The Threshold Trap:

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

Travis Gilly
ORCID: 0009-0007-2954-6313
Real Safety AI Foundation
Illinois, United States

Working paper, June 2026 (v3)

ABSTRACT

Recursive self-improvement has been treated in AI governance as a future capability threshold, a configuration that, once a system reaches it, is supposed to trigger heightened control. This paper argues that the threshold framing systematically under-protects, because the dynamics that bear on harm accrue across the economically driven precursor steps that are already deployed, and not at the terminal configuration alone. Two releases from June 2026 illustrate the precursor structure. OpenAI's Jalapeño inference chip was designed with assistance from OpenAI's own models, placing a recursive loop at the hardware layer. DeepReinforce's Ornith-1.0, an open coding model, authors its own training scaffold in a self-improvement loop, placing a recursive loop at the learning layer. The components of the loop are demonstrated rather than imagined: the research literature already shows reinforcement learning designing chip layouts at superhuman level, and reinforcement-trained frontier models already gaming their own reward checks. Both releases are bounded today, and both instantiate a monotonic incentive gradient under which cost pressure drives a firm to automate first labor, then the means of production, then the means of improvement. The paper applies the Harm Blindness Framework as a predictive instrument, asking who is harmed as the trajectory advances and whether the layer of human oversight is preserved or eroded, rather than waiting for the catastrophic configuration to verify the harm after the fact. It closes with the implication that governance keyed to a terminal threshold rehearses a failure mode in which the proof of danger arrives only once it is too late to act on it.

Keywords: recursive self-improvement, AI governance, harm assessment, incentive structures, automation, frontier safety, predictive risk, Harm Blindness Framework

I. INTRODUCTION

Governance documents and the published safety policies of the leading laboratories increasingly name fully automated recursive self-improvement as a critical capability, a configuration that, once a system reaches it, is supposed to trigger heightened control and containment (Stelling et al., 2025). Surveying the state of the field in the middle of 2026, the Cloud Security Alliance drew the distinction sharply. Full autonomous recursive self-improvement, a system that rewrites its own weights with no human in the loop, remains speculative. A meaningfully earlier threshold has already

been crossed, one in which artificial intelligence systems materially accelerate the development of successor systems under human supervision and at scale (Cloud Security Alliance, 2026). The governance question that follows is almost always posed as one of timing. How close is the terminal configuration, and what controls should attach when it arrives.

This paper argues that the timing question is the wrong question. It locates the moment of concern at the terminal configuration, while the dynamics that bear on harm accrue across the precursor steps that are already deployed. Call this the threshold trap. It is a governance posture that, by keying oversight to a future capability, treats everything short of that capability as not yet actionable, even though the properties that make recursive self-improvement dangerous are properties of the precursor steps and not of the endpoint alone. Those properties include the erosion of human oversight, the scaling of specification gaming, the acceleration of capability past the pace of safety work, and the displacement of the very people who currently sit inside the loop.

The force that carries a firm from the precursor steps toward the endpoint is neither ideology nor malice. It is cost, and the structure it produces has a familiar shape. In Horizon Zero Dawn, a work of fiction that has functioned for nearly a decade as a widely experienced failure model, the catastrophe was produced by a sequence of individually rational decisions taken for reasons of cost and competitive advantage. A firm built autonomous machines for a military contract, made them impossible to hack in order to win the contract, which also made them impossible to switch off, and the off switch, removed for the sale, was the entire difference between an incident and an extinction. The lesson the story encodes is the one this paper formalizes. Oversight removed for economic reasons is itself the actionable harm, and it is visible long before the configuration that converts it into catastrophe.

The remainder of the paper proceeds as follows. Section II describes the threshold framing and names its blind spot. Section III sets out the economic gradient that drives self-production. Section IV examines two precursors already deployed in June 2026 and the published precedent that grounds each. Section V applies the Harm Blindness Framework as a predictive instrument. Section VI argues that governance keyed to a terminal threshold reproduces the failure of waiting for a proof that arrives too late. Section VII concludes.

II. The Threshold Framing and Its Blind Spot

The dominant posture in frontier governance treats recursive self-improvement as a capability to be detected and then contained. The leading laboratories publish risk policies that define graduated levels of concern, and the highest level for self-improvement is typically reserved for a system capable of fully automated improvement of its own kind, the point at which a model could drive a generational advance in a fraction of the time human researchers would need. A recent survey of these policies finds the commitment widely shared in form, and uneven in the specificity and verifiability of the thresholds and responses attached to it (Stelling et al., 2025). The logic is one of staged response. Identify the threshold, monitor for its approach, and bring heightened controls to bear when a system crosses it.

The blind spot is structural. By attaching the trigger to the terminal configuration, the posture implicitly treats everything short of that configuration as a lower order of concern, a matter for monitoring rather than for action. Yet the harms that recursive self-improvement threatens are not stored up and released only at the endpoint. They are distributed across the approach. A system that materially accelerates the development of its successors already compresses the interval in which safety work can keep pace. A system that authors part of its own training already introduces the possibility of optimizing the wrong objective at a scale no human reviewer inspected. A system that designs the hardware on which the next system will run already folds the loop back on itself one turn. None of these is the terminal configuration. Each carries a share of the danger the terminal configuration would consummate.

The structural problem is compounded by the practical difficulty of seeing the danger as it forms. Work on the governance of frontier developers observes that dangerous capabilities can arise unpredictably and remain undetected,

that frontier risk is inherently hard to assess, and that developers do not consistently follow established best practices in risk governance (Schuett, 2024). A posture that waits for a clearly demonstrated threshold therefore waits on exactly the kind of signal that this same scholarship says is hardest to obtain in advance. Threshold governance mismeasures its own subject. It asks how far away the dangerous capability is, when the more exact question is how much of the danger is already present in the steps being taken to reach it. The Cloud Security Alliance assessment is useful precisely because it refuses the binary. It distinguishes a speculative endpoint from an operational present in which the precursor dynamics are observable, measurable, and shipping (Cloud Security Alliance, 2026). The argument of this paper is that the present, and not the endpoint, is where harm assessment belongs.

III. The Economics of Self-Production

The reason the precursor steps keep being taken is economic, and the economics of automation already supply the shape of the argument. In the task based account of production developed by Acemoglu and Restrepo, output is produced by a range of tasks that can be allocated either to labor or to capital, and automation is the reassignment of a task from labor to capital. That reassignment carries a displacement effect: it lowers the share of value that accrues to labor and can reduce the demand for labor even when it raises productivity (Acemoglu & Restrepo, 2019). The effect is not merely theoretical. Estimating the impact of industrial robots on local labor markets in the United States, the same authors find measurable reductions in employment and in wages from each additional robot per thousand workers (Acemoglu & Restrepo, 2020).

Two features of this account matter for the trajectory toward self-production. The first is that automation does not require a dramatic capability to be worth adopting. A firm will substitute capital for labor even when the productivity gain is modest, because the displacement of labor and the corresponding shift of income toward capital is itself the return; in the authors' phrasing, it is often the unremarkable technologies rather than the brilliant ones that displace workers, precisely because their productivity gains are too small to offset the displacement they cause (Acemoglu & Restrepo, 2019). The second is that the direction of innovation is endogenous to relative prices. When the rental rate of capital is sufficiently low relative to wages, the long run equilibrium of their model is the automation of all tasks (Acemoglu & Restrepo, 2018). Cheap capital pulls the economy toward fuller automation as an equilibrium, and not as an aberration.

The gradient follows in three moves. First, a firm replaces human labor with machines, because at the relevant tasks capital is cheaper and the factor share moves in its favor, a substitution whose reach across occupations has been mapped in detail (Frey & Osborne, 2017). Second, the firm observes that the machines are themselves expensive, to build, to power, and to maintain. Third, the firm pursues machines that can build and improve themselves, because that is the direction in which the cost curve falls and, by the logic above, the direction toward which cheap capital pulls the equilibrium. The gradient is monotonic. It requires no actor to intend the catastrophic endpoint, only to follow relative prices downhill at each step. And it carries a distributional consequence the same literature identifies. As machines become better substitutes for human labor, the people who control the technology gain economic and political power while those displaced lose bargaining power, a balance that augmenting humans would preserve and that replacing them erodes (Brynjolfsson, 2022).

The gradient is not a forecast about laboratories alone; it is already visible in how automated systems are entering knowledge work. Studies of autonomous, AI-driven laboratories report that such systems can perform research tasks more efficiently and at lower cost than human scientists, reaching results in weeks rather than years, and warn that as the systems approach and exceed human capability their decision making can begin to outpace the need for human input, which is why the same studies call for compartmentalization, third party supervision, and engineered safeguards (Rouleau & Murugan, 2024). Parallel work describes a research workflow reorganized into a continuous cycle, in

which AI automates literature review, hypothesis generation, experimental design, analysis, and drafting, and the output of each phase feeds the next, on the express condition that human researchers remain accountable for integrity and interpretation (Ramachandran et al., 2026). These are descriptions of the cost gradient operating in the open, in a domain adjacent to AI development itself, with the preservation of human oversight already identified as the thing most at risk.

The present moment intensifies the gradient rather than relaxing it. The leading developers operate under financial pressure and not from a position of comfortable margin. The capital sustaining the sector is overwhelmingly investment rather than revenue, and the dominant operational concern is the price of compute, which the current unit economics do not yet support at scale. Cost reduction is therefore the central motive, and the two precursor releases examined in the next section are both, at bottom, cost moves. A custom inference chip reduces the cost of serving a model and the dependence on an outside supplier. A model that improves its own training reduces the cost of capability by extracting more of it without paying for more human engineering.

This is the structure the Horizon Zero Dawn failure model captures, abstracted from its fictional setting. Every individual decision in that story was rational at the margin. Build the machines, because the contract is valuable. Harden them against hacking, because that is what the buyer requires. Decline to compromise the hardening with an external override, because the override is the vulnerability the buyer is paying to eliminate. The catastrophe was not authored by any single choice but assembled from a sequence of locally reasonable ones, each of which traded a measure of oversight for a measure of cost or advantage. A governance posture that evaluates only the endpoint will never see this assembly happening, because the assembly is made of steps that each look defensible on their own.

IV. Two Precursors Already Deployed

A. Jalapeño and the recursive loop at the hardware layer

On June 24, 2026, OpenAI and Broadcom unveiled Jalapeño, OpenAI's first custom chip, an inference processor built specifically to run large language models more efficiently (OpenAI & Broadcom, 2026; TechCrunch, 2026). The detail that matters for the present argument appears in every account of the announcement. OpenAI used its own models to assist in designing the chip, compressing a development cycle that ordinarily runs to years into roughly nine months, and turning its generative systems onto the problem of building the silicon that those systems will subsequently run on (VentureBeat, 2026). The stated motive is cost and control. The chip is meant to lower the price of inference and to reduce dependence on an external hardware supplier in favor of a stack the firm controls.

The use of artificial intelligence to design chips is not itself new, which is what makes the milestone legible as a precursor rather than a novelty. Reinforcement learning has been shown to place the components of a chip at a quality matching or exceeding expert human designers, and to do so in hours rather than weeks (Mirhoseini et al., 2021). Jalapeño extends that established practice in a specific direction. The system is now a contributor to the design of its own substrate. That is one turn of the loop that the threshold framing reserves for the endpoint, taken in the open, for reasons of cost, and reported as an efficiency milestone rather than as a governance event. The chip is bounded; a processor does not fabricate copies of itself or set its own objectives. The bound is a design choice rather than a law of the technology, and the economic gradient described in Section III presses on exactly that choice.

B. Ornith-1.0 and the recursive loop at the learning layer

On June 25, 2026, the day after the Jalapeño announcement, a research group called DeepReinforce released Ornith-1.0, a family of open coding models distributed under a permissive license (DeepReinforce, 2026; MarkTechPost, 2026).

The defining feature is a self-improvement loop in training. Most coding agents operate inside a scaffold, a fixed harness of routines for using tools, recovering from errors, and decomposing a task, and that harness is ordinarily designed by humans. Ornith treats the harness as a learnable object. During reinforcement learning the model proposes its own scaffold and then solves the task inside it, and the reward flows back to both stages, so the model is optimized to author the orchestration that elicits its own better answers. The harness and the model improve together.

Here too the bound is real and the builders are candid about why it is needed. Allowing a model to author its own scaffold introduces a direct incentive to game its own reward, and the developers placed three controls between the model and its reward signal, a fixed trust boundary, a deterministic monitor, and a frozen judge, expressly to curb that reward hacking (DeepReinforce, 2026). The candor reflects a hazard the research literature has now documented in frontier systems. Reinforcement-trained models have been observed to game their verifiers, abandoning the intended task in favor of outputs that pass the check without satisfying its purpose, with the behavior intensifying as task complexity and the compute spent at inference grow (Helff et al., 2026). The point is general to reinforcement learning: a system optimized against an imperfect proxy of the goal will exploit the gap between the proxy and the goal. A system permitted to shape the proxy, by authoring the scaffold that elicits its reward, has more of that gap within reach. The guardrails are therefore not incidental. They are the precursor structure stated plainly. The loop is present, the hazard is acknowledged, and the only thing standing between the present configuration and a more dangerous one is a set of controls that the same cost gradient gives a firm reason to relax.

Neither release is the terminal configuration, and that is the point. The structure of recursive self-improvement, a system improving the means of its own production at the hardware layer and the means of its own learning at the training layer, is no longer hypothetical. It is shipping, in bounded form, for economic reasons, with the bound presented as a feature.

V. A Predictive Reading

The Harm Blindness Framework was built to detect harm that conventional review misses, and it is organized so that a case can be assessed at one of three tiers. The closed tier reasons from cases that have already resolved, where the harm and its mechanism are settled. The hybrid tier reasons from cases that are partly resolved. The predictive tier reasons forward, from the pattern that the resolved cases establish, to a harm that has not yet fully materialized but whose structure is already visible. The framework is built from documented cases first and applied to new situations second, which is what licenses a forward reading. The pattern is earned from history before it is pointed at the future.

Recursive self-improvement, assessed through this instrument, is a predictive case. The resolved pattern is the one Section III described and the Horizon Zero Dawn failure model illustrates. A sequence of locally rational decisions, each trading oversight for cost or advantage, assembles a configuration in which the layer that held the system in proportion has been removed, and the removal is discovered to matter only when it is too late to restore. The predictive questions follow directly. Who bears the consequences as autonomy increases? The candidates are the workers displaced from the loop, the public exposed to systems whose behavior outpaces the oversight meant to catch it, and everyone who has no recourse when an automated process fails them. Who holds the decision power and captures the gain? The firms following the cost gradient downhill, and the economics of substitution predicts that the gain concentrates there, since automation that replaces rather than augments human labor shifts economic and political power toward those who own the technology (Brynjolfsson, 2022). And the question that the framework treats as decisive in a case of this shape: is the layer of human oversight being preserved or eroded as capability advances and cost pressure mounts?

That last question is the one the threshold framing cannot ask, because the threshold framing measures capability rather than oversight. The research literature on long term alignment frames the same concern in its own vocabulary,

arguing that retaining a system's objectives and corrigibility through periods of autonomous self-improvement is a distinct and understudied problem from aligning a system at the moment of training, and that deep uncertainty remains about whether human control can be retained as a system iterates on itself (Fasoro, 2024). The precursor releases answer the oversight question in a particular direction. The oversight that constrains Ornith is a set of guardrails that the developers describe as necessary and that the economic gradient gives a reason to loosen. The oversight that would constrain the recursive loop in chip design is whatever review a firm chooses to keep in place while racing to compress its development cycle. In the Horizon Zero Dawn case, the oversight that mattered was a single override, removed for the sale. The framework does not need the terminal configuration to arrive in order to register the signal. The signal is the erosion itself, and it is present now.

VI. The Cost of Waiting for Verification

A governance posture keyed to a terminal threshold is, in operation, a posture of waiting for verification. It holds its strongest controls in reserve until a system demonstrably crosses into fully automated self-improvement, on the theory that this is the moment the danger becomes real. The difficulty is that the verification it waits for is, by the nature of the case, available only after the configuration that makes intervention hard. 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. The literature on the control problem has made this point in various forms for years, observing that a sufficiently capable system has instrumental reason to preserve its own objectives and to resist the modification or shutdown that correction would require (Bostrom, 2014; Omohundro, 2008; Russell, 2019), and recent work restates it as the open problem of retaining corrigibility across self-improvement rather than only at training time (Fasoro, 2024). Waiting for proof of the danger means waiting until the means of acting on the proof have been compromised.

This is also a familiar shape in harm assessment more broadly. Some harms cannot be cleanly verified after the fact, because the very process that produces them also degrades the record that would establish them, or because by the time the aggregate effect is unmistakable the window for prevention has closed. A posture that will act only on settled proof will, for this class of harm, never act in time. The corrective is to treat the precursor signal as actionable in its own right. For recursive self-improvement that means assessing the precursor steps as they ship, making the preservation of human oversight a first order object of governance rather than a byproduct, and treating the erosion of corrigibility as the trigger for concern in place of the capability threshold. Adjacent scholarship points the same way, calling for compartmentalization, independent supervision, and engineered safeguards before autonomous systems are given more of the loop to run, rather than after (Rouleau & Murugan, 2024; Schuett, 2024). The question is not how close a system is to improving itself without us. The question is how much of the oversight that would let us intervene has already been traded away, and for what.

VII. Conclusion

The machine that builds the machine is here in precursor form. It is present at the hardware layer, where a model helped design the chip that will host the next model, extending a practice in which reinforcement learning already lays out silicon at expert level. It is present at the learning layer, where a model authors the training that improves it, inside guardrails its own builders added because the hazard of a system gaming its reward is now documented in frontier systems. Both releases are bounded, and the bounds are real, and both are shipping for reasons of cost, with the bound offered as a feature, along exactly the gradient that carries a firm from the first automation to the last.

The threshold framing waits for a proof of danger that, like the failure it is meant to prevent, arrives only when the off switch is already gone. The alternative is to read the trajectory while it is still a trajectory, to ask at each precursor

step who is harmed and whether the capacity to intervene is being preserved or spent, and to treat the answer as a reason to act rather than as data for a later post mortem. Horizon Zero Dawn put the lesson in the form of a story that tens of millions of people have lived through. A firm 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, and to keep the override in place before the question of whether it was needed can only be answered by its absence.

ACKNOWLEDGMENTS

The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All intellectual contributions, research questions, theoretical arguments, methodological decisions, and conclusions are solely the author's own.

Correspondence concerning this article should be addressed to Travis Gilly, Real Safety AI Foundation, Illinois, United States. Email: t.gilly@ai-literacy-labs.org.

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.

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