
Red Teaming A Narrow Path: ControlAI Policy Sprint

Diego Dorn
Independant

Camille Truchot
Independant

With
ControlAI & Apart Research

Abstract

While the 'Narrow Path' policy confronts the essential risk of recursive AI self-improvement, its proposed enforcement architecture relies on trust in a fundamentally non-cooperative and competitive domain. This strategic misalignment creates exploitable vulnerabilities.

Our analysis details six such weaknesses, including lack of verification, enforcement, and trust mechanisms, hardware-based circumvention via custom ASICs (e.g., Etched), issues with 'direct uses' of AI to improve AI, and a static compute cap that perversely incentivizes opaque and potentially risky algorithmic innovation.

To remedy these flaws, we propose a suite of mechanisms designed for a trustless environment. Key proposals include: replacing raw FLOPs with a benchmark-adjusted 'Effective FLOPs' (eFLOPs) metric to account for algorithmic gains; mandating secure R&D enclaves auditable via zero-knowledge proofs to protect intellectual property while ensuring compliance; and a 'Portfolio Licensing' framework to govern aggregate, combinatorial capabilities.

These solutions aim to participate in the effort to transform the policy's intent into a more robust, technically-grounded, and enforceable standard.

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

1 Introduction

This project addresses a central research question: *How robust are the Phase 0 regulatory policies proposed in A Narrow Path against practical challenges and evasions in preventing unsafe ASI development?* Our hypothesis is that while these policies provide a necessary foundation, significant verification weaknesses, definitional ambiguities, and regulatory loopholes undermine their effectiveness, particularly against sophisticated actors such as states, corporations, and rogue researchers motivated by competitive or strategic incentives.

We focus our analysis on two critical Phase 0 policies from *A Narrow Path*: the prohibition on developing and using AIs that improve other AIs, and the licensing regime and restrictions on the general intelligence of AI systems.

Our threat model first considers private corporations, and individual rogue researchers with varying strategic, competitive, or malicious motivations, but also covers states to some extent. We examined key failure modes, such as regulatory circumvention strategies, the opacity of private labs, and the intrinsic limitations of enforcement models relying primarily on self-reporting.

Taking a step back, our analysis also identified a cross-cutting weakness in all Phase 0 policies examined: the lack of viable **verification and enforcement mechanisms** for regulators, compounded by the significant challenges posed by relying solely on **national enforcement** in a highly globalized AI ecosystem. This dual gap seriously undermines the effectiveness of the proposed measures, leaving room for fraud, regulatory evasion, and “forum shopping” across jurisdictions.

Our contribution aims to strengthen these policies by offering concrete recommendations. This work supports the ongoing efforts by Control AI to inform lawmakers in the UK and US through the Direct Institutional Plan, providing a critical and pragmatic analysis to make ASI governance frameworks more robust, adaptable, and enforceable in practice.

2 Methods

Our critique of Phase 0 followed a structured, iterative process combining rapid ideation, threat modeling, and comparative policy analysis, grounded in both empirical reasoning and historical analogy.

We began by **generating a broad set of potential loopholes** through brainstorming and exploratory reasoning, leveraging both human insight and large language models for shallow evaluation. This allowed us to quickly surface weak points in the proposal's licensing logic, enforcement mechanisms, and assumptions about model capability control.

From this initial list, we **prioritized high-impact loopholes** based on plausibility, potential severity, and tractability of red teaming. We then **refined these loopholes**, specifying concrete exploit pathways and actor strategies that could systematically undermine the policy's goals.

Once these threats were formalized, we **conducted focused solution sprints**, generating and evaluating multiple potential patches for each loophole. For each proposal, we assessed feasibility and remaining challenges.

Finally, we synthesized our findings into a cohesive write-up, emphasizing clarity, justification, and policy relevance. This method balanced **creative adversarial thinking** with **grounded policy reasoning**, aiming to deliver proposals that are both innovative and actionable.

3 Results

Weakness 1: The policy lacks a viable verification mechanism.

Evidence: It's unclear how regulators would monitor compliance and detect a breach of the proposed measures, in particular *inside* a private lab. Regulators currently lack clear, practical mechanisms to independently verify whether entities are adhering to the proposed restrictions, especially when proprietary information and competitive advantage incentivize secrecy and obfuscation. The policy relies mainly on self-reporting, which is insufficient because organizations have strong motives to withhold or distort information.

Proposal 1A: Provide for investigative powers of the regulator:

- Power to order provision of any information or documents required for assessing compliance with the rules. Such a power already exists for supervisory authorities under the European Union General Data Protection Regulation (GDPR), as well as in the European Union Artificial Intelligence Act (AI Act) where the Commission may request the provider of the general-purpose AI model concerned to provide the documentation drawn up in accordance with its obligations.
- Power to conduct on-site inspections (announced & unannounced). This includes physical inspection of data centers (to verify hardware) and offices (to inspect security procedures). A precedent also exists in the GDPR.
- Power to access a model's weights, source code, and training/fine-tuning datasets for forensic analysis. Such power is already provided in the EU AI Act, as the Commission may request access to the general-purpose AI model.
- Power to mandate third-party audits. This leverages private sector expertise while maintaining regulatory oversight. Such a delegation power also already exists in the EU AI Act for the Commission's evaluations.

Proposal 1B: Enhance protection and incentives for whistleblowers: Whistleblowers are crucial to detect illicit or non-compliant behavior within organizations, especially when such behavior is difficult or impossible to detect externally. To encourage and protect them:

- Establish **strong legal protections**, including anonymity, anti-retaliation clauses, and employment safeguards.

- Introduce **positive incentives**, such as financial rewards or public-interest exemptions from NDA clauses.
- Build **secure, independent reporting channels**, e.g. encrypted online portals, or external ombuds services.

Proposal 1C: Logged R&D: All foundation model R&D should occur in a digitally secure environment where all code commits, data movements, and model weight changes are cryptographically logged. This is not real-time surveillance. It is an immutable ledger for *retrospective audits*.

Proposal 1C: Strict confidentiality obligations: All regulators and third-party auditors with access to sensitive data must be bound by strict confidentiality obligations.

Proposal 1D: API monitoring: The API requests could be monitored randomly to ensure that there is no breach of the rules. Such monitoring could be done by AI systems that would not record information but only alert in case of suspicious activity.

Remaining challenges:

- A government could **classify its own superintelligence research as a national security asset**, rendering domestic regulations irrelevant and creating an unlevel playing field. This “military loophole” creates a two-tier system: one for private actors, and another for state-run ASI efforts.
- The policy requires authorities to verify highly technical claims, from assessing a developer's R&D pipeline to auditing a model's latent capabilities. This necessitates **regulators staffed with world-class AI engineers**. We can anticipate difficulties in recruiting such profiles.
- The enforcement structure envisioned in “A Narrow Path” is entirely national in Phase 0. However, **without coordination at the international level, a risk of “forum shopping” emerges** where AI developers strategically relocate to jurisdictions known for their regulatory leniency or lack of enforcement resources. This could create pressure to lower standards across countries, undermining the policy’s global intent. To mitigate this, core verification and licensing obligations must eventually be harmonized, with mutual recognition frameworks, shared compliance data, and minimum enforcement capacity standards across jurisdictions.

Weakness 2: The "direct use" definition is unenforceable and incentivizes concealment

Evidence: The policy's reliance on the term "direct use" creates a definition that is both too broad and too narrow.

- **Overly Broad:** It bans established industry techniques like RLHF, where one model's output directly optimizes another. It also bans models used in adversarial red-teaming done between fine-tuning steps
- **Overly Narrow:** It does not ban the use of AI to develop algorithmic improvements that are not directly about "creating found systems". For example, advanced AI agents could be used to improve chip design, fast algorithms to share data across a network (which could be used for distributed training), or optimize compilers and kernels. It is unclear if this is enough to build superintelligence, but could provide a significant boost to algorithmic improvements.

Proposal 2A: Allow-list and block-list. Every use of AI in an AI context must be explicitly approved by a central authority (e.g. the IASC), and once approved or rejected is added to a public allow-list or block-list. This makes definitions of "direct use" and "found system" explicit, but requires a central authority to manage the list. This process would likely be slow and bureaucratic, but enable oversight into what is precisely done in terms of AI improving AI. A more robust proposal is likely 2B.

Proposal 2B: A tiered classification system to replace ambiguity with a clear, risk-based framework. Mandate a three-tier system for any use of AI within an AI development pipeline:

- **Tier 1 - Green (Permitted):** Negligible risk uses (e.g., typo correction, generic search). No approval required.
- **Tier 2 - Grey (Monitored):** Indirect acceleration risks (e.g., datacenter and GPU optimization, ML code debugging). Permitted, but all uses must be logged in a standardized, auditable (public?) format. This creates data trails for auditing and future analysis.
- **Tier 3 - Red (Restricted):** Direct recursive improvement risks (e.g., architecture generation, foundation model optimization). Banned by default, requiring a specific, high-scrutiny license for each instance.

This framework would not be static, and be subject to review at any time by the IASC.

Weakness 3: The "software-only" definition creates a hardware escape route

Evidence: The policy defines "found systems" exclusively as "software programs." This focus on software is a mistake. A company transforms a frontier AI into a hardware accelerator (ASIC/FPGA), directly embedding the weights into the hardware. They can then use it for inference, and task it to improve AI systems. This process is already done in some capacity by [Etched](#). By fabricating and then using this hardware, the company can achieve massive performance gains, all while technically complying with the policy. They are not using a found *software program*, but a found *hardware design*, circumventing the policy's intent entirely.

Proposal 3: Redefine "system" to be medium-agnostic. The restriction must apply to any AI-generated **design artifact**, whether software logic or hardware architecture.

Weakness 4: The fixed compute threshold perversely incentivizes algorithmic improvements

Evidence: The 10^{27} FLOP training limit is not a future ceiling; it is a present-day constraint, with models like Grok-3 already [approaching this scale](#) (est. 4.6×10^{26} FLOP), and the largest AI companies are likely to launch the first training runs above this ceiling this year. The effect of the FLOP limit will be to redirect companies to focus on algorithmic improvements, which are less understood, harder to track, and potentially introduce more unpredictable risks than brute-force scaling. The policy also fails to account for inference-time compute, where complex agentic behaviors can emerge without any new training.

Proposal 4A: Regulate "effective FLOP" (eFLOP). A static FLOP limit is easily gamed. Instead, a company's licensed compute budget for a training run should be dynamically adjusted by its algorithmic efficiency. Before a large run, the company must benchmark its new algorithm on a standardized reference model (e.g. how much compute does it take to train a model with the same benchmark performance as GPT-4). A demonstrated 2x efficiency gain on the benchmark would halve the company's 10^{27} FLOP allowance to 5×10^{26} eFLOP for that run. This forces the externality of algorithmic risk onto the developer's balance sheet.

It's important to note this proposal is a much stronger regulation than the previous one, as it bounds LLM capabilities, roughly at the current level. On the other hand, it takes 3 times less compute to train the same model every year [according to Epoch](#). In 10 years, this amounts to a scaling of $3^{10} = 59049$, or 4 orders of magnitude. We should expect at least this amount of algorithmic improvements, especially if it's the only way up that this policy allows.

Proposal 2C/4B: Directly bound the rate of algorithmic progress Instead of trying to regulate a cause of recursive self-improvement (AI improving AI, as mentioned in our Weakness 2), a policy could regulate the output of RSI: algorithmic improvements. New techniques could be allowed only if the efficiency gains to train a fixed model are below a certain threshold, which can increase over time. This would allow for instance to bound algorithmic improvement to 3x every year (as is currently the case) or lower. While eFLOPs cap total capability per run, this proposal would cap the annual rate of efficiency gains, directly targeting the recursive feedback loop.

Weakness 5: Absence of portfolio-based analysis enables AI swarms to bypass controls

Evidence: An actor could train dozens of individual AI models, each operating below the 10^{25} FLOP training limit required to trigger licensing. Because the current licensing system regulates models on an individual basis rather than the combined capabilities of integrated systems, these numerous smaller models could be orchestrated together as a "swarm" or collective system. This swarm could achieve capabilities that surpass the regulatory thresholds without any single model crossing the licensing criteria, effectively circumventing oversight and enabling powerful AI development without triggering the necessary licenses.

Proposal 5: Portfolio Licensing Framework

- **Trigger licensing based on aggregated portfolio:** Introduce a "cumulative threshold": if a developer operates several models (or services) the total of which reaches a FLOP threshold (even if no individual model crosses FLOP thresholds), a Portfolio License becomes mandatory. This could require periodic mandatory disclosures of all operational models.
- **License capability portfolios:** Limit each company to operating in only one of pre-defined high-risk categories under a standard license. Developing across several high-risk categories requires special, high-scrutiny permits.

Weakness 6: The verification framework lacks mechanisms to secure stakeholder buy-in

Evidence: Even with a verification system in place, it will fail if it is not trusted by key stakeholders. Companies need to trust that other companies follow the rules, especially when enforcement is hard, and country need to trust other countries enforce this proposal as well as they do. The proposal lacks the granular mechanisms needed to build this essential trust.

- **Deficit in Public Trust:** The public and international partners have no means to independently verify that large-scale compute resources are being used in compliance with the rules. This allows for suspicion that secret, non-compliant training runs are underway, eroding the basis for international stability.
- **Deficit in Corporate Trust:** Companies will view any audit as an unacceptable risk to their intellectual property. Without a protocol that protects trade secrets, they will have a powerful incentive to legally challenge, obstruct, or weaken the entire verification regime, rendering it ineffective.

Proposal 6: Establish a public compute registry. Mandate a live, public registry that maps significant compute clusters to their owner and current, high-level task allocation (e.g., "Allocated to Licensed Run #789", "Allocated to Below-Threshold Project 'Helios'", "General Commercial Use"). This provides part of the transparency needed to reassure countries and companies that others are playing the game fairly, and reinforces democratic oversight. This provides a simple, verifiable check on large-scale resource use without exposing sensitive data.

4 Discussion and Conclusion

Our red teaming analysis reveals that while the Phase 0 policies outlined in *A Narrow Path* represent a crucial foundation for ASI risk mitigation, they face significant practical vulnerabilities that could undermine their effectiveness in preventing the emergence of artificial superintelligence over the next 20 years. The most critical weakness across these policies is the absence of robust, legally mandated verification and enforcement mechanisms.

A regulatory framework is toothless without well-funded, technically sophisticated, and empowered authorities capable of investigating AI activities. Although *A Narrow Path* proposes a comprehensive set of rules, it does not fully grapple with the immense institutional and operational challenges involved in building such a regulator. Without substantial investment in regulatory capacity and, crucially, formalized international cooperation, Phase 0 risks being widely adopted in form but easily circumvented in practice.

Also, there is an urgent need for a harmonized international enforcement framework that prevents regulatory arbitrage and ensures consistent compliance worldwide.

Developers, motivated by competition and secrecy, could exploit gaps such as multi-model orchestration (“AI swarms”), fixed compute thresholds incentivizing algorithmic workarounds, and other elements highlighted in our analysis to continue advanced AI development under the radar. This highlights the need to close any gap in the proposed policy.

In conclusion, to meaningfully prevent ASI development for at least the next two decades, these Phase 0 policies must be complemented by a coordinated global governance regime that empowers regulatory bodies with the authority, resources, and cross-border cooperation necessary for effective oversight.

5 References

Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). *A Narrow Path: How to Secure Our Future. Control AI*. <https://www.narrowpath.co/>

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*.

European Parliament and Council. (2016). *General Data Protection Regulation (GDPR)*, Regulation (EU) 2016/679.

6 Appendix

LLM usage documentation

Prompt 1

<https://www.narrowpath.co/phase-0>

What are promising angles to red-team this policy proposal? Put yourself in the shoes of Zvi Mowshowitz.

Prompt 2

<https://www.narrowpath.co/phase-0>

I'm trying to make strong red-teaming of this policy. Assuming this policy has been drafted in detail, and applied everywhere in the world. How would you proceed to develop ASI as a rogue actor.

Some guidelines: Adopt an attacker minset, strongman the policy, brainstorm with different mindsets (e.g. find loopholes that Zvi would say, and exploit them), start by thinking deeply about how you can do this best

At the end, tell a few detailed stories about how you could build ASI in this context that exemplify specific loopholes.

Double check your work. Once you're done check that every point of your draft are consistent with every part of the proposed regulation. If not, go back and propose new ways to circumvent it. Stop only once you're confident there is no conflict between the regulation and your ideas.

Prompt 3

Which of those directions are the weakest points of this policy proposal:

<https://www.narrowpath.co/phase-0>

Evaluate the following directions and rank them by how important and true they can be. Some of those loopholes are good, others are bad.

Prompt 4

I have identified the following loophole in this "narrow path" policy:

- Orchestration of several models
- Compendium of small models
- combination of fine-tunes

Explain in more details to what extent they can be exploited (in phase 0 only).

My ideas to avoid exploitation of such loopholes:

- Forbid that a single company develops various models with different capabilities (inspired from antitrust rules)
- Forbid the combination of several models if the addition of the models reach the thresholds

Prompt 5

Yeah, so my goal is to explore those, to refine, make them more precise, and then put them in a format so that I can submit it to this Red Teaming Hackathon, and so have a list of flows and

proposals to submit that are interesting and useful for the judges and for the plan. The goal is to make this plan as strong as possible.

Before answering, think through your draft one more time to make it stronger, detect things that are incorrect, and opportunities to mentor me better in the 2 hours left for the hackathon. Don't tell me i'm good, make me and my proposition better.

Prompt 6

Great!. Could you write down a good writeup of those. Use weakness instead of loophole. Be specific (e.g. mention details that I mentioned or are in the proposal). Don't use —. Be concise and precise. To continue with Zvi: as precise as him, but slightly more concise.

For each weakness outline the solutions we've mentioned just after. Structure should be close to.

Weakness i

Evidence

Proposal(s)

Think about a draft, then critique it, and revise it. Do this twice.

Prompt 7

Okay, here's my first draft. Did I miss important things that I mentioned? Where could I make this stronger (more precise, more true, more specific...)? What would you encourage me to do, as the great mentor you are? Again, don't tell me my draft is great. I know it isn't, and it's surely not as great as it could be.

Prompt 8

What do you think of this idea :

Overall, the policies are not complete without investigation and enforcement powers of the competent authorities: regulations will not be complied with if the authorities do not have the power to audit, if whistleblowers are not encouraged to alert

Prompt 9

Focus on phase 0, what policy proposal would need verification by authorities? What powers should be given to the authorities for an efficient enforcement of the proposed policy (i.e. audit power, request documents, etc.)?

What example of such power already exist in US, EU or UK states ?

Prompt 10

Any last minute feedback. Can you check that the claims are substantiated? Can you draft me a precise abstract for this work? Do it as well as you know, as Zvi.

Prompt 11

I think you're overclaiming a bit at the start of the abstract, but I like the rest. Could you draft a better calibrated abstract (draft, critique, and revise it 3 times before answering). ultrathink