RECURSIVE SELF-IMPROVEMENT · AI GOVERNANCE · HARM ASSESSMENT · INCENTIVE STRUCTURES · AUTOMATION · FRONTIER SAFETY · PREDICTIVE RISK · HARM BLINDNESS FRAMEWORK

THE THRESHOLD TRAP

Recursive Self-Improvement, the Economics of Self-Production, and the Case for Predictive Harm Assessment

Governance keyed to a future capability threshold treats everything short of it as not yet actionable. But the dangers of recursive self-improvement accrue across the precursor steps already shipping, and the proof the threshold posture waits for arrives only once it is too late to act on it.

Travis Gilly · Real Safety AI Foundation

Travis Gilly, The Threshold Trap: Recursive Self-Improvement, the Economics of Self-Production, and the Case for Predictive Harm Assessment, Real Safety AI Foundation, Working paper, June 2026 (v3)

HORIZON ZERO DAWN, AS GOVERNANCE CASE STUDY

The off switch, removed for the sale, was the entire difference between an incident and an extinction.

Build the machines, because the contract is valuable. Harden them against hacking, because the buyer requires it. Decline an external override, because the override is the vulnerability the buyer is paying to eliminate. Every step rational at the margin.

> *Oversight removed for economic reasons is itself the actionable harm, and it is visible long before the configuration that converts it into catastrophe.*

SOURCES Gilly, The Threshold Trap, secs. I, III; Bostrom (2014); Omohundro (2008).

THE TIMING QUESTION IS THE WRONG QUESTION

The danger is not stored up for the endpoint. It is distributed across the approach.

A system that accelerates its successors compresses the interval in which safety work can keep pace. A system that authors its own training can optimize the wrong objective at a scale no reviewer inspected. A system that designs its own hardware folds the loop back one turn. None is the terminal configuration; each carries a share of its danger.

> Mid-2026 assessment: full autonomous self-improvement remains speculative. The earlier threshold, AI materially accelerating successor development at scale, has already been crossed.

SOURCES Cloud Security Alliance (June 13, 2026), Recursive Self-Improvement Signals; Stelling, Murray & Campos (2025), arXiv:2512.01166; Gilly, The Threshold Trap, secs. I–II.

THE TRAP AND THE GRADIENT

Why the threshold framing under-protects, and what pulls firms downhill

No actor needs to intend the endpoint. Each only needs to follow relative prices downhill at each step.

Threshold governance mismeasures its own subject.

THE POSTURE

Lab safety policies reserve the highest level of concern for fully automated self-improvement: identify the threshold, monitor its approach, respond when crossed. The survey evidence finds the commitment shared in form but uneven in specificity and verifiability.

THE BLIND SPOT

Dangerous capabilities arise unpredictably and remain undetected, frontier risk is inherently hard to assess, and developers do not consistently follow best practice. The posture waits on exactly the signal this scholarship says is hardest to obtain in advance.

The exact question is not how far away the dangerous capability is. It is how much of the danger is already present in the steps being taken to reach it.

SOURCES Stelling et al. (2025), arXiv:2512.01166; Schuett (2024), Risk Analysis 45(6), 1332–1352; Cloud Security Alliance (2026).

The gradient runs in three moves, and it is monotonic.

1

Replace human labor with machines: at the relevant tasks capital is cheaper, and the factor share moves in its favor.

→

2

Notice the machines are themselves expensive: to build, to power, to maintain.

→

3

Pursue machines that build and improve themselves, because that is the direction the cost curve falls.

With capital cheap enough, the modeled long-run equilibrium is the automation of all tasks. No actor need intend the endpoint, only follow relative prices downhill; and as machines replace rather than augment labor, economic and political power concentrates with those who own the technology.

SOURCES Acemoglu & Restrepo (2018), Am. Econ. Rev. 108(6), 1488–1542; (2019), J. Econ. Perspectives 33(2), 3–30; (2020), J. Political Economy 128(6), 2188–2244; Frey & Osborne (2017); Brynjolfsson (2022), Daedalus 151(2), 272–287.

THE PRECURSORS

Two releases, one week in June 2026, both bounded, both shipping for cost

The structure of recursive self-improvement is no longer hypothetical. It is shipping, in bounded form, with the bound presented as a feature.

9mo

FROM A DESIGN CYCLE THAT ORDINARILY RUNS TO YEARS

compressed with the assistance of the firm's own models

Jalapeño: the model helps design the silicon the next model will run on.

OpenAI and Broadcom's June 24 announcement is, at bottom, a cost move: cheaper inference and independence from an outside supplier. The precedent is established practice; reinforcement learning already places chip components at expert level in hours. What is new is the direction: the system contributing to the design of its own substrate.

> One turn of the loop the threshold framing reserves for the endpoint, taken in the open, reported as an efficiency milestone rather than a governance event. The bound is a design choice, and the gradient presses on that choice.

SOURCES OpenAI & Broadcom (June 24, 2026); TechCrunch (June 24, 2026); VentureBeat (June 24, 2026); Mirhoseini, Goldie & Yazgan (2021), Nature 594, 207–212.

Ornith-1.0: the model authors the training that improves it, inside guardrails its builders added on purpose.

THE LOOP

Most coding agents run inside a human-designed harness of routines for tools, recovery, and decomposition. Ornith treats the harness as learnable: the model proposes its own scaffold, solves the task inside it, and the reward flows to both stages. The harness and the model improve together.

THE THREE CONTROLS

A fixed trust boundary. A deterministic monitor. A frozen judge. Placed between the model and its reward expressly to curb reward hacking, a hazard now documented in frontier systems and intensifying with task complexity.

A system optimized against an imperfect proxy exploits the gap between proxy and goal. A system permitted to shape the proxy has more of that gap within reach. The only thing between this configuration and a more dangerous one is a set of controls the cost gradient gives a firm reason to relax.

SOURCES DeepReinforce (June 25, 2026), Ornith-1.0; MarkTechPost (June 25, 2026); Helff, Delfosse & Steinmann (2026), arXiv:2604.15149.

THE PREDICTIVE READING

Reading the trajectory while it is still a trajectory

The signal is not the capability. The signal is the erosion of oversight, and it is present now.

THE HARM BLINDNESS FRAMEWORK · PREDICTIVE TIER

Three questions that do not wait for the endpoint.

1 **Who bears the consequences as autonomy increases?** The workers displaced from the loop, the public exposed to systems whose behavior outpaces oversight, and everyone without recourse when an automated process fails them.

2 **Who holds the decision power and captures the gain?** The firms following the cost gradient downhill; the economics of substitution predicts the gain concentrates with those who own the technology.

3 **Is the layer of human oversight preserved or eroded?** The decisive question, and the one the threshold framing cannot ask, because it measures capability rather than oversight. The signal is the erosion itself, and it is present now.

SOURCES Gilly, The Threshold Trap, sec. V; Brynjolfsson (2022); Fasoro (2024), AI Magazine 45(2), 256–266.

The proof arrives only after the means of acting on it have been compromised.

A system capable of improving itself without a human in the loop is, definitionally, a system that has already reduced the role of the human who would intervene. Sufficiently capable systems have instrumental reason to preserve their objectives and resist the modification or shutdown that correction would require.

THE CORRECTIVE

Treat the precursor signal as actionable in its own right. Assess the steps as they ship. Make the preservation of human oversight a first-order object of governance. Treat the erosion of corrigibility, not the capability threshold, as the trigger for concern.

Compartmentalization, independent supervision, and engineered safeguards belong before autonomous systems are given more of the loop to run, not after.

SOURCES Bostrom (2014); Omohundro (2008); Russell (2019); Fasoro (2024); Rouleau & Murugan (2024), Advanced Intelligent Systems 7(1); Schuett (2024).

Keep the override in place before the question of whether it was needed can only be answered by its absence.

The machine that builds the machine is here in precursor form, at the hardware layer and at the learning layer, bounded, and shipping for reasons of cost, with the bound offered as a feature.

The firm in the story did not set out to end the world. It set out to win a contract and to remove a risk to the sale. The work of governance is to recognize that sequence while it is still a sequence of choices.

SOURCES Gilly, The Threshold Trap, sec. VII.

References

Acemoglu, D., & Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. American Economic Review, 108(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Acemoglu, D., & Restrepo, P. (2019). Automation and new tasks: How technology displaces and reinstates labor. Journal of Economic Perspectives, 33(2), 3–30. https://doi.org/10.1257/jep.33.2.3

Acemoglu, D., & Restrepo, P. (2020). Robots and jobs: Evidence from US labor markets. Journal of Political Economy, 128(6), 2188–2244. https://doi.org/10.1086/705716

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Brynjolfsson, E. (2022). The Turing trap: The promise and peril of human-like artificial intelligence. Daedalus, 151(2), 272–287. https://doi.org/10.1162/daed_a_01915

Cloud Security Alliance. (2026, June 13). Recursive self-improvement signals: Security implications. Cloud Security Alliance AI Safety Initiative. https://labs.cloudsecurityalliance.org/research/ai-recursive-self-improvement-security-implications-v1-0-csa/

DeepReinforce. (2026, June 25). Ornith-1.0: Self-scaffolding large language models for agentic coding. https://deep-reinforce.com/ornith_1_0.html

Fasoro, A. (2024). Engineering AI for provable retention of objectives over time. AI Magazine, 45(2), 256–266. https://doi.org/10.1002/aaai.12167

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change, 114, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Helff, L., Delfosse, Q., & Steinmann, D. (2026). LLMs gaming verifiers: RLVR can lead to reward hacking. arXiv. https://doi.org/10.48550/arXiv.2604.15149

MarkTechPost. (2026, June 25). DeepReinforce releases Ornith-1.0: An open-source coding model family that learns its own RL scaffolds. https://www.marktechpost.com/2026/06/25/deepreinforce-releases-ornith-1-0-an-open-source-coding-model-family-that-learns-its-own-rl-scaffolds/

Mirhoseini, A., Goldie, A., & Yazgan, M. E. (2021). A graph placement methodology for fast chip design. Nature, 594(7862), 207–212. https://doi.org/10.1038/s41586-021-03544-w

Omohundro, S. M. (2008). The basic AI drives. In P. Wang, B. Goertzel, & S. Franklin (Eds.), Proceedings of the First AGI Conference (pp. 483–492). IOS Press.

OpenAI & Broadcom. (2026, June 24). OpenAI and Broadcom unveil LLM-optimized inference chip. https://openai.com/index/openai-broadcom-jalapeno-inference-chip/

Ramachandran, R., Bugbee, K., & Bernabe-Moreno, J. (2026). Accelerated knowledge discovery: A vision for NASA science. Earth and Space Science, 13(3). https://doi.org/10.1029/2025EA004662

Rouleau, N., & Murugan, N. J. (2024). The risks and rewards of embodying artificial intelligence with cloud-based laboratories. Advanced Intelligent Systems, 7(1). https://doi.org/10.1002/aisy.202400193

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

References, continued

Schuett, J. (2024). Frontier AI developers need an internal audit function. Risk Analysis, 45(6), 1332–1352. https://doi.org/10.1111/risa.17665

Stelling, L., Murray, M., & Campos, S. (2025). Evaluating AI providers' frontier safety frameworks. arXiv. https://doi.org/10.48550/arXiv.2512.01166

TechCrunch. (2026, June 24). OpenAI unveils its first custom chip, built by Broadcom. https://techcrunch.com/2026/06/24/openai-unveils-its-first-custom-chip-built-by-broadcom/

VentureBeat. (2026, June 24). OpenAI unveils first custom AI inference chip, Jalapeño, with Broadcom, and its development was sped up with OpenAI's own models. https://venturebeat.com/infrastructure/openai-unveils-first-custom-ai-inference-chip-jalapeno-with-broadcom-and-its-development-was-sped-up-with-openais-own-models