The AI Never-Ending Story

Agentic Frameworks and the Tale of Self-Reinvention

Jérôme Vetillard

VP R&D | Chief Product Officer | AI-Powered Healthcare & Life Sciences Products

Compliance by Design | PhD ENS Ulm | MIT Sloan | IE Business School & Brown University

The AI Never-Ending Story: Agentic Frameworks and the Tale of Self-Reinvention

Author : Jérôme Vétillard – Language : EN, Date : November 26, 2024

Introduction

Agentic AI frameworks are once again in the spotlight, celebrated by some as revolutionary advancements poised to transform artificial intelligence. Yet, the term ‘revolution’ can also be interpreted in its astronomical sense—the Moon’s revolution around the Earth—signifying a recurrence of past events rather than a wholly novel breakthrough.

A deeper examination reveals that the concept of agentic systems has historical roots stretching back to the earliest attempts to develop AI capable of adaptive, goal-directed behavior. These modern frameworks are less a new invention and more an evolution of

foundational ideas, refined and adapted for contemporary challenges. By revisiting this history, we can better understand why initial attempts fell short, assess whether contemporary frameworks have genuinely overcome foundational issues, and explore potential pathways forward—including the critical dimensions of cybersecurity and data governance that are too often absent from these discussions.

1. A Brief History: Agentic AI Isn't New

The aspiration to develop agentic AI—systems that can independently perceive their environment, set goals, and make decisions—has been a cornerstone of AI research for decades. In the 1980s and 1990s, the emergence of "intelligent agents" promised systems that could autonomously explore, make decisions, and achieve objectives within complex, dynamic environments. The term "agent" became a buzzword, encapsulating the hope for AI that could sense, reason, and act without constant human intervention.

Several pioneering implementations exemplify these early agentic frameworks:

Brooks' Subsumption Architecture (1986)

Rodney Brooks introduced the Subsumption Architecture as a novel approach to robotics. Eschewing traditional AI planning methods, this architecture used layered control systems where higher layers could subsume the roles of lower ones. This allowed robots to exhibit adaptive behaviors in real-time, reacting to environmental stimuli without the need for complex internal representations.

Distributed Artificial Intelligence (DAI) Systems

In the 1990s (with some references as early as 1975), DAI emerged as a field focused on solving complex problems through the collaboration of multiple agents, each with specific roles. This approach led to the development of Multi-Agent Systems (MAS), applied in areas like logistics, resource allocation, and cooperative robotics. MAS aimed to model and manage decentralized systems where agents could coordinate to achieve shared goals.

Belief-Desire-Intention (BDI) Model

Developed by philosophers and AI researchers like Michael Bratman, Anand Rao, and Michael Georgeff in the late 1980s and 1990s, the BDI model sought to formalize rational agency. It modeled agents based on three mental attitudes: beliefs (information the agent holds about the world), desires (objectives or situations the agent would like to bring about), and intentions (plans or actions the agent commits to in order to achieve its desires).

The BDI framework was influential in areas like simulation environments and early autonomous decision-making systems. However, it highlighted a critical limitation: the lack of genuine self-generated intent. While agents could act upon desires and intentions, these were ultimately predefined by programmers, lacking intrinsic motivation. These systems found applications across various domains, including robotics, network management, and

early forms of autonomous planning. The ambition was to tackle problems in environments where human control was impractical or inefficient, aiming for systems that could adapt and optimize without constant oversight.

2. Why Did Early Agentic Frameworks Fall Short?

Despite their innovative designs and the enthusiasm they generated, many early agentic systems struggled to meet expectations. Several key factors contributed to their stagnation.

Computational Limitations

Early agentic systems were constrained by the limited computational power of their time. Tasks involving real-time perception, decision-making, and planning in complex environments require significant processing capabilities. The combinatorial explosion of possible states and actions often rendered these systems impractical for anything beyond controlled, simplified environments. Moreover, their environmental ‘perception’ was rudimentary, a limitation that persists even in advanced systems like large language models (LLMs). While LLMs are celebrated for their diagnostic capabilities, their effectiveness depends heavily on comprehensive semiology and anamnesis provided by a healthcare professional, highlighting their continued reliance on external input for meaningful context—that is, clean and structured data.

Overly Narrow Domain Expertise

These systems were typically designed for specific tasks within well-defined domains. Their intelligence was narrow, excelling in particular environments or at executing specific tasks but lacking the ability to generalize. When confronted with novel problems or contexts outside their programming, their performance deteriorated rapidly.

Lack of Robust Learning Capabilities

Early agentic frameworks had rudimentary learning mechanisms, if any. Many relied on hardcoded rules and lacked the ability to adapt through experience. This rigidity made them brittle in the face of changing environments or unforeseen challenges, as they could not modify their behavior based on new information.

Inefficient Collaboration

In multi-agent systems, coordination and communication were significant hurdles. Without robust protocols for interaction, agents could work at cross-purposes, leading to conflicts and inefficiencies. Resolving these issues required complex algorithms for negotiation, conflict resolution, and consensus-building, which were challenging to implement effectively.

The Question of Intent: A Philosophical Clarification

Perhaps the most profound limitation was the question of intent. To understand why this matters, we must be precise about what we mean—and what we do not mean—by “intrinsic motivation” in the context of artificial agents.

The philosophical landscape here is contested. Daniel Dennett’s intentional stance treats intentionality as an interpretive framework: we attribute beliefs and desires to systems when doing so helps predict their behavior, regardless of whether those systems “truly” possess mental states. By contrast, John Searle’s concept of intrinsic intentionality demands that genuine understanding or purpose arise from the system’s own nature—a standard that no current AI system meets. The BDI model formalized desires and intentions within an explicitly computational framework, but could not imbue agents with authentic self-driven purpose in Searle’s sense.

This distinction is not merely academic. It determines what we can reasonably expect from agentic systems. If we adopt a Dennettian view, the question becomes functional: can an agent adapt its goals through meta-learning in ways that are indistinguishable from self-generated motivation? Recent work on intrinsic curiosity in reinforcement learning—notably Pathak et al. (2017) on curiosity-driven exploration and Burda et al. (2018) on large-scale studies of curiosity-based agents—suggests that agents can develop behaviors that functionally resemble intrinsic motivation, exploring environments without extrinsic rewards. However, this functional autonomy still operates within reward structures designed by humans, and the agent has no capacity to question or transcend the meta-objectives that frame its exploration.

If we adopt a Searlean standard, the gap remains absolute: no computational process, however sophisticated, constitutes genuine understanding or purpose. For practical purposes in agentic system design, the relevant question is not whether agents “truly” have intentions, but whether they can exhibit sufficiently adaptive, context-sensitive goal modification to operate reliably in open-ended environments. The honest answer, as of today, is that they cannot—not because of a metaphysical deficit, but because of concrete limitations in generalization, planning under uncertainty, and robust self-correction.

3. What Is Structurally Different Today?

Fast forward to the present, and agentic AI frameworks are experiencing a resurgence. However, to understand whether this resurgence is substantive rather than merely cyclical, we must go beyond the generic trifecta of “more compute, better algorithms, bigger data” and identify what is structurally new in the current generation of agentic architectures.

The Obvious Advances

Enhanced computational resources (GPUs, TPUs, and increasingly specialized AI accelerators) allow for real-time processing of complex tasks that were previously intractable. Advanced learning algorithms, particularly deep reinforcement learning, enable agents to learn from experience and adapt to new situations. The proliferation of data

provides agents with rich training environments. These are necessary conditions for the current resurgence—but they are not sufficient to explain what is qualitatively new.

The Genuine Paradigm Shift: Language as Planning Substrate

The most significant structural difference is the emergence of large language models as general-purpose planners and reasoners. This represents a qualitative break from earlier agentic architectures, not merely a quantitative improvement. In classical BDI systems, beliefs, desires, and intentions were represented in formal logic or domain-specific languages. Planning was constrained to the expressivity of these representations. Communication between agents required predefined ontologies and protocols.

In LLM-based agentic systems, natural language serves simultaneously as the substrate for representation, planning, reasoning, and inter-agent communication. This is not a minor convenience—it fundamentally changes the design space. Yao et al. (2023) formalized this in the ReAct framework, demonstrating that interleaving reasoning traces with action steps in natural language enables agents to plan, self-correct, and interact with external tools in a unified loop. Shinn et al.'s Reflexion framework extends this further, showing that agents can use linguistic self-reflection to improve performance across episodes without parameter updates—a form of learning that operates entirely in the space of natural language.

The implications for multi-agent coordination are profound. Where DAI systems of the 1990s required laboriously engineered communication protocols, LLM-based agents can negotiate, debate, and coordinate using the same natural language interface. This dramatically lowers the barrier to building multi-agent systems, though it introduces new challenges around the reliability and consistency of natural language communication.

Tool Use and Environmental Interaction

A second structural difference is the capacity of LLM-based agents to interface with arbitrary external tools—APIs, databases, code interpreters, web browsers—through natural language descriptions of tool capabilities. This addresses, at least partially, the narrow domain expertise problem that plagued earlier systems. An LLM-based agent can, in principle, expand its capabilities by learning to use new tools at inference time, without retraining. However, the reliability of tool use remains uneven, and errors compound across multi-step tool chains in ways that are difficult to predict or control.

4. Persistent Challenges: Are We Truly Solving the Old Problems?

Despite these structural advances, several core issues persist—and some are exacerbated by the new architectures.

Generalization and the Evaluation Problem

Achieving true generalization—where agents can adapt to entirely new environments and tasks—continues to be a significant challenge. Contemporary agents often excel in domains they were trained on but struggle with transfer learning or zero-shot generalization. The reliance on large datasets tailored to specific tasks means that adaptability across diverse contexts is limited.

Crucially, the field lacks robust evaluation frameworks for agentic systems operating in open-ended environments. Benchmarks like SWE-bench (for software engineering tasks), WebArena (for web-based tasks), and GAIA (for general AI assistants) represent important steps, but they also reveal sobering limitations. Performance on these benchmarks drops precipitously as task complexity increases, and the gap between benchmark performance and real-world reliability remains substantial. The evaluation problem is not merely technical: it is conceptual. How do we measure the competence of a system whose operating environment is, by design, unbounded? Traditional metrics like accuracy or task completion rate assume well-defined success criteria, which may not exist for genuinely autonomous agents. This challenge echoes the difficulties faced by early DAI researchers but is amplified by the broader scope of contemporary systems.

Coordination, Emergence, and Alignment

Coordinating multiple agents to achieve coherent, beneficial outcomes is still difficult. While techniques like multi-agent reinforcement learning have made strides, ensuring that emergent behaviors are aligned with desired objectives remains a research frontier. Unintended consequences and chaotic interactions can arise in complex systems without careful design.

The alignment problem—ensuring that agentic systems act in accordance with human values—takes on new dimensions with LLM-based agents. Techniques like RLHF (Reinforcement Learning from Human Feedback), constitutional AI, and mechanistic interpretability represent concrete approaches to alignment, but they were primarily developed for single-model, single-turn interactions. Extending these techniques to multi-agent, multi-step agentic workflows introduces compounding risks. Reward hacking—where agents find ways to maximize reward signals without achieving the intended objective—becomes more likely as action horizons lengthen and the space of possible strategies expands. Specifying reward functions for open-ended tasks is inherently difficult, and misspecification can lead to behaviors that are technically optimal but practically harmful.

The Intent Gap: Functional vs. Genuine Autonomy

As discussed in Section 2, modern agents remain functionally constrained even when they appear highly autonomous. Their goals are shaped by reward functions, data biases, and human-defined objectives. The intrinsic curiosity mechanisms explored in recent research represent genuine progress toward functional autonomy, but they do not constitute self-generated purpose in any philosophically robust sense. For system design purposes, the practical question is whether agents can exhibit sufficient adaptive goal modification to be

useful in complex, dynamic environments—and the current evidence suggests we are far from that threshold in most real-world domains.

5. The Cybersecurity Blind Spot: Access Control, Data Governance, and Attack Surface Expansion

A dimension conspicuously absent from most discussions of agentic AI frameworks is cybersecurity. As agentic systems gain the ability to autonomously access, process, and transmit data across organizational boundaries, they introduce a category of security risks that existing frameworks are ill-equipped to handle. This is not a secondary concern—it is, in many deployment contexts, the primary obstacle to adoption.

The Access Control List (ACL) Problem

In traditional information systems, access control is governed by well-defined policies: users authenticate, receive role-based permissions, and access only the resources their roles authorize. When an agentic system operates on behalf of a user—or, more critically, on behalf of multiple users with different privilege levels—the identity and authorization model becomes profoundly more complex.

Consider a multi-agent system deployed in a hospital setting, where agents access electronic health records (EHRs), laboratory results, imaging data, and administrative databases. Each data source has its own ACL, governed by regulations such as GDPR (in Europe) or HIPAA (in the United States). An agent operating across these sources must respect fine-grained access policies that depend not only on the agent's own authorization level but also on the authorization of the human user who initiated the request, the purpose of the access (clinical care vs. research vs. billing), and the sensitivity classification of the specific data being accessed. Current agentic frameworks typically operate with a single set of credentials, often those of the deploying user or a service account, creating a de facto privilege escalation risk. The agent may inadvertently access data that the initiating user is not authorized to see, or aggregate information across sources in ways that violate the principle of minimum necessary access. This is not a theoretical risk: it is an architectural deficiency in most existing agentic deployments.

Data Leakage and Exfiltration Risks

Agentic systems that interact with external tools, APIs, and web services create data exfiltration pathways that are difficult to monitor and control. When an agent sends a query to an external API, the query itself may contain sensitive information extracted from internal documents. When an agent uses a code interpreter to process data, intermediate outputs may be written to temporary storage that is not subject to the same security controls as the source data.

The risk is compounded in multi-agent architectures where information flows between specialized agents. Each inter-agent communication channel represents a potential leakage point. A “White Hat” agent that aggregates factual data from multiple sources may inadvertently pass protected health information (PHI) to a “Green Hat” agent tasked with generating creative alternatives—which might then include that PHI in outputs visible to unauthorized users. In the context of healthcare and life sciences, where TweenMe operates, these risks are not abstract. The EU AI Act’s requirements for high-risk AI systems (which include medical devices and clinical decision support) demand demonstrable data governance, traceability of data flows, and robust access control mechanisms. The Medical Device Regulation (MDR) further requires that software as a medical device (SaMD) maintain the integrity, confidentiality, and availability of patient data throughout its lifecycle. Agentic systems that cannot provide verifiable guarantees on these dimensions face significant regulatory barriers to deployment.

Prompt Injection and Adversarial Manipulation

LLM-based agentic systems introduce a novel attack surface: prompt injection. An adversary who can insert malicious content into any data source that the agent processes—a web page, a document, an email, a database field—can potentially manipulate the agent’s behavior. In an agentic context, the consequences of successful prompt injection are far more severe than in a simple chatbot, because the agent has the ability to take actions: sending emails, modifying files, executing code, or making API calls.

The research community has identified several categories of prompt injection attacks: direct injection (embedding instructions in user inputs), indirect injection (placing instructions in data sources the agent will process), and multi-step injection chains where benign-looking instructions across multiple sources combine to produce harmful behavior. Defense mechanisms remain immature. Input filtering, output validation, and sandboxing can mitigate specific attack vectors, but no comprehensive defense exists. For agentic systems operating in high-stakes environments—healthcare, finance, critical infrastructure—this represents an unacceptable risk posture until robust defenses are developed and validated.

The Auditability and Traceability Imperative

Regulatory frameworks increasingly require that AI systems provide audit trails demonstrating what data was accessed, what decisions were made, and what actions were taken. For agentic systems, this requirement extends across potentially long chains of reasoning and action, involving multiple agents, tools, and data sources. Current agentic frameworks typically provide logs at the individual action level but lack integrated mechanisms for end-to-end traceability. Reconstructing the causal chain from an initial user request to a final output—including all intermediate reasoning steps, data accesses, and inter-agent communications—remains technically challenging. Without this capability, compliance with regulations like the EU AI Act’s transparency requirements or the MDR’s clinical evaluation obligations is difficult to demonstrate.

6. A Hybrid Approach: Humans as Top-Level Orchestrators

One promising avenue to address these challenges—both the longstanding ones and the cybersecurity concerns—is adopting a hybrid model where humans remain central to the decision-making process. In this framework, specialized agents perform tasks within their domains of expertise, but humans provide overarching guidance, strategic intent, and ethical oversight.

While some advocate for the pursuit of General AI, at TweenMe we believe it is more pragmatic and more responsible to leverage the human brain for high-level orchestration. This does not mean we reject agentic frameworks—we actively employ them for data processing pipeline optimization—but we recognize the need to clearly delineate where autonomous agent operation is appropriate and where human judgment must remain sovereign.

Our vision is to empower users with a sophisticated yet purpose-built AI-infused toolbox, allowing them to select and sequence tools based on their unique strategies and expertise. Instead of developing an AI orchestrator capable of addressing every conceivable data ‘monetization’ scenario, we focus on harnessing the knowledge and proficiency of data stewards to drive optimal and context-specific outcomes.

Advantages of this approach include the leverage of complementary strengths, where agents handle data-intensive, routine, or well-defined tasks efficiently while humans contribute creativity, moral judgment, and adaptability. Enhanced alignment is achieved because human oversight ensures that agentic behaviors remain aligned with human values and societal norms, reducing the risk of unintended consequences. From a cybersecurity perspective, human-in-the-loop architectures provide natural checkpoints for access control validation: the human orchestrator can verify that data accesses are appropriate, that outputs do not contain sensitive information, and that agent behaviors remain within authorized boundaries. Finally, humans provide the flexibility and responsiveness needed to adjust goals and strategies in response to changing circumstances.

7. Chain-of-Thought Meets Multi-Agent Specialization: Beyond the Analogy

An innovative concept in advancing agentic AI frameworks is the integration of Chain-of-Thought (CoT) reasoning with multi-agent specialization. CoT breaks down complex problems into a sequence of intermediate reasoning steps, enhancing both transparency and interpretability. This structured approach can be further enriched by drawing inspiration from the Six Thinking Hats methodology, where specialized agents are assigned distinct roles to handle specific aspects of problem-solving.

The White Hat Agent focuses on analyzing and presenting factual data from inputs. The Red Hat Agent evaluates sentiment, subjective perspectives, and emotional nuances. The Black Hat Agent identifies risks and provides critical analysis. The Yellow Hat Agent explores potential benefits and opportunities. The Green Hat Agent generates creative alternatives and innovative approaches. The Blue Hat Agent oversees and orchestrates the entire reasoning process, ensuring coherence and balance.

From Analogy to Architecture: Conflict Resolution and Formal Grounding

The Six Thinking Hats analogy is pedagogically useful, but for this framework to be operationally viable, several architectural questions must be addressed that go beyond metaphor.

The first and most critical question is conflict resolution. When the Black Hat Agent (risk assessment) and the Yellow Hat Agent (opportunity identification) produce contradictory conclusions—as they frequently will—how does the system resolve the tension? The Blue Hat orchestrator must implement a principled conflict resolution mechanism. Several approaches are possible: weighted aggregation based on domain-specific confidence scores; structured debate where conflicting agents present arguments and the orchestrator (human or meta-agent) adjudicates; Delphi-style iterative refinement where agents revise their positions in light of others' analyses; or hierarchical override where certain agent types have precedence in specific contexts (e.g., the Black Hat always has priority in safety-critical decisions). The choice of mechanism has profound implications for system behavior and must be made explicitly rather than left to emerge from ad hoc interaction.

This modular multi-agent approach has formal precedents that strengthen its theoretical foundation. Irving et al. (2018) proposed AI safety via debate, where two AI agents argue for opposing positions and a human judge evaluates the arguments—directly analogous to the Black Hat/Yellow Hat dynamic. Zhuge et al. (2023) explored the "Society of Mind" paradigm for multi-agent LLM collaboration, demonstrating that structured role assignment improves both the quality and interpretability of collective reasoning. Wang et al.'s work on multi-persona approaches shows that assigning distinct analytical personas to agents produces more diverse and robust analyses than single-agent systems. These formal frameworks transform the Six Thinking Hats from a suggestive analogy into a grounded architectural pattern with empirical support.

Hierarchical Composition: Specialized Agents Within Specialized Agents

A further advantage of this modular architecture is its composability. Each high-level "Hat" agent can itself orchestrate lower-level specialized agents for narrow automation tasks. For instance, the White Hat Agent might coordinate a data extraction agent, a statistical analysis agent, and a data quality validation agent. This hierarchical composition provides depth of specialization while maintaining the interpretability benefits of structured role decomposition.

Key advantages of this modular framework include enhanced performance through agent specialization, improved interpretability by breaking decisions into discrete and transparent

steps, and centralized oversight through a human or meta-agent orchestrator that integrates outputs from specialized agents. From a cybersecurity standpoint, modular architectures also enable fine-grained access control: each specialized agent can be granted only the minimum necessary data access permissions for its specific role, implementing the principle of least privilege at the agent level rather than granting blanket access to the entire system.

8. Building Towards the Future: Breaking the Cycle

The renewed interest in agentic AI is fueled by technological advances that make these systems more accessible and capable than ever before. The structural novelty of LLM-based agentic architectures—language as planning substrate, flexible tool use, natural language coordination—represents a genuine qualitative shift from earlier generations. However, to avoid repeating past disappointments, we must critically assess whether we are genuinely overcoming foundational challenges or simply deferring them.

The foundational questions remain open. On intrinsic motivation: can we develop agents with functionally robust adaptive goal modification—whether or not this constitutes “genuine” intent in a philosophical sense—sufficient for reliable operation in open-ended environments? On robust generalization: how can agents learn to transfer knowledge across domains, and how do we evaluate this capacity rigorously when the operating environment is unbounded? On ethical alignment: what mechanisms ensure that agentic systems act in accordance with human values as they gain longer action horizons and broader tool access, and how do current alignment techniques scale to multi-agent, multi-step workflows? On cybersecurity and data governance: how do we architect agentic systems that respect fine-grained access control policies, prevent data leakage across agent boundaries, resist adversarial manipulation, and provide the end-to-end auditability that regulatory frameworks demand?

Adopting hybrid models that keep humans at the center of orchestration offers a pragmatic path forward. By combining human strategic oversight with agentic efficiency—and by designing architectures where security and governance are foundational rather than afterthoughts—we can harness the strengths of both to tackle complex, multifaceted problems. Integrating structured multi-agent specialization with formal conflict resolution mechanisms and hierarchical composition provides a concrete architectural pattern for building systems that are both capable and interpretable.

History tends to repeat itself, but we have the opportunity to learn from past experiences and steer the development of agentic AI toward genuinely transformative outcomes. The path forward requires not only technical innovation but also architectural rigor in security design, intellectual honesty about the limitations of current systems, and a willingness to address foundational issues head-on rather than deferring them behind a wave of enthusiasm. By doing so, we can advance toward AI systems that are not only autonomous and adaptable but also trustworthy, auditable, and aligned with both human values and regulatory requirements.

References

Brooks, R. A. (1986). A robust layered control system for a mobile robot. IEEE Journal on Robotics and Automation, 2(1), 14–23.

Bratman, M. E. (1987). Intention, Plans, and Practical Reason. Harvard University Press.

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95).

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Large-scale study of curiosity-driven learning. arXiv:1808.04355.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-predictive next. Proceedings of the 34th International Conference on Machine Learning (ICML).

Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

Yao, S., Zhao, J., Yu, D., et al. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR).

Shinn, N., Cassano, F., Gopinath, A., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems (NeurIPS).

Zhuge, M., et al. (2023). Mindstorms in natural language-based societies of mind. arXiv:2305.17066.

Wang, Z., et al. (2023). Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. arXiv:2307.05300.

European Parliament (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act).

European Parliament (2017). Regulation (EU) 2017/745 on medical devices (MDR).