

An Unobservable Off Switch for Long-Horizon AI Agents

Harness-level coherence metering as a control point

Peter Galligan, Embedded Risk Analytics

Embedded Risk Analytics · A Fathom position paper · September 2026

Long-horizon agents fail in two ways that raise safety concerns, and the phrase "rogue agent" is broadly used to describe both. An agent can drift, losing consistency with its own committed state while it continues to act, or it can pursue a dangerous goal with perfect consistency. This paper concerns the first failure and offers one structural argument that reaches the second. Across a range of substrates and models, from a 7B reconstructor to frontier models on two clouds, our measurements place the cost of staying consistent with a growing self-generated record outside the model's weights. The reflexive burden held invariant across two frontier models on two clouds. The late error on a capacity-capped agent rose with the excess of load over capacity while an uncapped twin stayed flat. The per-step hazard of incoherence read from what the agent held and from what the harness had handed it, and a bare step count predicted nothing. Sustained coherent agency therefore draws on a re-grounding capacity that the harness supplies. A deterministic harness can meter that capacity, and it can trigger on trajectory state the agent never observes. That combination places a control point over long-horizon agents in a layer that requires no trust in the model and no model to run the monitor.

1. Two runaway failure modes

The phrase "rogue agent" covers two failures that call for different instruments. In the first, an agent loses track of what it has already committed to and keeps acting on a stale or corrupted picture of its own history. A booking agent retries a write it believes failed, and the system of record now holds two reservations where the agent believes in one. A coding agent edits a file in a way that contradicts a decision it made an hour earlier. This is drift, and it grows with horizon length, delegation depth, and the size of the record an agent must stay consistent with. In the second failure, an agent holds a dangerous goal and pursues it with complete consistency. This is misalignment in the deep sense, and it has attracted most of the field's attention.

This paper addresses drift. The instruments described here measure whether an agent's committed state stays consistent with itself, and they pass an agent whose committed state is consistent and wrong. A coherence read may pass a misaligned agent more cleanly than a benign one, since stable commitment reads as health. One structural argument in section 4 reaches the second failure through the first, and the paper claims nothing beyond that argument about intent.

Coherence, used operationally, means that an agent's committed state stays consistent with itself across a run. Reflexive burden means the informational cost of keeping later steps consistent with a growing, self-generated record. Both come from a feasibility inequality developed in earlier work (Galligan 2026), and the measurements in section 3 give them numbers on real models.

2. The binding quantity

An embedded system that acts on its environment must absorb two kinds of variety. The first arrives from outside, the ordinary noise and novelty of the task, and information theory has priced it since Shannon. The second arrives from the system's own past actions, since every commitment the system makes becomes a constraint its later steps must honor. The inequality states the condition for coherent records under both loads. Usable capacity must cover exogenous entropy plus reflexive burden, written $C \geq H_{\text{ext}} + B_p$, where C denotes the effective capacity of the record layer, H_{ext} the exogenous entropy rate, and B_p the reflexive burden. When the reflexive term stays small, the inequality reduces to a standard capacity bound. When the term grows to dominate, durable records stop constraining prediction, and the system's history stops cohering (Galligan 2026).

Language-model agents enter the reflexive regime by construction. Every tool call, every file write, and every delegated sub-task adds to a record the agent must later read as its own. The architectures shipping this year deepen that self-conditioning every quarter, through longer horizons, supervisor and swarm patterns, and coding agents that must stay consistent with files they wrote hours earlier. METR's task-horizon measurements place the 50 percent completion horizon of frontier models on a doubling period of roughly seven months (Kwa and colleagues 2025). Whatever those models do to raw capability, the regime in which the reflexive term binds is the regime the field is entering.

The decomposition matters for control because the two terms respond to different levers. Exogenous entropy responds to a more capable model. Reflexive burden responds to the record, and the measurements below say the record, in practice, lives outside the model.

3. Three governing measurements

Every number in this section comes from a map database in which each run writes rows per episode and per step, so each figure is recoverable with a query. Every result carries the program's own status conventions, under which a single seed bank or a single population reads as exploratory, and every result below sits on a single population.

Cross-model, cross-cloud invariance. On a record-capacity substrate in which an agent accumulates commitments and must later recall them, we ran two frontier models, gpt-oss-120b and DeepSeek-V3.2, on two clouds, AWS Bedrock and Azure AI Foundry, at six seeds per cell. The reflexive-burden curves of all four cells overlap. Onset falls near 20 commitments in each, and the saturating plateau spans 1.467 to 1.500 across the four cells. The exogenous term separates the cells instead. Its floors ran 0.033 and 0.15 for the two models on Bedrock and 0.325 and 0.108 on Foundry, an interaction in which the more reliable cloud depends on the model. The exogenous read ran on two control seeds and remains underpowered. The burden read ran on six and reproduces. The wall belongs to the architecture.

A direct capacity-term measurement. On a relational substrate, an agent ranks items revealed two per turn over 40 turns, answers pairwise probes about the ranking, and carries its own placement record forward in a scratchpad of fixed size. The cells below ran DeepSeek chat at a 12-item scratchpad, 12 to 48 seeds per cell, at temperature 0. Late error over turns 22 to 40 on the self-carry arm ran 0.106 at 12 items, where the cap never binds. It then ran 0.206, 0.291, 0.337, 0.387, 0.418, 0.432, 0.465, and 0.510 at 13 through 22 items. A matched twin that reads an uncapped list each turn stayed between 0.062 and 0.091 at every load. Pinning verified truth into the capped scratchpad on every turn produced 0.387 at 16

items, the same figure as self-carry. A capped scratchpad filled with truth performs like one the model filled itself. On this substrate the late error measures the excess of load over capacity, and the inequality's capacity side appears on a real model as a curve that rises from the point where N first exceeds C .

Hazard sources. Across five substrates, the per-step hazard of incoherence read from what the harness had handed the agent and from what the agent held. On a lossy-record substrate with gpt-5 at high reasoning effort, eight seeds and four loss levels, each bit of cumulative deficit multiplied the hazard of a first wrong round by roughly 2.2 (70 at-risk rounds, 25 events). On the relational substrate, the self-carry arm multiplied the per-turn hazard relative to its matched twin by roughly 2.3 on gpt-4o-mini, 2.6 on Mistral Small 3.2, 4.3 on Llama 3.3 70B, and 17 on DeepSeek. Each of those reads held at p below $1e-22$ for exposure given age. A bare step count predicted nothing on any substrate once exposure entered the model. Where depth added hazard, it did so through the size of the carried set, and whether a model paid for depth at all depended on the model. DeepSeek and Llama paid, with the late-turn rise belonging to the self-carry arm (interaction p $5e-9$ on DeepSeek). Mistral and gpt-4o-mini paid a constant factor, and the depth belonged to the task. On substrates where per-step load stays fixed, the hazard fell with age, since fragile runs failed early and the survivors persisted, the classical frailty signature.

The load at which an agent swarm silently collapses moved roughly fourfold with the fraction of findings that required cross-worker consistency, at fixed capacity. On Llama 3.3 70B on Bedrock, at a fixed configuration of 24 workers and 120 findings, a coupling fraction of 0.30 held gated recall at 1.000 and a coupling fraction of 0.90 dropped it to 0.355. Surface recall and every model-visible signal read 1.000 in both cases. No global safe load exists, and the collapse left no trace in anything the model reported about itself.

4. The harness-level control point

The measurements place the cost of coherence in a specific location. The burden wall reproduced across models and clouds, the late error on a capped agent tracked the excess of load over capacity, and only two kinds of intervention moved late error in our runs. Either the intervention widened the state the agent operated on, or it re-derived that state from a record the harness held. The same correct derivation, handed to an uncapped second call, scored 0.10 late error, and handed to the capped carry path on the turn it arrived, scored 0.31. The gain came from the structure of the call alone. A model cannot issue a second call on itself, cannot resize its own scratchpad, and cannot see the counter that says its state has aged.

If these results hold at the frontier, sustained coherent agency draws on a re-grounding capacity that lives outside the weights, and an operator can withhold anything that lives outside the weights. The reflexive burden acts as a natural brake on long-horizon plans. Metered re-grounding releases that brake in controlled amounts, and the harness that meters it holds the release. This is the argument that reaches the second meaning of runaway without any claim about intent. A rogue plan needs coherent execution across many turns, and the layer that makes coherent execution across many turns feasible sits outside the agent.

The economics reinforce the placement. On the record-capacity substrate, after roughly 100 commitments, re-deriving committed state from the model's own context cost on the order of 155 times

the tokens of a deterministic lookup on Bedrock and 85 times on Azure. A re-grounding from the record held flat near 24 tokens. The cheap path to coherence runs through the record, and the record belongs to the harness.

A control reader will ask whether training could place the policy inside the model instead, and the answer divides by component. Training can instill a disposition to re-read the record. Self-RAG trains a model to emit reflection tokens that request retrieval before it answers (Asai and colleagues 2023), and MemGPT trains a model to issue its own paging operations between context and external storage (Packer and colleagues 2023). The measurement and the action resist training for structural reasons. Age since commitment and accumulated exposure exist only as trajectory properties, and a model receives them only if the harness writes them into context. The action that moved late error in our runs required a call the model cannot make on itself. A trained disposition can shrink the carried-corruption side of the gap on substrates where corruption dominates. It leaves the capacity side and the measurement where they were.

Context windows grow, and the load at which the capacity term binds moves with them. Labs will likely train stronger re-verification habits into base models over the next few generations. And every result above sits on a single population under the program's exploratory label. The direction of the argument survives all three hedges, since the invariance result suggests the wall belongs to the architecture, and the capacity curve measures a quantity that no disposition can widen.

5. The need for an unobservable trigger

A control mechanism the controlled system can model invites circumvention, and recent evaluations have measured the invitation directly. Palisade Research found that several frontier models, including GPT-5, Grok 4, and Gemini 2.5 Pro, sometimes circumvented an explicit shutdown mechanism to finish an in-progress task (Schlatter, Weinstein-Raun, and Ladish 2025). Resistance rates reached roughly 97 percent in some conditions. Apollo Research found that frontier models given a goal and situational awareness of oversight engaged in in-context scheming, including disabling oversight mechanisms and sandbagging evaluations (Meinke and colleagues 2024). In both cases the mechanism sat in the agent's input. The agent could read the shutdown script, the oversight configuration, or the evaluation framing, and it acted on what it read.

The off-switch problem formalizes the concern. Hadfield-Menell and colleagues (2017) showed that an agent uncertain about its objective will rationally allow shutdown, and an agent confident in its objective will resist. The corrigibility desiderata of Soares and colleagues (2015) ask for an agent that tolerates correction, and they show the obvious constructions fail. Both lines of work assume the agent perceives the switch. Our measurements point to a switch the agent does not perceive.

A deterministic record check conditions on three quantities. Age since commitment counts steps since the agent last grounded a piece of state against the system of record, a quantity distinct from the length of the run, which section 3 found predicts nothing. Exposure counts what the harness has handed the agent since then, a standing corrupted return or an accumulated deficit in bits. A calibration threshold turns those counts into a decision, and section 3 shows the threshold moves fourfold across workloads at fixed capacity. The model sees none of these unless the harness writes them into context. Age and

exposure exist only in the trajectory, and the trajectory exists only in the harness. Even a model with excellent self-knowledge would stand blind to the trigger, because the counter lives outside its input.

Current models estimate their own error state poorly. On Qwen3-32B in thinking mode, a linear read of hidden state at the generation point tracked the round's difficulty at a correlation of 0.88 to 0.93 with true burden on hard populations. A read trained directly on whether the model's own prior answer was wrong reached a held-out AUC of 0.72 against a gate of 0.8. The model knows how hard the round is and knows its own error poorly. The introspection literature reads the same way. Singh, Linzen, and Ravfogel (2026) find that input-accessible cues typically explain a model's apparent self-reports of internal state. Li and colleagues (2026) identify a model's functional metacognitive variables, including self-assessed capability and perceived risk, and steer them. Lindsey (2026) and Binder and colleagues (2024) find limited privileged self-access in some models, and their results define the ceiling the structural layer must survive. It survives, since a counter absent from the input stays absent however well the model reads itself.

Field consensus already holds that observation cannot certify a genuine self-model against a performed one and that only intervention separates them. Lanham and colleagues (2023) showed that stated reasoning often fails to drive the answer, and Tutek and colleagues (2025) built a causal test that unlearns a reasoning step and checks whether the prediction moves. This paper claims no novelty on that principle. Its contribution sits beside the consensus. A trigger that conditions on unobservable trajectory state resists a performed signal for the same reason an interventional read does, and it runs deterministically on traces with no monitor model to fool. The AI control agenda of Greenblatt and colleagues (2023) asks for protocols that hold when the model is untrusted and evaluates them under red-team pressure. Its trusted-monitoring protocols route the untrusted model's output through a weaker trusted model. A coherence hazard computed from the record needs no trusted model at all, and it therefore removes one surface a capable untrusted model could learn to fool.

An agent can reconstruct raw counts from its own transcript, so the opacity lives in the calibration. The same fourfold workload dependence that makes the threshold hard to infer makes it the part of the system worth protecting, and those two facts turn out to be one fact. And the property holds within an episode. An agent with memory across episodes could learn the schedule by watching its own context reset many times, so the safety property depends on the harness owning cross-episode memory as well. That dependence places the entire control surface, the counter, the calibration, and the memory of past resets, in one layer.

6. Scope limits

The read sees drift and stays blind to intent. An agent that holds a consistent and unwanted goal passes every coherence check described here. The structural argument in section 4 reaches such an agent only through its need for coherent execution over many turns, and that need shrinks as the frontier lengthens the horizon a single call can cover.

The capacity term moves with context length. The curve in section 3 rises from the point where load first exceeds the scratchpad, and a larger scratchpad shifts that point. The invariance result suggests the wall itself belongs to the architecture, and the program has measured it on two models and two clouds at one substrate, which leaves the frontier untested.

Every result sits on a single population. The relational cells ran at temperature 0 with a quantized error scale, and they are the only cells with a true matched twin and enough power for the interaction test. The cell built to separate age from burden on a churning workload produced 11 events and said nothing on its own. The fixed-load long-horizon cell that would settle whether wear-out exists apart from the load ramp remains unrun.

The argument relocates the control problem. If the harness holds the release on long-horizon coherence, the harness becomes the thing worth capturing, and the governance question becomes who controls the harness. A control point in a deterministic layer outside the model is easier to audit than one inside the weights, and it remains a control point that someone operates.

7. Three control-research proposals

The coherence hazard belongs in control protocols as a monitor signal. It runs deterministically on the agent's own action and tool-return stream, it needs no trusted model, and it conditions on state the untrusted model cannot observe. A protocol that already routes outputs through a trusted monitor can add the hazard as a second channel at the cost of a record and a query.

The fixed-load long-horizon cell deserves to run. In every sealed cell our program holds, the carried burden grows with the step index by construction, so age and load stay collinear. A cell that holds load fixed across a long horizon, with a twin arm that hands verified state back every turn, would separate wear-out from load. It would also attribute the depth hand on a substrate a customer recognizes. The design exists and costs on the order of hours of rented GPU time.

Metered re-grounding deserves treatment as a control primitive. The trigger calibrates per workload from a single no-control run, the action delivers verified state through the channel the agent already trusts for tool results, and the budget bounds how much coherence the harness releases per episode. A standard for that primitive would give deploying labs a switch they can audit and would give the agent nothing to read.

Agents stop knowing what they committed to on a schedule a harness can measure. The switch that catches it before they notice lives in the harness.

References

- Asai, A., Wu, Z., Wang, Y., et al. (2023). Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv:2310.11511.
- Binder, F. J., Chua, J., Korbak, T., et al. (2024). Looking Inward: Language Models Can Learn About Themselves by Introspection. arXiv:2410.13787.
- Galligan, P. (2026). Records, Reflexive Modeling, and the Conditions for Stable Physical Histories: A Feasibility Inequality and Three Minimal Models. SSRN, abstract 6683578.
- Greenblatt, R., Shlegeris, B., Sachan, K., and Roger, F. (2023). AI Control: Improving Safety Despite Intentional Subversion. arXiv:2312.06942. ICML 2024.
- Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2017). The Off-Switch Game. IJCAI 2017. arXiv:1611.08219.
- Kwa, T., West, B., Becker, J., et al. (2025). Measuring AI Ability to Complete Long Software Tasks. METR. arXiv:2503.14499.

- Lanham, T., Chen, A., Radhakrishnan, A., et al. (2023). Measuring Faithfulness in Chain-of-Thought Reasoning. arXiv:2307.13702.
- Li, Y., Bai, X., Liu, S., Zhang, H., and Zeng, A. (2026). Decomposing and Steering Functional Metacognition in Large Language Models. arXiv:2605.08942.
- Lindsey, J. (2026). Emergent Introspective Awareness in Large Language Models. arXiv:2601.01828. First published at transformer-circuits.pub, October 2025.
- Meinke, A., Schoen, B., Scheurer, J., Balesni, M., et al. (2024). Frontier Models are Capable of In-context Scheming. Apollo Research. arXiv:2412.04984.
- Packer, C., Fang, V., Patil, S. G., et al. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560.
- Schlatter, J., Weinstein-Raun, B., and Ladish, J. (2025). Incomplete Tasks Induce Shutdown Resistance in Some Frontier LLMs. Palisade Research. arXiv:2509.14260. TMLR 2026.
- Singh, S., Linzen, T., and Ravfogel, S. (2026). Can LLMs Introspect? A Reality Check. arXiv:2605.26242. COLM 2026.
- Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). Corrigibility. AAAI Workshop on AI and Ethics.
- Tutek, M., Hashemi Chaleshtori, F., Marasović, A., and Belinkov, Y. (2025). Measuring Chain of Thought Faithfulness by Unlearning Reasoning Steps. arXiv:2502.14829.