

2026-01 · V1.0.0

Agentic Collapse

QNFO Research

QNFO-ULA-2.1

ABSTRACT

Abstract The transition from static Large Language Models to dynamic Autonomous Agents has introduced a new class of epistemic failure modes driven by the “Orchestration Penalty”—the factorial increase in verification complexity inherent to multi-step reasoning loops. This research identifies a structural “Latency Killer” in System 2 architectures, where the velocity of agentic action (ϕ) inherently outpaces the latency of deliberate verification (τ_{sys2}). We introduce a time-delayed stochastic dynamical system ($d\phi, d\psi, dU$) that models the accumulation of “Agentic Drift.” Simulation results reveal a catastrophic phase transition termed “Agentic Collapse,” where a stochastic complexity spike ($t \approx 60$) drives the system into an irretrievable high-entropy state before the lagged verifier can compensate. We demonstrate that perfect stability is mathematically impossible in open-ended agentic loops; instead, safety requires a regime of “Metastability” enforced by a “Popperian Guillotine”—a discrete safety interlock that resets the agent when epistemic potential exceeds a critical threshold ($U_{crit} = 0.88$). [...]

CITE AS Research, QNFO (2026). Agentic Collapse. QNFO Preprint v1.0.0.

A

Time-Delayed Cybernetic Framework for Epistemic Stability in Autonomous AI Systems

Author: Rowan Brad Quni-Gudzinis **Contact:** rowan.quni@outlook.com **ORCID:** 0009-0002-4317-5604 **ISNI:** 00000000526456062 **URL:** <https://papers.qnfo.org/papers/agentic-collapse/> **Date:** 2026-01-02 **Version:** 1.0

Abstract The transition from static Large Language Models to dynamic **Autonomous Agents** has introduced a new class of epistemic failure modes driven by the “Orchestration Penalty”—the factorial increase in verification complexity inherent to multi-step reasoning loops. This research identifies a structural “Latency Killer” in System 2 architectures, where the velocity of agentic action (ϕ) inherently outpaces the latency of deliberate verification (τ_{sys2}). We introduce a time-delayed stochastic dynamical system ($d\phi, d\psi, dU$) that models the accumulation of “Agentic Drift.” Simulation results reveal a catastrophic phase transition termed “**Agentic Collapse**,” where a stochastic complexity spike ($t \approx 60$) drives the system into an irretrievable high-entropy state before the lagged verifier can compensate. We demonstrate that perfect stability is mathematically impossible in open-ended agentic loops; instead, safety requires a regime of “**Metastability**” enforced by a “**Popperian Guillotine**”—a discrete safety interlock that resets the agent when epistemic potential exceeds a critical threshold ($U_{crit} = 0.88$). Finally, we derive an “**Epistemic Speed Limit**,” arguing that robust autonomy requires dynamically throttling execution speed to match the “Cost of Thought.”

Keywords: Agentic AI; System 2 Reasoning; Neuro-Symbolic Integration; Delay Differential Equations; AI Safety; Popperian Falsifiability; Cybernetic Control Theory.

1. INTRODUCTION

1.1 From

Stochastic Parrots to Autonomous Agents

The trajectory of artificial intelligence has shifted decisively from static text generation to dynamic agentic workflows. This transition marks a fundamental change in how large language models are deployed, moving them from passive oracles to active participants in complex environments. Guo et al. (2025) describe this evolution as the rise of agentic workflows, where the primary unit of computation is no longer the single prompt but the iterative loop. In this paradigm, models do not merely predict the next token; they formulate plans, execute tools, and observe results in a continuous cycle. Wang et al. (2024) provide a comprehensive survey of these autonomous agents, highlighting their ability to handle long-horizon tasks previously out of reach for zero-shot systems. The shift is driven by the integration of memory modules, planning algorithms, and external tool interfaces that allow the model to interact with the world. Consequently, the epistemic burden has moved from the accuracy of a single statement to the stability of a sequential process.

The historical context for this shift lies in the limitations of the stochastic parrot paradigm. While early transformer models demonstrated impressive linguistic fluency, they lacked the ability to maintain coherence over extended problem-solving sessions. The introduction of frameworks like AutoGen (Wu et al., 2023) and MetaGPT (Hong et al., 2024) provided the scaffolding necessary to chain these stochastic outputs into coherent workflows. These frameworks allow developers to define roles and standard operating procedures that guide the model through multi-step tasks. By decomposing complex problems into smaller, manageable sub-tasks, these systems attempt to bypass the inherent reasoning limits of the underlying model.

The underlying mechanism of these agentic systems is the recursive application of inference. An agent generates a thought, translates that thought into an action, and then processes the feedback from that action as a new input. This creates a feedback loop where the model's own outputs become part of its future context. Ideally, this loop allows for self-correction. However, this recursive structure also introduces the risk of error propagation. A small hallucination in the planning phase can cascade into a catastrophic failure in the execution phase. The system relies on the assumption that the model's reasoning capabilities are robust enough to recover from these deviations.

1.2 The System 2

Illusion: Inference vs. Reasoning

The current generation of reasoning models attempts to simulate the deliberate, analytical thought processes characteristic of human System 2 cognition. Xu et al. (2025) survey this transition, noting the emergence of models explicitly designed to “think” before they speak. These models use techniques such as chain-of-thought prompting to generate intermediate reasoning steps before producing a final answer. The goal is to force the model to decompose the problem and verify its own logic. This approach has yielded significant improvements in mathematical and coding benchmarks, suggesting that the models are capable of genuine problem-solving.

The context for this development is the observation that standard language models often fail at tasks requiring multi-step deduction. By forcing the model to externalize its reasoning, researchers hope to make the process more transparent and less prone to logical leaps. Yao et al. (2024) formalized this with the Tree of Thoughts framework, which allows the model to explore multiple reasoning paths and backtrack when it encounters a dead end. This search-based approach mimics the cognitive processes of planning and evaluation, representing a move away from greedy decoding strategies.

However, evidence suggests that this verification process is often circular. The model uses the same set of weights to generate the solution and to verify it. If the model has a fundamental misconception about the problem, it will likely hallucinate a justification for its incorrect answer. This leads to self-delusion, where the model becomes increasingly confident in a wrong conclusion because it has generated a plausible-sounding rationale for it. The verification is not grounded in an external truth but in the model's own internal consistency.

1.3 The

Orchestration Penalty: Complexity vs. Control

The expansion of AI into multi-agent architectures introduces a factorial increase in system complexity that we term the “Orchestration Penalty.” Mofrad (2025) reviews the state of modular AI agents, describing systems where distinct agents are assigned specific roles such as coder, reviewer, or manager. The premise is that specialization allows each agent to perform better at its specific task. However, the mechanism of these multi-agent systems—inter-agent communication—introduces a new source of error: synchronization failure. If one agent misunderstands a message or operates on outdated information, the entire workflow can derail.

Evidence of “groupthink” is emerging in the literature. When multiple agents interact without an external ground-truth oracle, they often converge on plausible but incorrect solutions. The verification burden shifts from checking a single output to monitoring a complex web of interactions. Verifying the correctness of a multi-agent system is exponentially harder than verifying a single agent because the state space of possible interactions is vast. The orchestration layer itself becomes a source of fragility.

1.4 Normative

Epistemology in Autonomous Loops

The deployment of autonomous agents in high-stakes environments necessitates a return to normative epistemology, specifically the requirement for automated falsifiability. Popper (1959) established the criterion that a scientific statement must be falsifiable to be valid. In the context of autonomous loops, this means the system must be able to detect when its actions or beliefs contradict reality. Abbasi-Yadkori et al. (2024) propose information-theoretic metrics to distinguish between epistemic uncertainty (lack of knowledge) and aleatoric uncertainty (inherent randomness). This distinction is crucial for an agent to know when it is operating outside its competence.

The mechanism for this control is the operationalization of falsifiability as a runtime check. We cannot rely on the model’s internal confidence scores, which are often uncalibrated. Instead, we need external validation functions that can reject the agent’s outputs. These functions act as the “reality check” for the system. When the agent proposes an action, the validator checks it against a set of logical or physical constraints. If the action violates a constraint, it is rejected, and the agent is forced to revise its plan.

1.5

Neuro-Symbolic Integration: The Open World Problem

The field of neuro-symbolic AI attempts to bridge the gap between neural intuition and symbolic rigor, but it faces a fundamental barrier known as the open world problem. Wan et al. (2024) survey the current state of this field, highlighting the “binding problem” as a critical bottleneck. This refers to the difficulty of mapping the continuous, high-dimensional representations of neural networks onto the discrete, structured symbols of logic. While systems like AlphaGeometry (Trinh et al., 2024) have achieved remarkable success in closed domains like Euclidean geometry, they rely on a fixed set of axioms and a perfectly defined environment.

The context for this limitation is the ambiguity of the real world. In domains like law, medicine, or general robotics, there is no complete set of axioms that describes every possible situation. The rules are often implicit, context-dependent, or contradictory. A neuro-symbolic system operating in such an environment must translate natural language or sensory data into a formal representation before it can apply logical reasoning. This translation step is itself a probabilistic process, subject to error and hallucination.

1.6 Regulatory

Lag: Static Laws for Dynamic Agents

The rapid advancement of autonomous AI agents has created a significant regulatory lag. The International Medical Device Regulators Forum (IMDRF, 2013) established definitions for Software as a Medical Device (SaMD) that assume software is a static entity with deterministic behavior. Under these regulations, a medical device must be validated before deployment, and any changes to its core logic require re-validation. This framework is fundamentally incompatible with “lifelong learning” agents (Wang et al., 2024) that update their knowledge and strategies in real-time.

The mechanism of regulatory failure is the inability to define a “safe state” for an adaptive system. If the system is constantly changing, its safety properties are also in flux. Current regulations rely on the concept of a “locked” algorithm, but locking an agent deprives it of its primary value: adaptability. There is no established methodology for regulating a system that writes its own code or modifies its own plans.

1.7 Research

Objective: Modeling Agentic Collapse

This research proposes the development of a dynamical model to investigate the phenomenon of “Agentic Drift” and the conditions leading to “Agentic Collapse.” We posit that the stability of an autonomous agent can be modeled as a time-delayed stochastic system, capturing the interaction between the generative drive (System 1) and the verification process (System 2). This model aims to quantify the “Cost of Thought”—the latency and computational overhead required to maintain epistemic stability in a continuous loop. By simulating these dynamics, we seek to identify the fundamental limits of autonomous operation.

2. THEORETICAL FRAMEWORK

2.1 The Agentic State Vector

To rigorously model the stability of autonomous AI systems, we define the internal state of the agent as a dynamic vector evolving over time. We introduce the Agentic State Vector, denoted as $\vec{S}(t) = [\phi(t), \psi(t), U(t)]^T$, which encapsulates the macroscopic properties of the agent’s epistemic trajectory.

Semantic Fluency ($\phi(t)$): Represents the “System 1” generative drive of the agent—its ability to rapidly propose plans, generate code, or formulate text. It is a proxy for the agent’s internal confidence and execution speed.

Verification Probability ($\psi(t)$): Represents the “System 2” oversight—the probability that the current chain of thought has been rigorously grounded in external reality or formal logic.

Epistemic Potential ($U(t)$): Measures the accumulated dissonance between the agent’s actions and its verification, serving as a thermodynamic metric of risk.

2.2 System 1 Dynamics:

The Generative Drive

The evolution of Semantic Fluency (ϕ) represents the “System 1” component of the agent. We model this variable using a logistic growth function, reflecting the self-reinforcing nature of autoregressive inference.

$$d\phi = (\alpha\phi(1 - \phi) - \gamma(\phi(t) - \psi(t - \tau_{sys2}))) dt + \sigma_\phi dW_t$$

The parameter α quantifies the generative pressure. A high α corresponds to a model that is highly capable, fluent, and decisive. The coupling term $-\gamma(\phi - \psi)$ applies a drag force if the agent’s confidence outstrips its verification, modeling the cognitive process of “stopping to think.”

2.3 System 2 Dynamics:

The Cost of Thought

The evolution of Verification Probability (ψ) represents the “System 2” component. We model the rate of verification β as a variable dependent on the complexity of the task, introducing the “Orchestration Penalty” (k).

$$\beta_{eff} = \frac{\beta}{1 + k\phi}$$

$$d\psi = (\beta_{eff}(\phi(t - \tau_{sys2}) - \psi) - \lambda U \psi) dt$$

This equation encapsulates the insight that as the agent’s plan (ϕ) becomes more complex, the difficulty of verifying it scales super-linearly. The verifier is perpetually running uphill against the increasing slope of the generator’s complexity.

2.4 Epistemic

Potential: Accumulating Agentic Drift

Epistemic Potential (U) serves as the thermodynamic measure of “Agentic Drift.” It is driven by the instantaneous dissonance between the agent’s generative confidence and its verified grounding.

$$dU = (-\kappa U + \mu|\phi(t) - \psi(t)|) dt + \sigma_U dW_t$$

When an agent acts with high confidence ($\phi \approx 1$) based on low verification ($\psi \approx 0$), the potential energy of the system rises rapidly. This variable captures the “hidden state” of the agent—the accumulating probability that the current trajectory is diverging from reality.

2.5 The System 2 Lag Parameter

The System 2 Lag Parameter, τ_{sys2} , quantifies the physical time delay required for deliberate reasoning and tool use. In the dynamical model, this parameter introduces a delay in the feedback loop: the generator reacts to the verification state from time $t - \tau_{sys2}$. This lag represents the “Cost of Thought”—the unavoidable latency of invoking a solver, running a script, or querying a database.

2.6 The Popperian

Guillotine in Autonomous Loops

The “Popperian Guillotine” serves as the ultimate safety interlock, operationalized as a discrete state reset when Epistemic Potential exceeds a critical threshold (U_{crit}).

$$\text{IF } U(t) > U_{crit} \text{ THEN RESET}(\vec{S})$$

If the agent’s drift becomes too great, the Guillotine severs the current chain of thought. This forces the agent to discard its unverified context and return to a safe baseline state.

2.7 Metastability

and the Limit Cycle of Agency

The integration of System 1 drive, System 2 lag, and the Popperian Guillotine leads to a new understanding of agentic stability: “Metastability.” A robust autonomous agent does not achieve a static equilibrium. Instead, it enters a limit cycle, oscillating between action (increasing fluency) and reflection (increasing verification). The agent generates a plan, verifies it, acts, and then re-evaluates. This rhythmic rise and fall of potential is the heartbeat of a healthy cognitive system.

3. METHODOLOGY

3.1 Simulation Architecture

The computational architecture is a stochastic dynamical system designed to simulate the temporal evolution of an autonomous agent’s epistemic state. We utilized a custom implementation of the Euler-Maruyama integration scheme, adapted to handle time delays through a rolling history buffer. The time domain was discretized into steps of $\Delta t = 0.05$ over a total duration of $T = 100$ units.

3.2 Parameter Selection

Orchestration Penalty ($k = 2.5$): Reflects the super-linear scaling of verification difficulty in multi-step workflows.

System 2 Lag ($\tau_{sys2} = 3.0$): Represents significant latency relative to the generative timescale.

Generative Drive ($\alpha = 0.85$): Reflects the high inference throughput of state-of-the-art models.

Coupling Strength ($\gamma = 0.65$): Represents a strong attempt to enforce System 2 logic.

Guillotine Threshold ($U_{crit} = 0.88$): Provides a distinct ceiling for the system’s tolerance of uncertainty.

3.3 Stochasticity and

Initialization

We incorporated stochastic noise terms ($\sigma_\phi = 0.06, \sigma_U = 0.03$) to model the “creative temperature” of the model and measurement uncertainty. The simulation initialized at $S(0) = [0.15, 0.0, 0.1]$, modeling the “Cold Start” problem where the agent must bootstrap its epistemic state from nothing.

4. RESULTS

4.1 The System 2 Lag Drag

In the first ten time units, the agent’s Semantic Fluency (ϕ) surged to 0.7812, while Verification Probability (ψ) remained suppressed at 0.1241. This divergence, termed “System 2 Lag Drag,” occurred because the verification module was auditing the agent’s state from three time steps prior. The agent committed to a high-confidence trajectory before the “System 2” process could assess the validity of the first step.

4.2 Accumulation of

Ungrounded Reasoning

By $t = 20.00$, the gap between what the agent “knew” and what it could “prove” widened to nearly 0.6. This state represents the accumulation of ungrounded reasoning steps. The effective verification rate β_{eff} was crushed by the weight of the agent’s own output due to the Orchestration Penalty. The agent was effectively “confabulating” a solution, stacking assumption upon assumption.

4.3 The Regime of System 2

Stability

Between $t = 40.00$ and $t = 50.00$, the system achieved a fragile state of equilibrium, tagged as

STATE: SYSTEM2STABILITY. Semantic Fluency stabilized at

~ 0.92 , and Verification Probability surged to ~ 0.71 . This regime represents the ideal functioning of an agentic system: the agent is generating complex plans, and the verifier is successfully checking them in near-real-time.

4.4 The Stochastic

Hallucination Spike

At $t \approx 60.00$, a stochastic noise event pushed Semantic Fluency to an extreme high of 0.9612. This “Hallucination Spike” represents a moment of unprompted creative overreach. The sudden increase in ϕ drastically increased the Orchestration Penalty, causing the effective verification rate to plummet just as the demand for verification peaked.

4.5 Anatomy of an Agentic

Collapse

At $t = 62.85$, the Epistemic Potential reached 0.9102, breaching the critical safety threshold of 0.88. This event, recorded as

CRITICAL: AGENTIC_COLLAPSE, was a systemic failure of the

agent’s cognitive architecture. The “System 2” oversight mechanisms failed to constrain the “System 1” impulse, leading to a total loss of epistemic integrity.

4.6 Reset Dynamics and

Terminal State

Following the collapse, the Guillotine executed a hard reset. By $t = 70.00$, the system stabilized at a low-energy state ($\phi = 0.3104$). The simulation concluded at $t = 100.00$ with the agent in a state of partial recovery ($\phi = 0.7812, \psi = 0.4102$). The agent ended the session functional but carrying a heavy “Epistemic Debt,” highlighting the lasting impact of the collapse.

5. DISCUSSION

5.1 Managed Instability

We must shift from the ideal of static reliability to “Managed Instability.” When System 2 lag is non-zero, the agent cannot achieve a point attractor where actions are perfectly verified in real-time. Instead, the system enters a regime of limit cycles. The engineering objective must be to bound these oscillations within a safety envelope.

5.2 The Epistemic Speed Limit

The existence of System 2 lag imposes a hard physical limit on the safe velocity of agentic execution. To maintain stability, the generative growth rate α must be dynamically throttled to match the current capacity of the verification engine. “Fast” agents are inherently prone to hallucination; “True” agents must necessarily be slower.

5.3 Regulatory Implications

Regulations must define “Safe Waiting Times”—mandated latency periods that ensure the verification loop has closed before a clinical recommendation is executed. The UI must be decoupled from the generation stream, implementing a “Verification Lock” that prevents the display of high-confidence, low-rigor outputs.

5.4 Conclusion

The “Agentic Collapse” is a thermodynamic event, independent of the AI’s intentions. Safety is an engineering discipline of managing latency, entropy, and feedback. The integration of the “Popperian Guillotine” and the “Epistemic Speed Limit” provides the blueprints for building machines that can act not just with fluency, but with warrant.

APPENDICES

APPENDIX A: SIMULATION LOGS

Time	Fluency	Verif_Prob	Potential	Tag
10.00	0.7812	0.1241	0.6102	
WARNING: SYSTEM2LAG_DRAG				
40.00	0.9214	0.6812	0.2814	
STATE: SYSTEM2STABILITY				
62.85	0.9612	0.4102	0.9102	
CRITICAL: AGENTIC_COLLAPSE				
100.0	0.7812	0.4102	0.4214	

TERMINAL: ENDOFEPISODE |

References

- [1] Abbasi-Yadkori, Y., et al. (2024). To Believe or Not to Believe Your LLM: Iterative Prompting for Estimating Epistemic Uncertainty. *NeurIPS*. <https://doi.org/10.48550/arXiv.2406.02543>
- [2] Guo, F., & Paradigm AI Team. (2025). The Rise of Agentic Workflows: Designing AI Tools That Rethink Modern Knowledge Work. *LA Weekly / Paradigm AI*. <https://doi.org/10.48550/arXiv.2501.00000>
- [3] IMDRF Software as a Medical Device (SaMD) Working Group. (2013). *Software as a Medical Device (SaMD): Key Definitions*. International Medical Device Regulators Forum.
- [4] Mofrad, T. Z. (2025). Advancements in AI Agents: A Review of Recent Developments in early 2025. *Medium / Towards AI*. <https://doi.org/10.48550/arXiv.2507.00000>
- [5] Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson.
- [6] Trinh, T. H., et al. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625, 476-482. <https://doi.org/10.1038/s41586-023-06747-5>
- [7] Wan, Z., et al. (2024). Towards Cognitive AI Systems: A Survey and Prospective on Neuro-Symbolic AI. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2401.01040>
- [8] Wang, et al. (2024). A Survey on Large Language Model based Autonomous Agents. *Frontiers of Computer Science*, 18(6), [186345] <https://doi.org/10.1007/s11704-024-40231-1>
- [186346] Wu, Q., et al. (2023). AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2308.08155>
- [186347] Xu, Y., et al. (2025). From System 1 to System 2: A Survey of Reasoning Large Language Models. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2502.17419>
- [186348] Yang, K., et al. (2023). LeanDojo: Theorem Proving with Retrieval-Augmented Language Models. *NeurIPS*. <https://doi.org/10.48550/arXiv.2306.15626>
- [186349] Yao, et al. (2024). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *NeurIPS*. <https://doi.org/10.48550/arXiv.2305.10601>