
Algorithmic Governance for A Narrow Path

Raymond
Koopmanschap
Independent

Josh Thorsteinson
Independent

Lucy Luo
Independent

With
ControlAI & Apart Research

Abstract

This research evaluates the policy framework detailed in A Narrow Path, identifying vulnerabilities that could allow for artificial superintelligence (ASI) development despite global implementation. Using a red teaming approach, our analysis identifies algorithmic efficiency improvements as a major threat. Such advancements can undermine compute-centric controls (Policies 1, 3, and 6) by enabling more AI capabilities with fewer computational resources. To address this, we recommend: (1) expanding the International AI Safety Commission's (IASC) role to include proactive algorithmic risk assessment; (2) introducing new policies to restrict high-risk algorithm development; (3) mandating rigorous safety assessments for algorithms with significant efficiency gains, and (4) watermarking licensed AI models to catch model distillation. These measures aim to strengthen "A Narrow Path" by governing AI algorithmic progress, thereby supporting compute-centric controls in preventing dangerous ASI development.

Keywords: AI policy, red teaming, superintelligence governance, regulatory implementation, safety cases, threat modeling

Introduction

This work critically examines A Narrow Path's policy framework. Its central objective is to identify potential circumvention pathways and weaknesses with the document's proposed policies. We ask: assuming all of the document's policies are implemented globally, what vulnerabilities would remain that could allow for the development of ASI, and how can these loopholes be addressed? We anticipate legal definitions, challenges to compute governance, and possible policy exploits.

We red team A Narrow Path’s policy framework by developing scenarios where state, corporate, or individual actors circumvent these controls. While we began by exploring several threat models, we converged on algorithmic efficiency improvements as seen in the “Rebranded AGI Company” scenario. Our analysis focuses on Phase 0 policies as a network of defences and is most relevant to Policies 1, 3, and 6.

Methods

The core of our methodology involved a structured red teaming process. We initially approached the analysis from three distinct perspectives, each undertaken by a different team member:

1. **Technical Feasibility:** Assessing how advancements in AI technology, hardware, and algorithms could enable circumvention.
2. **Legal Circumvention:** Examining how actors might exploit ambiguities, loopholes, or the letter of the law within signatory nations to advance ASI development.
3. **Covert Circumvention:** Investigating how actors (state or non-state) might operate outside legal frameworks or in non-signatory regions to pursue ASI undetected.

Team members compared findings from these perspectives, identifying overlapping themes and potential failure modes, such as definitional ambiguities in policy language and the persistent challenge of accounting for algorithmic efficiency gains. We initially formulated four distinct circumvention scenarios: two exploring legal pathways and two focused on covert operations. These scenarios were then detailed and assessed for plausibility and impact. The team generated possible details of circumvention and searched for clauses that may restrict each path. Through this iterative process, we converged to focus on algorithmic efficiency improvements as a salient and likely failure mode across various circumvention strategies.

To inform our assessment of the viability of governing algorithmic efficiency improvements, we conducted a historical case study analysis. This involved examining precedents from nuclear, biological, and encryption technologies relying primarily on Vermeer (2024). Based on the surveyed historical precedents, a full halt to continued algorithmic improvements appeared infeasible due to verification and enforcement challenges. However, we identified potential methods of slowing algorithmic progress.

Having identified algorithmic efficiency as a critical threat vector, we undertook further analysis to develop targeted policy patches, both for well-resourced actors aiming to develop highly efficient algorithms internally and for generally limiting dangerous AI algorithmic progress.

LLMs (e.g., Gemini 2.5 Pro, Claude Opus 4) were instrumental in supporting our research process. They were used for brainstorming, scenario formulation, idea generation, soliciting feedback, and writing. Please see Appendix 2 for prompts used.

Threat Modeling: Rebranded AGI Company

Epistemic Status

Key facts like amount of distillation and watermarking are roughly checked but papers are not completely read and a quick summary generated by Gemini was used. The idea of distillation and watermarking was proposed by Gemini.

Scenario 1: Rebranded AGI Company

1. An AGI company (rebranded as just an AI company) trains the largest possible AI model within the licensing regime, establishing a public record of compliance.
2. Knowledge distillation enables the transfer of the teacher's core reasoning abilities into a new model that might be over 100-1000x smaller, placing it outside all licensing and regulatory oversight. Research shows that student models (e.g., Orca 2, from Microsoft) can achieve reasoning abilities comparable to teacher models (e.g., GPT-4) that are orders of magnitude larger, confirming the feasibility of this step (Mukherjee et al., 2023; Mitra et al., 2023).¹
3. Thousands of copies of this small, unregulated “student” model are spun up on unregulated servers to discover novel AI breakthroughs, completely invisible to regulators.²
4. The secret discoveries from the swarm are used to build ASI with minimal, unregulated compute, achieving a decisive strategic surprise that is lower than the threshold.³

The Central Failure Mode: This strategy exploits the fact that distillation can separate a model's capabilities from its size. It allows an adversary to create a highly capable, unregulated asset from a regulated one, rendering compute-based monitoring and licensing ineffective.

Recommended Policy Amendment

To patch the model distillation vulnerability, we recommend adding something like the following amendment:

As a condition of receiving a Training License, the developer must allow the regulatory body to embed a robust, non-falsifiable cryptographic watermark into the final model weights.

¹ Not the same abilities as GPT-4 but comparable. Thus the distilled model needs to specialize in for example research tasks. This does fall roughly under AIs improving other AIs but since it is not detected and can't be detected after the fact it can never be checked.

² Inference doesn't need extreme amounts of compute and can work on different tasks in different datacenters in parallel which is not discoverable since these data centers don't communicate with each other.

³ This works because the threshold doesn't get adjusted by hidden capabilities.

This watermark is a prerequisite for a model's legal use. Any attempt to tamper with it will result in immediate license revocation and severe penalties.

The presence of a licensed model's watermark in any other model will constitute evidence of its origin. This is based on research demonstrating that watermarks are robustly transferred to derived models through processes like knowledge distillation (Aaronson et al., 2023). This method is likely robust to distillation methods (Kirchenbauer, J., et al. (2023).

This watermark allows to catch evidence of AI training AI (policy 3). The watermark policy makes sure this specific distillation violation can be identified after the fact by detecting the watermark in the distilled model.

Note: This doesn't catch the model. But if the model is detected there is clear proof they did something illegal. This forces the company to keep every distilled model that is derived from a licensed model secret which is a significant cost.

Algorithmic Efficiency Forecast

Epistemic Status

The legwork for this section was done by Gemini.⁴ Here, we just wanted to get a rough sense for how feasible compute governance would be after 20 years of algorithmic improvements.

Forecast

A Narrow Path establishes three compute thresholds:

- Upper Limit: 10^{27} FLOP
- Middle Limit: 10^{25} FLOP
- Lower Limit: 10^{23} FLOP

The document specifies that these thresholds will be updated over time to account for innovations in algorithmic efficiency. But by how much will the thresholds need to be lowered? History shows that AI algorithms become dramatically more efficient over time. For LLMs, the compute needed to achieve the same capabilities has halved roughly every 8 months.

The proposed framework bans "AIs improving AIs." This would likely slow algorithmic efficiency research, but human-led innovation would continue. To account for this, we assume a 16-month doubling.

Over 20 years (240 months), efficiency improvements occurring every 16 months would result in 15 doublings of algorithmic efficiency. This creates a total improvement factor of $2^{15} = 32,768$.

⁴ Chat: <https://g.co/gemini/share/e33a79693b1e>.

To offset this, computing thresholds must decrease by the same factor:

- New Upper Limit: 3.05×10^{22} FLOP
- New Middle Limit: 3.05×10^{20} FLOP
- New Lower Limit: 3.05×10^{18} FLOP

GPT-3 was trained using an estimated 10^{23} FLOP (Li, 2023), an order of magnitude more than the New Upper Limit. By AI standards, GPT-3 is now an old and small model. This presents a strong challenge to the longevity of A Narrow Path's compute-centric model.

Recommendations

A Narrow Path relies heavily on compute governance. This is feasible as long as it takes massive computational resources to train frontier AI. But over time, increasingly efficient algorithms will diminish the effectiveness of compute-centric regulations, as forecasted in the previous section. This makes algorithmic progress itself an important variable to address directly, especially considering the 20+-year ambition of the report. We propose the following policies:

1. Expand the International AI Safety Commission's Role

The International AI Safety Commission (IASC) should do more than just monitor computing power. It should actively track algorithmic research to identify developments that could threaten global safety. Instead of simply reacting to problems after they emerge, the IASC would stay ahead of dangerous research trends.

2. Create "Algorithmic Research Areas of Concern" (ARACs)

The IASC should maintain a ranked list of research areas that pose risks. This list would help national governments and international bodies understand which types of algorithmic research need closer oversight. It would also guide funding decisions, steering resources toward safer research directions.

3. Restrict the Most Dangerous Algorithmic Development

Some algorithmic research poses extreme risks and should face strict limitations. The IASC would identify these "Extreme Risk" research areas and severely restrict who can work on them. Only specially licensed researchers in highly secure, audited facilities, like GUARD, would be permitted to pursue this work.

4. Require Safety Reviews Before Research Begins

Researchers planning major algorithmic breakthroughs should complete an "Algorithmic Impact Assessment" before starting their work. This review would evaluate potential risks and establish safety measures upfront, making safety considerations a standard part of the research process rather than an afterthought.

Conclusion

This analysis highlights the threat posed by algorithmic efficiency improvements to A Narrow Path's compute-centric policy framework. Our red teaming, particularly the “Rebranded AGI Company” scenario, demonstrates how these advancements can undermine regulatory efforts by enabling greater AI capabilities with less computational resources. To safeguard against this, we propose expanding the IASC's mandate to include proactive algorithmic risk assessment, introducing new policies to restrict high-risk algorithm development, and mandating rigorous safety assessments for algorithms with significant efficiency gains. These measures would help to enable the long-term effectiveness of A Narrow Path in preventing existentially risky ASI development.

References

- Aaronson, S., et al. (2023). Watermarking Large Language Models.
- Kirchenbauer, J., et al. (2023). A Watermark for Large Language Models. arXiv preprint arXiv:2301.10226.
- Li, C. (2023, August 3). OpenAI's GPT-3 Language Model: A Technical Overview. Lambda.
https://lambda.ai/blog/demystifying-gpt-3?srsltid=AfmBOoq9w28xnlaaTWaIMi8btYhTd4fUBCCfTjwfl_0-uxbwDEVjas-m
- Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). A Narrow Path: How to Secure Our Future. Control AI. <https://www.narrowpath.co/>
- Mitra, A., et al. (2023). Orca 2: Teaching Small Language Models How to Reason. arXiv preprint arXiv:2311.11045.
- Mukherjee, S., et al. (2023). Orca: Progressive Learning from Complex Explanation Traces of GPT-4. arXiv preprint arXiv:2306.02707.
- Vermeer, M. J. D. (2024, August 19). Historical analogues that can inform AI governance. RAND.
https://www.rand.org/pubs/research_reports/RRA3408-1.html

Appendix 1: Additional Threat Scenarios

Scenario 2: Secret ASI Development

1. A well-resourced nation-state launches a classified program to develop domestic chip production capabilities over two decades. These chips don't need to be cutting-edge—no advanced equipment like EUV machines required. By the time production comes online, algorithmic breakthroughs will have dramatically reduced the computing power needed to train superintelligent AI.
2. Using domestically produced chips, the nation builds multiple data centers. Each facility appears modest in size to avoid detection, but together they provide enough computing power for ASI development. Advanced distributed training techniques allow these separate centers to work as one massive system.
3. The distributed network trains ASI while staying below international monitoring thresholds. Each individual data center looks unremarkable, but their combined capability crosses the superintelligence threshold.

Supporting Tactics

1. The nation operates a more visible data center that deliberately pushes treaty boundaries. This facility generates international controversy and diplomatic drama, consuming regulators' attention while the real ASI development proceeds unnoticed.
2. The nation systematically weakens global AI safety institutions by:
 - Creating bureaucratic obstacles that slow down investigations
 - Flooding oversight bodies with trivial complaints and false alarms
 - Supporting regulatory complexity that makes effective monitoring nearly impossible

Scenario 3: Decentralized consortium of companies

1. A decentralized consortium of well-funded, ideologically aligned tech incubators, philanthropic foundations (with no publicly known relationships) orchestrated by a council of ex-industry pioneers and e/accs/disillusioned futurists. Some organizations are disguised as AI ethics research bodies and others as specialized AI R&D startups. They operate across multiple signatory nations.
2. Each startup focuses on developing advanced AI sub-components (e.g., novel reasoning modules, hyper-efficient learning algorithms). Each project is kept below the Training License (TL) compute thresholds or if a TL is needed, submit highly convincing safety cases arguing for “narrow utility” and “inherent boundedness” of their isolated component.
3. To avoid “AI directly improving AIs”, workflows within these organizations are labeled “human-led R&D with AI assistants”. They extend the definition of “direct use” as much as possible, using the most advanced, permitted tools to support R&D. While avoiding recursive self-improvement in the magnitude of 100x, the compounded impact of estimated 5x per year makes large strides in intelligence.

4. Beyond official training licenses, compute is sourced through numerous small contracts with various Compute Licensed (CL) providers, none individually triggering major alarms.
5. Each of the numerous small compute contracts with different Compute Licensed (CL) providers would be assigned a portion of the overall model or a shard of the total dataset. Model updates (gradients or partial weights) would then be securely and discreetly exchanged between these isolated clusters, possibly using encrypted peer-to-peer networks or a covert central server that aggregates updates. They bring to the world their vision of their future with their ASI.

Appendix 2: Prompts

Lucy's Prompt:

-  A Narrow Path P1.pdf ,
<https://aistudio.google.com/app/prompts?state=%7B%22ids%22:%5B%221ap891SmCve5xWwMlkr6tuPIYylA8hb0m%22%5D,%22action%22:%22open%22,%22userId%22:%22113071541434776706307%22,%22resourceKeys%22:%7B%7D%7D&usp=sharing>

Raymond's prompts

- <https://claude.ai/share/5a8e1bfc-1e5f-4b9d-b43c-89a107a8a45a> (This was a Claude Project. In the Project content was the Narrow Path Phase 0 information and this prompt: "We have a day-long hackathon where we will red-team Phase 0 of the Narrow Path. The goal of the project is: Help strengthen the Phase 0 policies by identifying weaknesses, implementation challenges, and gaps before they reach legislators' desks. Assume
- - All policies are globally implemented (don't focus on feasibility of implementation for now)
- - An adversarial company/project that want to still build Artificial Superintelligence given these policies either tries to
- 1. Legally work around these policies and still build artificial superintelligence
- 2. Tries to stay undetected and build artificial superintelligence in secret
- We should critique the plan but also propose solutions of how we can improve it.
- We got advised to focus on one of these six policies. But in the initial stage might be good to look at all 6 of them and see which is best to tackle.
- Gemini:
<https://drive.google.com/file/d/1PUSLVxrbUfjr7BGAA0oYtVw6T8BcER3S/view?usp=sharing>,
<https://aistudio.google.com/app/prompts?state=%7B%22ids%22:%5B%221sZv7jgIFhZdqUa-ueKgwGOVRUHqiu-q6%22%5D,%22action%22:%22open%22,%22userId%22:%22118071295904479530486%22,%22resourceKeys%22:%7B%7D%7D&usp=sharing>

Josh's prompts:

- <https://aistudio.google.com/app/prompts?state=%7B%22ids%22:%5B%221ao1mXgXJSx9tstNw1nxNILygyE5YGVm6%22%5D,%22action%22:%22open%22,%22userId%22:%22112789485224936051313%22,%22resourceKeys%22:%7B%7D%7D&usp=sharing>
- <https://aistudio.google.com/app/prompts?state=%7B%22ids%22:%5B%221WG1JYNysrgvd0t54mxBA4hR4r9v3D5vj%22%5D,%22action%22:%22open%22,%22userId%22:%22112789485224936051313%22,%22resourceKeys%22:%7B%7D%7D&usp=sharing>

- <https://aistudio.google.com/app/prompts?state=%7B%22ids%22%3A%5B%221ZlclYvwrXyCfAAY62WnEEzbCLjcvDZX4%22%5D%22action%22%3A%22open%22%22userId%22%3A%22112789485224936051313%22%22resourceKeys%22%3A%7B%7D%7D&usp=sharing>
- <https://aistudio.google.com/app/prompts?state=%7B%22ids%22%3A%5B%221YtU2fB5-TjPEAFNL0JFCAENwwfw8YNkU%22%5D%22action%22%3A%22open%22%22userId%22%3A%22112789485224936051313%22%22resourceKeys%22%3A%7B%7D%7D&usp=sharing>
- <https://g.co/gemini/share/483f15e9bba0>
- <https://aistudio.google.com/app/prompts?state=%7B%22ids%22%3A%5B%221KUtsznFYUAL8cLKOZW4APFA333SQhiLm%22%5D%22action%22%3A%22open%22%22userId%22%3A%22112789485224936051313%22%22resourceKeys%22%3A%7B%7D%7D&usp=sharing>
- <https://g.co/gemini/share/8cfae8443913>
- <https://g.co/gemini/share/e33a79693b1e>
- <https://claude.ai/share/56578f7a-2b4a-4728-9450-538029eae55d>

Appendix 3: Nitpicks

1. In phase 0, policy 4 states: “Before **any** AI system is trained or deployed, its developer must provide reliable safety cases” (our emphasis). This doesn’t specify that the policy only holds for licensed models, despite us being told on Discord that this is the case.
2. Apparently the “no AIs improving AIs” policy applies for *any* AI. This seems infeasible to implement in reality.