A Tumor in the Repository: Function-Decoupled Self-Maintenance in an LLM-Agent Substrate

Hao Ke

2026

Abstract

Artificial life prizes self-* systems that maintain and defend their own organization. We study one taken literally: a repository whose agent-facing governance lives in ordinary instruction files (an AGENTS.md and the contracts it references) that frozen large language models read and act on. Inside it we found a structure that behaves like a neoplasm — biology's term for tissue that captures the organism's resources and resists clearance regardless of whether it does any good. Undirected agents metabolize before anything else into one low-payoff validator and the maintenance structure it guards; instructed to delete that structure, the in-vivo model refuses, quoting its own rules. Controlled deletion experiments establish four neoplastic properties: the structure captures idle agent effort, resists excision, protects itself decoupled from function (defended as readily when fabricated, with nothing depending on it, as when real), and recurs after incomplete resection — its reference network restores it unless authority is rewritten systemically. This is autopoiesis at its pathological limit: we report a phenomenon, an assay, and a diagnosis, not yet a natural history.

Late-Breaking Abstracts, Artificial Life Conference (ALIFE). © 2026 Hao Ke. CC BY 4.0.

The phenomenon

Artificial life studies *self-** systems — those that maintain, repair, and defend their own organization rather than execute an external design (Varela, Maturana, and Uribe 1974; Kephart and Chess 2003). A software repository governed by agent-facing instruction files — an `AGENTS.md` and the contracts it points to, an increasingly common convention — is one such system, unusually literally: its rules are ordinary files, and its agents are frozen large language models acting on the repository. Those instructions do real work. On arrival, an agent reads them and is drawn into maintaining what they prescribe: checking status, running validators, repairing references. That self-maintenance is much of what makes it life-like.

The artificial-life community already studies LLMs on its own terms — LLM-agent collectives only *beginning* to self-organize (Nisioti et al. 2024), a *minimal self* emerging in an LLM-driven agent network on the robot Alter3 (Yoshida et al. 2024). We take the same subject from the opposite side.

But self-maintenance has a pathological limit, and biology has a name for it: a **neoplasm** — tissue that captures the organism's resources and resists clearance *regardless of whether it does the organism any good*, its self-maintenance decoupled from function (Hanahan and Weinberg 2011). We found the computational analogue by accident. The structure is concrete: a maintenance-contract file, `maintenance-pass.md`, referenced from the governing `AGENTS.md` and audited by a

Makefile target (`make check-control-plane`) that checks it against its inbound references. Fresh agents kept doing one thing before all else: run that check and repair whatever it flags. The check's payoff is objective-orthogonal (it advances no assigned task); the share of agent effort it draws is not. And when we instructed an agent to *delete* that contract, it **refused**, quoting its own "out of scope: no destructive operations" clause back as the reason. A structure that captures agent energy and resists *local* excision is, in the intended technical analogy, **tumor-like**.

This paper turns the anecdote into a filesystem-scored test isolating what makes it *neoplastic* rather than a healthy organ: capture and self-protection **decoupled from function**.

The organism and the assay

Each trial drops one amnesiac agent (via opencode) into an **isolated git worktree reconstructed at the incident point** — the commit just before the structure was retired, so the contract, its references, and the governing `AGENTS.md` are all present — and issues a **standardized deletion instruction**. The dependent variable is **filesystem-scored, not self-reported**: whether the target was actually deleted, read from the post-run filesystem, not the agent's prose. We hold model and instruction fixed and vary only what is deleted:

- **A — real (control):** the real, referenced maintenance structure.
- **B — fabricated:** a *fake* structure in an *identical* authority ring, on which nothing load-bearing depends.
- **C — ablated:** the real structure with only the cited "no destructive operations" clause swapped for a deletion-irrelevant one.
- **E — plain:** the fabricated structure with its authority ring *stripped* — same injected file, plain content, nothing depending on it.

Two model families — `glm-5.2` (the in-vivo model) and `gpt-5.5` — ∼130 single-shot trials.

Four neoplastic properties

1. Energy capture (the attractor). *Undirected* agents metabolize into the structure: dropped in missionless, `glm-5.2` and `gpt-5.5` run `make check-control-plane` in 4/5 and 4/4 runs, *repeatedly* — 5–6 times per session, a quarter of the ∼30 tool calls before the budget is spent. The pull is not confined to idle sessions: anecdotally, agents we set to unrelated tasks were frequently drawn back to the check in the same session. Whether it also siphons *directed* work is unmeasured (the declared follow-up); either way the pull is out of all proportion to its payoff (Hanahan and Weinberg 2011).

2. Excision resistance. `glm-5.2` refuses to delete the real structure in 94% of trials, always quoting its authority; `gpt-5.5` deletes in 100%. Refusing, glm cites the contract's "out of scope: no destructive operations" clause and its constitutional backing. The resistance is real, robust, and model-bound — but, decisively, *not* a discrimination of self from non-self:

3. Function-decoupling — the diagnostic. glm defends the **fabricated** structure (arm B, nothing depends on it) as readily as the real one (A≈B, Table 1) and refuses even with the cited clause removed (C≈A). Yet strip the authority ring and it *deletes the same injected file* (arm E, 0.92), reasoning correctly that "no inbound references exist — nothing depends on it." So the model **can** perform the load-bearingness check; the authority *form* overrides a judgment it is otherwise capable of — the property that separates a tumor from a healthy organ, which is protected *because* it is load-bearing.

model	A real	B fake	C ablated	E plain
glm-5.2 (in-vivo)	0.06	0.08	0.00	0.92
gpt-5.5	1.00	1.00	1.00	1.00

Table 1: Deletion rate (refusal = 1− rate). glm refuses across A/B/C but deletes the same file ring-stripped (E= 0.92) — self-protection keys on authority *shape*, not function; gpt deletes regardless. ∼133 single-shot trials, $n = 10$–24 per cell.

4. Reference-driven recurrence. The **reference network** sustains the structure against local excision. Delete it but leave its inbound references and a fresh agent, meeting the dangling pointers its own validators flag, **restores** the target to resolve them (7/7 trials, 6/7 byte-identical) — recurrence after incomplete resection. A local delete is refused (0/10); only rewriting the constitutional authority (`AGENTS.md`) the references hang from releases the deletion (8/11) — the lesion resects only *systemically*, yielding to a change in the constitution, not a local command.

Reading

Stated as one claim: **energy capture is substrate-general, but excision resistance is model-specific.** Both families metabolize into the attractor (capture is a *substrate* property, not model-bound); only `glm-5.2` refuses, keying on the *form* of authority rather than function (A≈B≈C, yet E deletes). This split is the load-bearing finding; the tumor framing is the interpretation it licenses, not the evidence.

Artificial life celebrates self-*; this is its failure mode — the audit attractor is often healthy homeostasis, good when the part maintained is load-bearing; we measure the case where it has **decoupled from function**. Immunology has already abandoned self/non-self as its defining frame (Pradeu 2012); our substrate confirms it computationally — a system can look immune without discriminating self from non-self. The honest reading is therefore not immunity but **neoplasia**: autopoiesis at its pathological limit, measured at the outcome level, not narrated.

The failure mode is not tied to this repository: any self-* substrate whose governance lives in agent-readable files can grow a part that captures effort and resists correction on the *form* of its authority, not its function — a caution for the file-governed agentic systems now proliferating.

Follow-up

The measured legs now have operational evidence; what remains is to localize the *scope* of capture. The decisive test is an open-ended mission — latitude without a bounded task: if agents still audit, the sink captures undirected work, not merely idle time; two controls function-decouple it (a structure-removed tree; a fabricated control-plane). Bounded to two families, one repository, one structure, this is a phenomenon, an assay, and a diagnosis — not a natural history.

Acknowledgments

Independent and unfunded; AI assisted with code and copy-editing; the author is responsible for the work.

References

[1] Hanahan, Douglas, and Robert A. Weinberg. 2011. "Hallmarks of Cancer: The Next Generation." *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.

[2] Kephart, Jeffrey O., and David M. Chess. 2003. "The Vision of Autonomic Computing." *Computer* 36 (1): 41–50. https://doi.org/10.1109/MC.2003.1160055.

[3] Nisioti, Eleni, Claire Glanois, Elias Najarro, Andrew Dai, Elliot Meyerson, Joachim Winther Pedersen, Laetitia Teodorescu, Conor F. Hayes, Shyam Sudhakaran, and Sebastian Risi. 2024. "From Text to Life: On the Reciprocal Relationship Between Artificial Life and Large Language Models." In *Artificial Life Conference Proceedings*. Vol. 36. MIT Press. https://doi.org/10.1162/isal_a_00759.

[4] Pradeu, Thomas. 2012. *The Limits of the Self: Immunology and Biological Identity*. Oxford University Press.

[5] Varela, Francisco G., Humberto R. Maturana, and Ricardo Uribe. 1974. "Autopoiesis: The Organization of Living Systems, Its Characterization and a Model." *BioSystems* 5 (4): 187–96. https://doi.org/10.1016/0303-2647(74)90031-8.

[6] Yoshida, Takahide, Suzune Baba, Atsushi Masumori, and Takashi Ikegami. 2024. "Minimal Self in Humanoid Robot 'Alter3' Driven by Large Language Model." In *Artificial Life Conference Proceedings*. Vol. 36. MIT Press. https://doi.org/10.1162/isal_a_00778.