
Red Teaming A Narrow Path: A Critical Analysis

Shivam Arora, Ph.D.

Trajectory lab

With
ControlAI & Apart Research

Abstract

ControlAI has developed "A Narrow Path" - the first comprehensive plan to address extinction risks from Artificial Superintelligence (ASI). In this document we review, critique, and red-team *Phase 0: Safety* policies of the proposed plan. We found out that these policies are well-intentioned but lack sufficient grounding in technical understanding, resulting in significant gaps in their design and applicability. Our analysis highlights how, even when fully implemented, these policies leave room for strategic evasion that undermines their original purpose.

● Introduction

ControlAI has developed "A Narrow Path" - the first comprehensive plan to address extinction risks from Artificial Superintelligence (ASI). This is a three phase plan addressing: Safety, Stability, and Flourishing. In this document we review, critique, and red-team *Phase 0: Safety* policies of the proposed plan. Our goal is to strengthen the policies by identifying weakness, implementation challenges, and gaps.

Given the limited time for this sprint, we will focus on the first four policies in the document.

- **Methods**

We break down our analysis in three major parts. First we analyze the definition from a technical perspective. We examine the definitions and ideas presented critically focussing on implementation feasibility. Secondly, we use threat models to assess the effectiveness of policy. We cite various studies, policies, empirical experiments wherever applicable and available to our reasonings.

- **Results**

We want to start by acknowledging that writing effective policy is inherently challenging. It's rarely possible to capture every nuance or foresee all contingencies in advance, especially when default assumptions must stand in for lived complexity. We critique with intention to strengthen the policies.

We will go through each policy one by one:

Policy 1: Prohibit the development of Artificial Superintelligence

According to this policy,

“The development, creation, testing, or deployment of artificial superintelligence systems is prohibited. It is prohibited to knowingly participate in the development of, build, acquire, receive, possess, deploy, or use, any superintelligent AI.”

Here, Artificial Superintelligence is defined as: Any AI system that significantly surpasses human cognitive capabilities across a broad range of tasks.

Implementation feasibility:

The lack of metrics to define ASI render the definition open to interpretation making it hard to define the scope of the work.

The section addressing the scope in the policy document heavily refers to superintelligence itself creating circular reasoning, for example: “This prohibition extends to research aimed at producing artificial superintelligence, enhancement of existing AI systems that could result in artificial superintelligence, and the operation or transfer of superintelligence-related technologies”

- This makes it hard to concretely define the scope. And effectively makes it hard to distinguish between *improving the AI systems at any given time* and improving the AI systems for the *development of ASI*.

Most researchers such as Nick Bostrom¹ and others have hypothesized the emergence of superintelligence from the improvement of current AI systems with a sudden leap. Thus many AI systems can be simply a step away from ASI but still not qualify the definition of ASI, as stated in the document.

Policy effectiveness:

1. Under the assumption that this policy is adopted and implemented, any threat actors such as rogue AI labs with the mission to develop ASI systems, can reformulate their mission and research to be pertaining to improvement of their current AI systems. This way they can effectively be working with the goal to eventually develop ASI, essentially evading the goal of the policy by still following it.

2. Further, this policy incentivizes the AI labs to be *legally compliant but ethically evasive*. Many such restrictive policies related to Nuclear weapons, Genome editing, Chemical and Bio-weapons have resulted in labs and states working on these areas covertly². This behavior is even more high stakes in the case of AI since we rely on AI labs for understanding of new emergent behaviours and capabilities. And safety research which any inspection or regulatory body with need is not possible without their collaboration, making it necessary that no policy encourages the culture of secrecy and non-sharing of knowledge.

Suggestions:

If such a policy has to be implemented, we suggest the following alternative that captures the essence of policy while avoiding many of the limitations.

- We recommend that this policy be stated more precisely, with a set of benchmarks.
- The goal of those benchmarks should be to test certain capabilities that are considered near but a safe-distance away from ASI. We should restrict development of AI systems that surpass that benchmark. Since most AI companies have market incentive to claim the leadership for these benchmarks, it will ensure that we stay in the known regarding the increasing capabilities of AI systems.

¹ Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

² There are various sources to back-up this claim but I am sure an expert in this field can find a better reference for this than I can.

Policy 2: Prohibit AIs capable of breaking out of their environment

According to this policy, “AI systems capable of unauthorized access and the intentional development of AI systems with unauthorized access capabilities are prohibited”

Here unauthorised access is defined as: “Accessing a computer without authorization and/or exceeding the scope of authorized access, either to access information without permission, cause material harm or obtains something of value (e.g., compute time); in general, this should follow precedents created by the Computer Fraud and Abuse Act in the United States and its foreign counterparts.”

Implementation feasibility:

This policy prohibits AI systems with certain capabilities. This policy fails to take into account the technical aspects of these systems. Most AI systems being developed currently are *general-purpose systems* where by definition they are not trained for a particular use case but instead they are trained to be good at generalizing for general purpose tasks.

Policy effectiveness:

1. Under the assumption that this policy is adopted and implemented, any threat actors such as rogue AI labs with the mission to develop AI systems that are capable of unauthorised access, can simply train their models for a variety of tasks such as tool use, programming, reasoning etc which are benign tasks from the perspective of this policy but once trained these model, by design, will generalise to the capability to gain unauthorised access, cyber attacks.

A recent report by Google on Adversarial Misuse of Generative AI³ describes the use of their generative AI for cyber-attacks by state actors. This is a demonstration of the general capabilities of these models. Note that bad actors could elicit these capabilities despite guardrails. Hence rouge labs and bad state actors can easily escape this policy by training their models for general capabilities rendering this policy obsolete.

2. The policy states that “punishing those who willfully create harmful capabilities for an AI model to gain unauthorized access.”

This policy does not states the scope for the AI safety researchers in AI labs working for with the agenda of *Model organisms of misalignment*⁴, where part of the goal is to essentially train the model to do that.

Suggestion:

³Google, [Adversarial Misuse of Generative AI](#)

⁴ Hubinger, Nicholas Schiefer, Carson Denison, Ethan Perez [Model organisms of misalignment](#)

We believe that such capabilities and behaviour should be included in the first policy as a separate policy for a particular behaviour for *general-purpose* models is obsolete.

Policy 3: Prohibit the development and use of AIs that improve other AIs

This policy states that: “The direct use of found systems to build new found systems, or improve existing found systems, is prohibited.”

They define found systems as: Software programs which haven’t been written by hand by human developers, but which instead have been found through mathematical optimization.

Implementation feasibility:

Based on the given definition, all code produced by any given LLMs is a *found system*. Most big companies have already updated their tech stacks to include code editors that can write from a part of code to complete code. Most companies CEO have stated that more than 30% of their code is AI generated⁵⁶. Most tech companies aim for automation and AI is simply going to be a part of the link in this automation. This makes implementing this policy as stated hard to specify scope for.

Policy effectiveness:

Under the assumption that this policy is adopted and implemented, any threat actors such as rogue AI labs, who want to circumvent this process will hire token human employees for superficial oversight where most of the code is automated via AI.

Suggestion:

We believe a separate definition of *found system* is not warranted as it is equivalent to *AI generated code* and can add unnecessary confusion, unless the original intent for the definition was otherwise.

We believe that this is an important policy and with slight modification can be restricted to the most dangerous case of AI systems improving itself without creating a compliance

5

https://www.tomshardware.com/tech-industry/artificial-intelligence/microsofts-ceo-reveals-that-ai-writes-up-to-30-percent-of-its-code-some-projects-may-have-all-of-its-code-written-by-ai?utm_source=chatgpt.com

6

https://www.businessinsider.com/ai-code-meta-microsoft-google-llamacon-engineers-2025-4?utm_source=chatgpt.com

complexity and barrier for most tech companies. Although such policies are very hard to implement and monitor for compliance.

Policy 4: Require a valid safety case for deployment of AI systems

This policy states that before any AI system is trained or deployed, its developer must provide reliable safety cases showing convincingly that the planned AI system will not develop any capability that is legally prohibited or restricted in the current jurisdiction.

Here Safety case is defined as: An argument for why an Ai system will not do a prohibited action, which can be made before training, deploying, and running the AI.

Implementation feasibility:

The policy states that: “In the more advanced cases, it might be necessary to provide detailed mechanistic models of how the AI system works, for example for arguing that a SoTA LLM such as Claude or GPT-4 wouldn’t use any modeling of human psychology, since it definitely has the data, objectives, and incentives to learn how to do it and use it in practice.”

Safety cases are much needed but as stated in the policy, these requirements might be out of scope even theoretically. Even with sufficient progress in Mechanistic interpretability, we might not be able to make claims as stated in the policy for the safety case.

Policy effectiveness:

Such a policy if implemented would be quite effective at ensuring safety.

Suggestion:

Different tools and techniques might be necessary to provide a feasible safety case.

● Discussion and Conclusion

The current document has great potential to prevent ASI development for 20+ years. But it needs to be updated to improve the technical depth of the policies. We need good policies to prevent catastrophic and existential risks from advanced AI systems. But, we need to be cautious as a bad policy implementation can also have net negative impact for the safety from advanced AI systems in the longer run.

● References

- Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). A Narrow Path: How to Secure Our Future. Control AI. <https://www.narrowpath.co/>
- Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.
- Apart Research Hackathon Submission

Google, [Adversarial Misuse of Generative AI](#)

Hubinger, Nicholas Schiefer, Carson Denison, Ethan Perez [Model organisms of misalignment](#)

- Appendix

LLMs usage: Used to reformulate a few sentences grammatically for better clarity.