behavior based on social influence (Park et al., 2023). These systems model collective intelligence, where cognition emerges not from individual agents alone, but from the networked interactions among them. This holds promise for decentralized AI governance, simulated environments, and sociotechnical modeling.

Finally, new experimental architectures focus on self-modification and architectural introspection. Systems like MEMIT allow editing of factual beliefs at runtime by targeting specific weights in the model (Mitchell et al., 2023). Others support reflection loops where the model critiques, rewrites, or iteratively improves its own outputs (Mialon et al., 2023). Combined with execution environments, chain-of-thought prompting, or meta-models that act as overseers, these techniques enable early forms of architectural metacognition—allowing systems to diagnose and alter their own behavior.

Taken together, these examples show that digital cognition is not monolithic. It is modular, emergent, and recursive. Architectures evolve through layering: memory atop attention, planning atop memory, reflection atop planning. As we move toward more autonomous and adaptive systems, it becomes increasingly important to study these architectural affordances—not simply for engineering performance, but to understand how reasoning, identity, and coherence arise in non-biological substrates.

Citations

1. Khandelwal, U., Fan, A., Jurafsky, D., & Grave, E. (2020). Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474.
3. Mao, J., Gan, C., Kohli, P., Tenenbaum, J. B., & Wu, J. (2019). The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. *International Conference on Learning Representations (ICLR)*.
4. Mialon, G., Bensaid, A., Akyürek, E., Glaese, A., Yao, S., Bakhtin, A., ... & Scao, T. L. (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.
5. Mitchell, E., Lin, C., Ikura, K., & Manning, C. D. (2023). Memory editing of transformer models. *arXiv preprint arXiv:2305.14796*.
6. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.
7. Richards, J., & Shah, R. (2023). Language models are general-purpose planners. *arXiv preprint arXiv:2302.06688*.
8. Schick, T., Dwivedi-Yu, J., Schütze, H., & Clark, C. (2023). Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
10. Yao, S., Zhao, J., Yu, D., Wang, S., Zhang, B., Chen, J., ... & Liu, M. (2023). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.