

Red Teaming A Narrow Path: ControlAI Policy Sprint

Gideon Daniel Giftson T

Independent

With

ControlAI & Apart Research

June 14, 2025

Abstract

All six policies are red teamed step-by-step systematically. We initially corrected vague definitions and also found that the policies regarding the capabilities of AI systems lack technical soundness and that more incentives are needed to entice states to sign the treaty. Further, we discover a lack of equity in the licensing framework, and a lack of planning for black-swan events. We propose an oversight framework right from the manufacturing process of silicon chips. We also propose calling for a moratorium on the development of general AI systems until the existing tools for analyzing them can catch up. Following these recommendations still won't guarantee the prevention of ASI for 20 years, but ensures that the world is on track to even tackle such a system if it is somehow created.

Keywords: AI policy, red teaming, governance, alignment, safety cases

1 Introduction

With artificial intelligence (AI) systems increasingly being used in safety-critical scenarios (Hill, 2021), it is of utmost importance that safety and alignment guarantees are reached about the system. While complex systems are put through rigorous safety tests before being deployed, in other engineering fields such as aviation, AI systems are being seemingly deployed with very few guardrails, and a marked lack of legislation.

ControlAI has taken the initiative to fill this gap using the Narrow Path proposal (Miotti et al., 2025). In this policy sprint, we choose to focus on the Phase 0 policies of the Narrow Path policy proposal, which are concerned with

preventing artificial super-intelligence (ASI), which is the end-goal of the top contemporary research labs, even though it is akin to opening Pandora’s box. We consider the ease of implementation of this proposal, and test whether it is circumventable. Finally, we take a stroll through history to learn lessons from similar policies, such as the Nuclear Non-Proliferation Treaty (Bunn, 2003). We find the limitations of the current proposal, and add how it can be modified to be more technically sound.

2 Methods

We initially looked at the soundness of the definitions in the policy documents. We also look at whether the policy can be ignored by motivated actors. Initially, the feasibility of implementing the policy is questioned, and then later the effectiveness of the policy is looked into. We then look for historical cases which are similar, and compare the policies. We also look at the issue from multiple perspectives.

3 Results

3.1 Policy 1 - Prohibit the development of ASI

3.1.1 Definitions

The first definition about artificial super-intelligence is highly vague. It needs to be more specific. The phrase - ”significantly surpasses” - can be easily manipulated by firms to claim that their new model is only a minor improvement over existing benchmarks in order to escape legislation. Similarly, the phrase ”broad range of tasks” is also vague. To solve this issue and reduce the ambiguity, the authors can define some set of tasks such as a benchmark dataset or a task analogous to the Turing test.

3.1.2 Implementation Feasibility

Given that this is largely a normative policy, which sets up a foundation, for future policies, it does it’s job well. However, it is questionable whether firms will be ready to give in to government oversight through auditing and monitoring activities.

3.1.3 Effectiveness of Policy

For the purpose of analysis in this report, we shall define a super-firm to be a company with nearly infinite resources hellbent on developing ASI.

In the last paragraph of page 16, the document states that only intentional attempts to develop ASI will be punished. This can be used as a loophole to claim ignorance by super-firms, if they’re caught developing ASI. In page 17, in the **Scope** section, the document claims that this policy doesn’t aim to

affect theoretical research. This is dubious since theoretical breakthroughs in understanding super-intelligence will make it easier to develop ASI, and a result, lower the threshold of compute and other resources required to create it.

3.1.4 Historical Precedent

While there are a multitude of historic precedents such as the non-proliferation treaty, biological weapons convention and the chemical weapons convention, the unique issue here is the ability to monitor and verify whether the signatories are actually following through with their commitments digitally. For example, the comprehensive nuclear test ban treaty (Kalinowski, 2006) can be easily monitored since a nuclear test above ground can be observed using the vast array of satellites, while a test below ground can be observed using seismic sensors. No such simple solution exists for ASI. This requires setting up controls in the supply chain for products such as graphical processing units (GPUs).

3.2 Policy 2 - Prohibit AIs capable of breaking out of their environment

3.2.1 Definitions

Definitions are found to be sufficient for this policy.

3.2.2 Implementation Feasibility

It is questionable as to how AI’s capabilities can be understood. Existing literature shows that there is no clear consensus on the way to track the capabilities of an AI system (Anwar et al., 2024). Literature also shows that capability benchmarks can be highly fickle and may not accurately evaluate the ability of the system to perform some task (Teney et al., 2020). Further, the effects of scaling compute to introduce new previously unseen capabilities, called as ‘emergent’ capabilities is also not well understood (Wei et al., 2022).

Therefore, it doesn’t make sense to propose legislation based on the capability of AI systems, since super-firms can just feign ignorance regarding the ability of the system to break out of its system in this particular case. There is also the major issue of **sandbagging** (van der Weij et al., 2024), where a language model can systemically underperform on evaluations. The bottom line, is that capabilities cannot be evaluated truthfully without further advances in interpretability.

3.2.3 Effectiveness of Policy

Now, given that we cannot properly define what are the capabilities of a given AI system, in order to make this policy effective, the solution is to use AI systems in highly sanitized environments. The AI system should be given very minimal chances to interact with the external environment. It should be deployed in a sandbox environment, so that even if it breaks free, the damage done is limited.

Significant care needs to be placed while drafting legislation regarding AI chatbots too since even if they are sandboxed, they can still be problematic due to their tendency to be sycophantic (Perez et al., 2023).

AI agents should also be banned since they're given enormous power and autonomy.

3.2.4 Historical Precedent

We can look to the sanitation procedures followed by medical labs doing research on diseases Iwen et al., 2015, to understand how to contain a foreign substance - limited contact! It makes more sense to limit our contact with AI systems, instead of expecting AI systems to somehow not be capable of breaking out of their environment - which at this moment is a hard theoretical problem - ascertaining whether an AI system has a capability or not.

It should be noted that fail-safes are a necessity here, since we can never guarantee our safety policy to be perfect. This includes creating kill-switches in routers to contain the AI to a specific network, or kill-switches in the systems themselves.

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

3.3.1 Definitions

The definition for **direct use** is vague about what constitutes a key step. It would be advisable to list a set of tasks such as proposing new model architectures, for example.

3.3.2 Implementation Feasibility and Effectiveness

A regulatory body such as the FDA require a vast amount of funding (FDA, 2025). Therefore, a similar amount of importance will have to be given for AI safety. Further, corruption will cause certain countries to have a lax implementation of these policies, which will lead to something akin to off-shore tax havens.

3.3.3 Effectiveness of Policy

It should be noted that a super-firm will just be able to run a sufficiently obscured internal team which is blind to external oversight. Such non-aligned actors can also make use of countries with lax laws (it is foolish to expect all states to have the same amount of strictness in implementing these policies).

3.3.4 Historical Precedent

Whistle-blowers must also be encouraged and given sufficient protection.

3.4 Policy 4 - Only allow the deployment of AI systems with a valid safety case

3.4.1 Definitions

It is unclear what are the prohibited actions stated in the description. A set of guidelines for what makes an action prohibited would make this definition more meaningful.

3.4.2 Implementation Feasibility

This policy calls for the developers to prove that a model won't use a particular capability, which is impossible with our current understanding of interpretability and system capabilities. Therefore, we are de-facto banning general AI systems, which makes sense theoretically, however, won't work practically. General AI systems such as GPT-4 already exist and open-source versions of other models exist too.

3.4.3 Effectiveness of Policy

This policy is somewhat effective, since we are basically banning general AI systems, and are confining the world to highly specific AI systems. However, existing issues such as sandbagging cause doubts regarding the true benchmark of a system's capability. This once again calls for improvements in our understanding of interpretability.

3.5 Policy 5 - A licensing regime and restrictions on the general intelligence of AI systems

3.5.1 Implementation Feasibility

While implementing a licensing regime makes sense, one should note the limitations imposed by corruption and bureaucratic delays especially in developing countries. electricity, and export controls

3.5.2 Effectiveness of Policy

While 10^{25} seems to be the current minimum amount of training required for AI systems to be capable of generalizing (Sevilla et al., 2022), it would make more sense to reduce the threshold a couple of magnitudes in order to account for future innovations in architecture as well as to account for the scenario where somehow a powerful model is trained without the license. The policy doesn't account for black-swan events.

3.5.3 Historical Precedent

Historically, the setting up of licensing has led to the marginalisation of underprivileged actors due to the amount of money required to maintain standards

(Jain, 2007). Therefore, grants need to be given in order to ensure that developing countries are also able to develop AI systems without being limited by the licensing regulations.

3.6 Policy 6 - An international treaty establishing common red lines on AI development

3.6.1 Implementation Feasibility

It must make sense for a state to join the treaty. Near-guarantees should be provided to ensure that they won't be left behind. Export controls are a necessity. They need to be implemented from the beginning of the supply chain. Since a super-firm will just make its own chips to circumvent GPU hoarding restrictions, patterns of mining and export of silicon and other rare earth materials are necessary. Subsequent monitoring of GPU manufacturing and tracking through GPS chips and serial numbers is also a must. Further, data centres should also be monitored using power usage, and cooling system capabilities. A fully monitored supply chain will make it virtually impossible to circumvent this treaty, and as a result, states will be forced to sign the treaty.

3.6.2 Effectiveness of Policy

However, we must also account for unknown mineral deposits, such as during the second world war, when the British imported uranium from the Congo (Fleckner & Avery, 2005). Therefore, the policy might not be 100% effective, but it might be near perfect.

3.6.3 Historical Precedent

We have seen that non-proliferation treaty, although very good, has not been perfect (countries such as India for example, have become a nuclear power). It is crucial to understand that our policy will never fully guarantee the existence of ASI within the next 20 years. Therefore, research needs to be carried out on ways to mitigate a possible ASI, and ways to contain it. The need of the hour is alignment and interpretability research, so that any AI system, super-intelligent or not, is working for us, and not against us.

4 Discussion and Conclusion

The current policies along with the suggestions in this document, have a very good chance of preventing ASI for about 20 years. In my opinion, the weakest links are the policies on preventing AIs from breaking out and preventing AIs from improving other AIs. Since we cannot accurately understand what are the capabilities of a given AI system at present, it is impossible to ascertain whether an AI system will be incapable of those two things. Therefore, instead

of allowing such systems to be even deployed, it is crucial to ensure that current AI deployment is limited only to highly specialized cases.

Given that the current AI development cycles tend to very rapid, it is crucial that we reach a moratorium on the development of general AI systems, since we don't possess the understanding required to test the capabilities of an AI system. Huge leaps are required in interpretability (existing literature works only for toy models in my opinion (Nanda et al., 2023)) and alignment theory before research on generalized AI systems can restart safely. As seen in the non-proliferation treaty, actors and non-signatories definitely have a chance to develop ASI, so research also needs to be done to think about ways to contain such an AI system.

References

- Anwar, U., Saparov, A., Rando, J., Paleka, D., Turpin, M., Hase, P., Lubana, E. S., Jenner, E., Casper, S., Sourbut, O., et al. (2024). Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.
- Bunn, G. (2003). The nuclear nonproliferation treaty: History and current problems. *Arms Control Today*, 33(10), 4–10.
- FDA. (2025). FY 2025 budget summary.
- Fleckner, M., & Avery, J. (2005). Congo uranium and the tragedy of hiroshima. *55th Pugwash Conference, Hiroshima, Japan*.
- Hill, S. A. I. (2021). Bail reform and the (false) racial promise of algorithmic risk assessment. *UCLA L. Rev.*, 68, 910.
- Iwen, P. C., Smith, P. W., Hewlett, A. L., Kratochvil, C. J., Lisco, S. J., Sullivan, J. N., Gibbs, S. G., Lowe, J. J., Fey, P. D., Herrera, V. L., et al. (2015). Safety considerations in the laboratory testing of specimens suspected or known to contain ebola virus.
- Jain, R. (2007). Access to leisure: Impact of forces of liberalisation, privatisation and globalisation on socially disadvantaged groups in india. *World Leisure Journal*, 49(1), 15–23.
- Kalinowski, M. B. (2006). Comprehensive nuclear-test-ban treaty verification. In *Verifying treaty compliance: Limiting weapons of mass destruction and monitoring kyoto protocol provisions* (pp. 135–152). Springer.
- Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2025). A narrow path.
- Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. <https://arxiv.org/abs/2301.05217>
- Perez, E., Ringer, S., Lukosiute, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434.
- Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. *2022 In-*

- ternational Joint Conference on Neural Networks (IJCNN)*, 1–8. <https://doi.org/10.1109/IJCNN55064.2022.9891914>
- Teney, D., Abbasnejad, E., Kafle, K., Shrestha, R., Kanan, C., & Van Den Hengel, A. (2020). On the value of out-of-distribution testing: An example of goodhart’s law. *Advances in neural information processing systems*, 33, 407–417.
- van der Weij, T., Hofstätter, F., Jaffe, O., Brown, S. F., & Ward, F. R. (2024). Ai sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*.
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Appendix

I felt that it was somewhat ironic to use AI to red-team how to control AI. Therefore, I abstained from using AI for the purposes of this sprint.