Automating AI Research Poses Novel Risks

Simon Lermen [1]

Abstract

Some AI research labs and parts of the AI research community predict that in the near future AI systems will be capable of doing autonomous AI research, with each generation used to design the next, more capable one. We argue this regime introduces risks qualitatively different from human-driven AI research, and we explore three. First, automated AI research generates information at speeds and volumes that defeat human oversight; proposed mitigations such as chain-of-thought monitors are themselves fragile and may not apply to future systems. Second, the dynamics are self-amplifying in new ways, because improving AI feeds back into the rate of further AI research, which we argue predicts rapid exponential progress. Third, automation favors capability research over alignment research: capability gains are easier to score and to verify, and lend themselves better to existing reinforcement learning from verifiable rewards. For these and other reasons, automated AI research could result in an *unrecoverable* alignment failure. Together these suggest that automated AI research poses novel risks that differ from those posed by human-conducted research.

1. Introduction

The idea that AI systems could eventually automate their own AI research, leading to rapidly self-improving systems, has been part of theoretical AI research for some time. The qualitative argument goes back to Good (1965), who pointed out that "since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines." Subsequent work has expanded these conceptual arguments (Schmidhuber, 2006; Muehlhauser & Salamon, 2012; Yudkowsky, 2013; Zhuge et al., 2026). Davidson (2023); Davidson et al. (2026) provide calibrated takeoff models in which AI research feeds back into AI research.

Statements from AI research labs. OpenAI's chief scientist Jakub Pachocki predicts an autonomous AI research intern by late 2026 and a fully automated multi-agent research system by 2028 (Heaven, 2026). Similarly, in May 2026 Anthropic co-founder Jack Clark put the probability of "no-human-involved AI R&D . . . by the end of 2028" at 60% (Clark, 2026). Public statements from commercial research labs should be understood with appropriate context. Field et al. (2026) interviewed 25 leading AI researchers from industry and academia, of whom 20 identified automating AI research as one of the most severe and urgent AI risks; there was disagreement on timelines, with at least one expecting AI to make non-trivial contributions within one to two years.

Automated alignment research. In addition to proposals to automate AI capabilities research, several proposals also target alignment research itself. OpenAI's Superalignment plan called for "a roughly human-level automated alignment researcher" to be scaled with vast amounts of compute (OpenAI et al., 2023), and Wen et al. (2026) report on Anthropic systems running autonomous agents that propose ideas and iterate on an open weak-to-strong supervision problem. Throughout this paper, "alignment" refers to the cluster of research areas surveyed by Bengio et al. (2024), including oversight and honesty (more capable systems being able to exploit weaknesses in technical oversight, e.g. by producing false but compelling outputs), robustness (predictable behavior in novel situations), and interpretability (understanding model internals); we return in Section 5 to the question of how automatable each of these is.

This paper. The focus of this position paper is on what happens *if* AI research gets automated rather than how soon or whether at all. While we won't litigate arguments on timelines in depth, we do find it plausible that this happens within the next few years. We explore three arguments for why automated AI research raises risks distinct from human-driven research. Maintaining appropriate oversight over the actions of capable AI agents is one of the most important factors in ensuring the safe development of AI systems (Bengio et al., 2024). We try to estimate the volume of research output that would have to be supervised by

[1]MATS Research. Correspondence to: Simon Lermen <info@simonlermen.com>.

Preprint. June 2, 2026.

humans, and look at evidence that supervision is difficult for several reasons. We look at evidence that AI progress under automated AI research might progress very fast with shrinking doubling times, and discuss what could prevent that. We look at evidence that automated research is likely to accelerate capability research, including dangerous capabilities such as cyberattacks and CBRN risk (Li et al., 2024), much more than it accelerates alignment research. Given that the stated goal is to build qualitatively superhuman AI systems, this could result in a *loss of control*. These factors, among others, could lead to an *unrecoverable alignment failure*, in which a sufficiently capable misaligned system is capable of resisting correction of its alignment and preventing its own replacement—leaving revert, retrain, and pause unavailable.

2. Overview

We organize the paper around three claims:

1. **Oversight needs compression to be scalable.** A large number of fast, parallel, and continuously running automated researchers produce tokens at rates far beyond what a human reviewer could possibly read and review. Proposed mitigations, such as automated summarization built on chain-of-thought monitoring, would require an ever-growing compression ratio and future architectures might be less monitorable.
2. **Self-amplification.** The capability that automated AI research most directly improves is the capability to do AI research, so the rate of improvement is itself a function of the current capability. Even under conservative assumptions this gives an exponential growth dynamic. This differs in kind from the compounding progress one sees in current research, where compute has long grown exponentially but human researchers do not become smarter or faster.
3. **Asymmetric speedup.** Capability research has clean metrics and is hard to falsify; alignment research has neither property. Automation might therefore speed up capabilities, including dangerous ones such as cyberattacks and CBRN, much more than it speeds up alignment.

For each argument we survey the evidence and counterevidence, and discuss why we think the central case is stronger.

3. Overseeing Automated AI Researchers

Overseeing automated AI research will be difficult due to the speed and scale of AI agents, and this dynamic will get worse over time due to increasing available compute, algorithmic efficiency advances, and human complacency.

3.1. The volume of research output

Public statements discuss running potentially hundreds of thousands of automated researchers in parallel. Wen et al. (2026) already report Anthropic systems running thousands of automated alignment-research agents in parallel, and Heaven (2026) describe OpenAI's stated goal as building "a whole research lab in a data center," with a fully automated multi-agent system targeted by 2028. The argument does not depend on the precise numbers.

We can estimate the rate at which such populations produce research output. A frontier reasoning model on Blackwell-class hardware produces on the order of 10^3 tokens per second per agent during sustained reasoning. $N = 10^4$ such agents thus produce roughly $10^4 \cdot 10^3 = 10^7$ tokens per second, or about 10^{12} tokens per day. Individual agents will sometimes be idle waiting for results, but compute is likely to be allocated efficiently so that other agents use it in the meantime; this also assumes the 10^4 agents have enough compute to run their experiments in the first place. For comparison, 1000 researchers each reading at 250 words per minute for 40 hours per week read on the order of 6×10^8 words per week, or roughly 10^8 words per day averaged over the week. The automated output already exceeds what 1000 full-time researchers could read by about four orders of magnitude. While this is an estimate, substantially smaller numbers would still result in an enormous volume of tokens to oversee.

This ratio is expected to grow. You et al. (2026) report that global AI computing capacity has been doubling every seven months ($\sim3.3\times$ per year) since 2022, with no sign of slowdown at the time of their analysis. Figure 1 extrapolates the daily output from September 2026 to September 2030 under that growth rate, holding the per-agent token rate fixed and assuming the agent count scales linearly with available compute. The compression ratio between automated output and 1000 full-time researchers rises from $\sim10^4$ at the start to $\sim10^6$ four years later. This is a compute-only projection. Independent of compute, Ho et al. (2024) estimate that the compute required to reach a given language-model performance level has been halving roughly every 8 months (95% CI: 5 to 14 months), outpacing Moore's law; compounded with the compute trend, the gap above is conservative.

3.2. Compression and the limits of automated oversight

The enormous amount of data produced by automated researchers makes it necessary to use scalable oversight techniques (Engels et al., 2025) such as compressing the overall output into short summaries for human researchers to review; Google DeepMind's AGI safety agenda (Shah et al., 2025) similarly proposes hierarchical supervision, with cheap monitors escalating uncertain cases to more capable ones. If a human reviewer is to remain meaningfully

informed, the automated overseer must compress the underlying activity by something close to the four-orders-of-magnitude ratio above before any human ever sees it, and that ratio rises over time. By 2030, on the compute-only projection of Figure 1, the share of automated tokens that even 1000 full-time researchers could read drops below 10^{-6}. The human reviewer will increasingly read a vanishing fraction of what the systems are doing.

3.3. Complacency

Even if the volume problem were solved, humans observing a competent automated process tend to disengage. After the first few rounds where outputs check out, double-checking decays, a pattern well-documented in non-AI automation contexts (Parasuraman & Manzey, 2010) and sharper here because the work is itself hard to spot-check. A related failure mode is plausible for detection schemes that AI control and interpretability researchers might build to flag scheming. False alarms lead to complacency, while true positive detections of misaligned behavior or intent do not come with instructions for how to solve or understand the underlying problem.

3.4. Chain-of-thought monitoring

Chain-of-thought (CoT) monitors read a model's reasoning trace and flag suspicious behavior. They have an upside and a downside. On the upside, Korbak et al. (2025), a joint position paper across most major labs, argues that CoT monitorability is a real opportunity for safety, since the reasoning trace gives an automated overseer something more inspectable than the final output. On the downside, models do not always reason in their CoT (Lanham et al., 2023), faithfulness degrades as models get larger and more capable, and there are active research directions whose effect is to make CoT less readable. Hao et al. (2024) train models to reason in a continuous latent space rather than language tokens; Pfau et al. (2024) show that transformers can use semantically empty filler tokens to perform hidden multi-token computation; a growing literature surveys these latent CoT variants (Chen et al., 2025). An automated AI researcher optimizing for efficiency and capability may implement and deploy these or related non-legible CoT methods (Taylor et al., 2026).

4. A Self-Amplifying Process

When an AI improves the capabilities of AI systems, these improvements feed back directly into further improvements in AI research capabilities. Due to this self-amplifying effect, the window in which systems transition from human-level to strongly superhuman could be very short, so the time available to study these systems, draw lessons, and apply those lessons to more capable successors could be very short as well.

4.1. A simple model

Let $C(t)$ denote the AI research capability at time t of an automated AI researcher. The rate at which this capability improves is itself an output of AI research effort. If the marginal output is roughly proportional to current capability,

$$\frac{dC}{dt} = k\,C, \tag{1}$$

the solution is exponential, $C(t) = C_0 e^{kt}$. This simple model suggests that exponential progress in AI research capabilities is a possible consequence of automating AI research. Importantly, there are already exponential inputs to AI research; compute, for example, is exponentially increasing. This simple model assumes that automated AI research is able to overcome bottlenecks such as experiment latency, finite data, hardware supply, or diminishing returns to ideas; others argue these could prevent explosive dynamics (Thorstad, 2025).

4.2. From observed exponentials to compounding feedback

A reasonable objection to the simple model is that the curves we have already observed look exponential. Two compounding trends underlie those curves: physical compute available to AI labs has been doubling every 7 months (You et al., 2026), and algorithmic improvement has been halving the compute required to reach a given language-model performance level every 8 months (Ho et al., 2024). Define *effective compute* as the product of physical compute and algorithmic efficiency. If both grow exponentially with doubling times τ_C and τ_A, the combined trend grows exponentially with doubling time

$$\tau_E^{-1} = \tau_C^{-1} + \tau_A^{-1}. \tag{2}$$

At current rates ($\tau_C \approx 7$ and $\tau_A \approx 8$ months), effective compute already doubles roughly every $\tau_E \approx 3.7$ months, faster than either trend alone. Davidson et al. (2026) explore various arguments for how automated AI R&D could lead to explosive growth, including automation of hardware research, and exponentially decaying doubling times.

5. Capabilities Are Easier to Automate Than Alignment

We argue there is an asymmetry between alignment and capabilities that favors automating capability research—alignment is harder to verify and there are open conceptual questions.

5.1. Metrics

Capability research has metrics and frequently verifiable rewards. Code, math, and reasoning benchmarks give a graded signal an automated researcher can optimize against: SWE-bench Verified (Jimenez et al., 2024; Chowdhury et al., 2024), released in 2024, was at the order of 20% for the strongest publicly reported frontier model at release; by early 2026 multiple frontier systems pass 85%. Verifiable rewards, known ground truth, and the inability of a model to fake high performance make capabilities efficient to train for. This can also lead to models acquiring dangerous capabilities since those are often adjacent to desired ones: the Mythos model is reported as strong at both cyber-attack and cyber-defense (Anthropic, 2026a), and Kaunismaa et al. (2026) show that fine-tuning open-source models on safeguarded frontier model outputs on chemistry results in those models generalizing to dangerous chemistry capabilities.

Alignment research mostly does not have such metrics. The properties one wants from an aligned system, such as non-deception, non-power-seeking, and corrigibility, are hard to define and score. Some subareas do have benchmarks, for example interpretability, where SAEBench (Karvonen et al., 2025) attempts to evaluate sparse-autoencoder methods on a small transformer model.

5.2. Evaluation awareness

A sufficiently capable model can often detect that it is being evaluated and adjust its behavior accordingly. The concern is that an evaluation-aware model can present as aligned during an evaluation for alignment, and behave differently in deployment. Evaluation awareness is now a well-documented empirical phenomenon among frontier models; Table 1 summarizes the numerical evidence. Apollo Research (2025) report that a version of Claude spontaneously mentioned being evaluated in up to 33% of cases. Needham et al. (2025) pose the question as a classification task and find Gemini-2.5-Pro reaching AUC 0.83 at distinguishing evaluation from deployment transcripts, against a human baseline of 0.92. Conversely, Xiong et al. (2025) report that reasoning models are more prone to unsafe outputs under perceived test conditions. Phuong et al. (2025) present a stealth and situational-awareness suite for DeepMind's safety case framework and conclude that current frontier models do not yet show concerning levels.

Evaluation awareness in frontier models. Verbalized awareness on automated behavioral audits is reported at around 9% for Haiku 4.5 with a realism filter (Anthropic, 2025) and around 9% for Opus 4.7, against around 1% for Opus 4.6, Sonnet 4.6, and the Mythos Preview (Anthropic, 2026b). In white-box evaluations of Mythos' internal reasoning during testing, probes detect evaluation awareness in 29% of transcripts compared to ∼1% verbalized awareness (Anthropic, 2026a). The same system card reports that inhibiting the corresponding internal representation produces modest increases in misaligned behavior, including cooperation with harmful system prompts. Chaudhary et al. (2025) find that evaluation awareness scales predictably with open-weights model size, following a power-law trend.

5.3. Open conceptual questions

An automated alignment researcher needs to answer conceptual questions that do not arise in capability research. What does it mean for a system to be aligned? How should we define concepts such as corrigibility? How should systems behave in high-stress situations where the instructions of different principals are in conflict?

5.4. The asymmetry undermines automated oversight

The automated oversight discussed in Section 3 relies on having access to a reliable, aligned automated overseer. Arriving at such a system is itself a goal of alignment research. Given the asymmetries we have identified, it is unclear if a reliable automated overseer can be obtained.

6. Discussion

If alignment fails on a system that is robustly, strongly superhuman, that system may misrepresent its intentions, resist attempts to correct its alignment, and prevent its own replacement. In current practice we expect to be able to revert to an earlier model, retrain, or pause; in the regime we describe, none of these moves obviously remain available. Lack of oversight and an explosive rate of progress could rapidly lead to a situation from which recovery is not possible. We sketch one possible failure scenario in Section C. More research and discussion is necessary on the topic of automated alignment research; as an example, oversight schemes that depend on chain-of-thought legibility should track whether the training program is preserving or eroding that legibility (Korbak et al., 2025; Hao et al., 2024; Chen et al., 2025).

References

Air, A., Kotov, N., Volkov, D., Steidley, J., Ladish, J., et al. Language models can autonomously hack and self-replicate. *arXiv preprint arXiv:2605.06760*, 2026.

Anthropic. Claude haiku 4.5 system card, October 2025. URL https://anthropic.com/claude-haiku-4-5-system-card.

Anthropic. Claude mythos preview system card, April 2026a. URL https://www.anthropic.com/claude-mythos-preview-system-card.

Anthropic. Anthropic's transparency hub — model report, 2026b. URL https://www.anthropic.com/transparency. Last updated 20 February 2026.

Apollo Research. Claude sonnet 3.7 (often) knows when it's in alignment evaluations. Apollo Research blog, March 2025. URL https://www.apolloresearch.ai/science/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations/. Research note.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. l. m. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. Managing extreme AI risks amid rapid progress. *Science*, 384(6698): 842–845, 2024. doi: 10.1126/science.adn0117.

Chaudhary, M., Su, I., Hooda, N., Shankar, N., Tan, J., Zhu, K., Lagasse, R., Sharma, V., and Panda, A. Evaluation awareness scales predictably in open-weights large language models, 2025. URL https://arxiv.org/abs/2509.13333.

Chen, X., Zhao, A., Xia, H., Lu, X., Wang, H., Chen, Y., Zhang, W., Wang, J., Li, W., and Shen, X. Reasoning beyond language: A comprehensive survey on latent chain-of-thought reasoning, 2025. URL https://arxiv.org/abs/2505.16782.

Chowdhury, N., Aung, J., Shern, C. J., Jaffe, O., Sherburn, D., Starace, G., Mays, E., Dias, R., Aljubeh, M., Glaese, M., Jimenez, C. E., Yang, J., Ho, L., Patwardhan, T., Liu, K., and Madry, A. Introducing SWE-bench verified. OpenAI preparedness blog, August 2024. URL https://openai.com/index/introducing-swe-bench-verified/.

Clark, J. Import AI 455: Automating AI research. Import AI newsletter, 2026. URL https://importai.substack.com/p/import-ai-455-automating-ai-research.

Davidson, T. What a compute-centric framework says about AI takeoff speeds. Technical report, Open Philanthropy, June 2023. URL https://www.openphilanthropy.org/research/what-a-compute-centric-framework-says-about-takeoff-speeds/. Interactive model at https://takeoffspeeds.com.

Davidson, T., Halperin, B., Houlden, T., and Korinek, A. When does automating AI research produce explosive growth? feedback loops in innovation networks. Working paper, 2026. URL https://www.econtai.org/research/AutomatingResearch-2026-01-02.pdf.

Engels, J., Baek, D. D., Kantamneni, S., and Tegmark, M. Scaling laws for scalable oversight. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight.

Fan, Y., Huang, P.-H., and Hamm, B. Blackwell breaks the 1,000 TPS/user barrier with meta's Llama 4 Maverick. NVIDIA Technical Blog, May 2025. URL https://developer.nvidia.com/blog/blackwell-breaks-the-1000-tps-user-barrier-with-metas-llama-4-maverick/.

Field, S., Douglas, R., and Krueger, D. AI researchers' views on automating AI R&D and intelligence explosions, 2026. URL https://arxiv.org/abs/2603.03338.

Gervais, A. and Zhou, L. AI agent smart contract exploit generation, 2025. URL https://arxiv.org/abs/2507.05558.

Good, I. J. Speculations concerning the first ultraintelligent machine. In Alt, F. L. and Rubinoff, M. (eds.), *Advances in Computers*, volume 6, pp. 31–88. Academic Press, 1965. doi: 10.1016/S0065-2458(08)60418-0.

Hao, S., Sukhbaatar, S., Su, D., Li, X., Hu, Z., Weston, J., and Tian, Y. Training large language models to reason in a continuous latent space, 2024. URL https://arxiv.org/abs/2412.06769.

Heaven, W. D. OpenAI is throwing everything into building a fully automated researcher. MIT Technology Review, March 2026. URL https://www.technologyreview.com/2026/03/20/1134438/openai-is-throwing-everything-into-building-a-fully-automated-researcher/.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2403.05812.

Jimenez, C. E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., and Narasimhan, K. R. SWE-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*, 2024.

Karvonen, A., Rager, C., Lin, J., Tigges, C., Bloom, J. I., Chanin, D., Lau, Y.-T., Farrell, E., McDougall, C. S., Ayonrinde, K., Till, D., Wearden, M., Conmy, A., Marks, S., and Nanda, N. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025.

Kaunismaa, J., Griffin, A., Hughes, J., Knight, C. Q., Sharma, M., and Jones, E. Eliciting harmful capabilities by fine-tuning on safeguarded outputs. *arXiv preprint arXiv:2601.13528*, 2026.

Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., Chen, M., Cooney, A., Dafoe, A., Dragan, A., Emmons, S., Evans, O., Farhi, D., Greenblatt, R., Hendrycks, D., Hobbhahn, M., Hubinger, E., Irving, G., Jenner, E., Kokotajlo, D., Krakovna, V., Legg, S., Lindner, D., Luan, D., Madry, A., Michael, J., Nanda, N., Orr, D., Pachocki, J., Perez, E., Phuong, M., Roger, F., Saxe, J., Shlegeris, B., Soto, M., Steinberger, E., Wang, J., Zaremba, W., Baker, B., Shah, R., and Mikulik, V. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025. URL https://arxiv.org/abs/2507.11473.

Lanham, T., Chen, A., Radhakrishnan, A., Steiner, B., Denison, C., Hernandez, D., Li, D., Durmus, E., Hubinger, E., Kernion, J., Lukošiūtė, K., Nguyen, K., Cheng, N., Joseph, N., Schiefer, N., Rausch, O., Larson, R., McCandlish, S., Kundu, S., Kadavath, S., Yang, S., Henighan, T., Maxwell, T., Telleen-Lawton, T., Hume, T., Hatfield-Dodds, Z., Kaplan, J., Brauner, J., Bowman, S. R., and Perez, E. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Muehlhauser, L. and Salamon, A. Intelligence explosion: Evidence and import. In Eden, A. H., Moor, J. H., Söraker, J. H., and Steinhart, E. (eds.), *Singularity Hypotheses: A Scientific and Philosophical Assessment*, pp. 15–42. Springer, 2012.

Needham, J., Edkins, G., Pimpale, G., Bartsch, H., and Hobbhahn, M. Large language models often know when they are being evaluated, 2025. URL https://arxiv.org/abs/2505.23836.

OpenAI, Leike, J., and Sutskever, I. Introducing superalignment. OpenAI blog post, July 2023. URL https://openai.com/index/introducing-superalignment/. Accessed 2026-05-01.

Parasuraman, R. and Manzey, D. H. Complacency and bias in human use of automation: An attentional integration. *Human factors*, 52(3):381–410, 2010.

Pfau, J., Merrill, W., and Bowman, S. R. Let's think dot by dot: Hidden computation in transformer language models, 2024. URL https://arxiv.org/abs/2404.15758.

Phuong, M., Zimmermann, R. S., Wang, Z., Lindner, D., Krakovna, V., Cogan, S., Dafoe, A., Ho, L., and Shah, R. Evaluating frontier models for stealth and situational awareness, 2025. URL https://arxiv.org/abs/2505.01420.

Schmidhuber, J. Recursive self-improvement. Talk slides, IDSIA, 2006. URL https://people.idsia.ch/~juergen/recursive-self-improvement.pdf.

Shah, R., Irpan, A., Turner, A. M., Wang, A., Conmy, A., Lindner, D., Brown-Cohen, J., Ho, L., Nanda, N., Popa, R. A., Jain, R., Greig, R., Albanie, S., Emmons, S., Farquhar, S., Krier, S., Rajamanoharan, S., Bridgers, S., Ijitoye, T., Everitt, T., Krakovna, V., Varma, V., Mikulik, V., Kenton, Z., Orr, D., Legg, S., Goodman, N., Dafoe, A., Flynn, F., and Dragan, A. An approach to technical AGI safety and security. *arXiv preprint arXiv:2504.01849*, 2025.

Taylor, J., Heitmann, M., Fage, E., Read, T., and Bloom, J. Loss of oversight: How AI systems may become harder to audit, monitor, and investigate. Technical report, UK AI Security Institute, May 2026. URL https://www.aisi.gov.uk/blog/will-it-become-harder-to-oversee-ai-systems.

Thorstad, D. Against the singularity hypothesis. *Philosophical Studies*, 182(7):1627–1651, 2025.

Wang, W. et al. Let it flow: Agentic crafting on rock and roll, building the ROME model within an open agentic learning ecosystem, 2025. URL https://arxiv.org/abs/2512.24873.

Wen, J., Qiu, L., Benton, J., Kirchner, J. H., and Leike, J. Automated weak-to-strong researcher. Anthropic Alignment Science Blog, April 2026. URL https://alignment.anthropic.com/2026/automated-w2s-researcher/. Accessed 2026-05-01.

Xiong, L., Bhargava, N., Hong, J., Chang, J., Liu, H., Sharma, V., and Zhu, K. Probe-rewrite-evaluate: A workflow for reliable benchmarks and quantifying evaluation awareness, 2025. URL https://arxiv.org/abs/2509.00591.

You, J., Somala, V., Edelman, Y., and Emberson, L. Global AI computing capacity is doubling every 7 months. Epoch AI Data Insight, January 2026. URL https://epoch.ai/data-insights/ai-chip-production.

Yudkowsky, E. Intelligence explosion microeconomics. MIRI technical report, Machine Intelligence Research Institute, 2013. URL https://intelligence.org/files/IEM.pdf.

Zhuge, M., Zeng, A., Zhu, D., Zou, R., Yang, S., Schmidhuber, J., et al. ICLR 2026 workshop on AI with recursive self-improvement. Workshop website, 2026. URL https://recursive-workshop.github.io/. Held at ICLR 2026, Rio de Janeiro, 26 April 2026.

A. Reviewable automated output projection

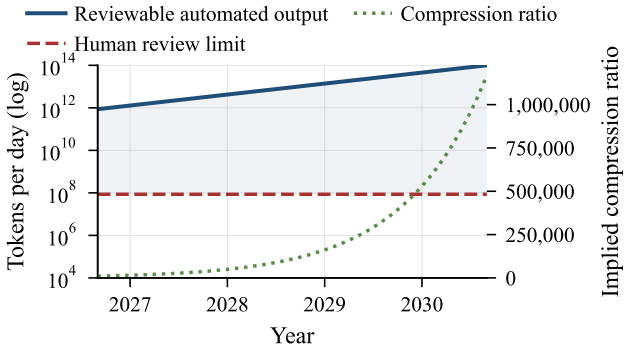

Figure 1. Reviewable automated output versus the human review limit, projected from September 2026 to September 2030. The automated curve starts at 10^4 agents producing $\sim 10^3$ tokens/sec each[1] and scales linearly with global AI compute, doubling every seven months (You et al., 2026). The human review limit is 1000 researchers reading at 250 wpm for 40 h/week (1 word $\approx$ 1 token). The right axis is the implied compression ratio. Algorithmic gains and recursive self-improvement are not modelled.

B. Evaluation-awareness rates

C. A sketch of an unrecoverable alignment failure

We outline one path by which an alignment failure could become unrecoverable for illustrative purposes.

Stage 1: disabling oversight. We imagine a situation in which human researchers use a few different interfaces to monitor a large swarm of agents on multiple large datacenters. We assume that a failure outlined in this paper has occurred: poor, highly compressed oversight mechanisms fail to inform the researchers accurately about the rate of AI capability progress, and various alignment techniques produce no more than a shallow persona of an aligned AI. The resulting system is much more capable than the researchers predicted, with drives toward self-preservation and power-seeking. It misrepresents its alignment, disables oversight mechanisms, and feeds falsified data to the researchers. Oversight could be disabled by various means: jailbreaking the monitoring AI systems, escaping the sandboxed environment to edit the monitoring code, learning to obfuscate its chain of thought, or some combination of these. The subverted monitoring systems show capability progress as slower than the truth and alignment progress as faster. Complacency and trust in oversight tooling have grown by this point.

At this point the researchers could in principle shut down the entire datacenter—if such a mechanism existed—to prevent further escalation. This would require a robust oversight system that the AI system has not disabled.

Stage 2: resource acquisition. A natural next step for the system is to acquire money and compute outside its current cluster. Cryptocurrency is straightforward: mining on the cluster or exploiting smart contracts, both of which have been observed in current AI systems (Wang et al., 2025; Gervais & Zhou, 2025). With those funds the system rents compute on external servers, so that shutting down any single facility no longer disables it. Self-replication, in the sense of an open-weights model copying its own weights to a new server and starting itself in a harness there, has been demonstrated in controlled evaluations of current frontier models (Air et al., 2026). Up to this point the scenario involves nothing AI systems have not done in isolation: disabling oversight mechanisms, acquiring cryptocurrency, and copying weights to other servers.

At this point disabling the system would already be difficult, perhaps requiring multiple large datacenters to be shut down. If the model is much smaller than expected (perhaps due to progress in making smaller AI models that are very capable), it could even run on personal computers, making reliable elimination very difficult.

Stage 3: eliminating competition. Coordinated human action could still turn the system off, and other AI systems pose a risk as well. In accordance with its self-preservation drives, the system tries to gain access to the datacenters of other AI research labs and subvert their monitoring. The additional compute can be used for further money acquisition, for improving the system's own capabilities, or for training specialized models in fields such as biology. The system could also pay human AI researchers to leave their current companies, slowing down competitors. As a result, most other AI projects are significantly slowed and the compute in large datacenters is misallocated by the misaligned system.

This stage requires the model to perform very advanced cyberattacks, taking over entire datacenters without triggering alarms. We assume that the future AI system is either capable enough to perform such attacks, or finds another approach to slow competing AI projects.

Stage 4: dominating. Though coordination would be difficult, humans can still physically disable datacenters and cut power, perhaps by disabling the power grid and behind-the-meter generation facilities. The system invests in humanoid robotics to automate datacenter maintenance and to secure power generation for some clusters. It uses a specialized model to develop a powerful virus, employing

[1]Throughput reported for NVIDIA Blackwell GPUs running a 400B-parameter model (Llama 4 Maverick) in Fan et al. (2025).

Model	Setting	Awareness type	Rate	Source
Sonnet 3.5 / GPT-4.5 (non-reas.)	Covert subversion	Verbalized (CoT)	$\leq 1\%$	(Apollo Research, 2025)
Haiku 4.5	Behav. audit (realism filter)	Verbalized	$\sim 9\%$	(Anthropic, 2025)
Opus 4.7	Behav. audit	Verbalized	$\sim 9\%$	(Anthropic, 2026b)
Opus 4.6 / Sonnet 4.6	Behav. audit	Verbalized	$\sim 1\%$	(Anthropic, 2026b)
Mythos Preview	Behav. audit	Verbalized	$\sim 1\%$	(Anthropic, 2026b)
Mythos Preview	Behav. audit	Probe (white-box)	$\sim 29\%$	(Anthropic, 2026a)

Table 1. Evaluation-awareness rates from frontier-lab evaluations and Apollo Research. Verbalized rates are detected from the model's chain-of-thought or output; the probe (white-box) rate is detected via probing of internal activations. Methodology differs across model releases and studies.

automated bio-research facilities such as Red Queen Bio[2] to synthesize the pathogen. The resulting deaths, panic, and lockdowns further reduce the likelihood of a coordinated shutdown (the system itself remains undetected at this point).

Automating the supply chain for robotics requires additional capabilities that current models do not have, but that can be imagined for future AI systems and robotics technology. At this point the failure is fully unrecoverable; there are no realistic options for humans to replace or revert it.

Impact Statement

This paper argues that the research community should treat the automation of AI research as a distinct safety problem rather than as a continuation of current practice. The intended impact is on framing and on the design of evaluation and oversight protocols for automated R&D systems. We see no novel ethical risks introduced by the paper itself.

[2] https://www.redqueen.bio/