
A Narrow Line Edit: ControlAI Policy Sprint

Aidan Kierans
University of Connecticut

Alexandra Bradley
University of Connecticut

With
ControlAI & Apart Research

Abstract

Rather than explore specific policy questions in depth, we analyzed the presentation of the “Narrow Path” Phase 0 proposal as a whole. We considered factors like grammar, style, logical consistency, evidential support, comprehensiveness, and technical context.

Our analysis revealed patterns of insufficient support and unpolished style throughout the proposal. Overall, the proposal failed to demonstrate the rigor and specificity that is typically found in effective policy proposals. With effort to address these oversights (aided by our thorough annotations), the proposal could be significantly improved. These changes will also allow for deeper, narrower policy analysis to be integrated more effectively than is currently possible. For this reason, we expect our findings to multiply the efficacy of this policy sprint.

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

• Introduction

While reading the Phase 0 proposal to decide which policies to analyze, we found that many parts of the proposal were insufficiently developed to benefit from deep policy analysis. Our hypothesis is that, in order to influence policymakers to advance good policy, proposals shared with them should be polished, well-argued, clear in their recommendations, and backed up by arguments and citations that anticipate objections.

Thus, we expect the Narrow Path proposal will be more successful as a communication tool if we show how to polish it and which arguments are incomplete. Likewise, wherever it is overly light on citations or overlooks common objections, we can make the policies themselves more successful by pointing these issues out.

There were many good ideas in the Introduction and the Phase 0 of this report. That said, we were also able to find many opportunities for improvement throughout the Introduction and Phase 0 of the report. As a result of this we pivoted to giving feedback on the writing of these sections as a whole, including suggestions to improve readability, brevity, and more assertive and accurate language. We believe that after implementing our recommendations and raising the standards of clarity and rigor throughout the proposal, deeper analysis of the policies' potential successes and failures will become more tractable. Until those changes are made, thorough attempts to address specific gaps in the policies may be difficult to combine, since they likely address different or shifting understandings of what the proposal currently recommends.

Furthermore, if the rigor of the proposal is improved according to our recommendations, we expect policymakers (and their staffers) to be more amenable to the proposed policies. We also expect them to understand the proposal better and feel that their time and attention are being respected. These factors are critical to effective policy advocacy -- if Control AI succeeds in these areas while arguing for effective policies, lawmakers are significantly more likely to hear and address their concerns appropriately.

● **Methods**

Our planned methodology was to read through the policy proposal, make note of potential topics for improvement or deeper response, and then choose one topic for a detailed and well-researched response. However, we found many more gaps and areas needing development than we expected. Due to our respective backgrounds in technology policy and writing, we felt that the prevalent stylistic errors and structural issues that kept distracting us would likely have a similar effect on the intended audience.

To address this pressing issue, we combed through the Introduction and Phase 0 of the Narrow Path proposal, commenting on stylistic issues as we found them. In order to cover more ground in this single-day sprint, we focused on highlighting instances of recurring stylistic oversights in their first few instances rather than every single occurrence. Likewise, to enable the structural issues to be considered in aggregate, we focused on identifying gaps throughout all of Phase 0. This was made possible by drawing from our relevant background knowledge rather than supplying detailed investigation for each gap we found.

● **Results**

Our detailed analysis is linked in the Appendix. In general, we observed the following patterns:

1. Implementation Feasibility Gaps:
 - a. Many of the proposed policies appear to require vast, sweeping action from lawmakers. These actions are unlikely to be taken without specificity as to who is responsible for what. Without stating which institutions should handle each

element—distinguishing, for example, between legislative and executive actions, laws and standards, and international differences in such institutions—no institution is likely to take responsibility. A better proposal would list appropriate institutions by name wherever relevant. By doing so, the authors would make the practical steps of implementing these policies appear less onerous.

- b. Some of the technical capabilities that these policies seek to limit, such as self-improving AI, are already possible with existing software. Regulators are generally hesitant to impose rules that limit existing practices that have yet to cause harm, especially in technology policy. These limits are thus unlikely to be adopted without clarification to exclude existing methods.
2. Policy Effectiveness Weaknesses:
- a. We found that many policies had too few definitions for the terms they used. Generally speaking, regulations must carefully define even “common” terms in order to limit interpretations that would evade lawmakers’ intentions. Since the proposal aims to guide international regulation, it is doubly important that its definitions are airtight, as any variations in interpretation between nations could open regulatory loopholes.
 - b. Due to the other gaps in implementation feasibility and evidential support that we noted, we anticipate significant barriers to the careful and uniform adoption of these policies that would be necessary to prevent circumvention. This is exacerbated by the shallow implementation details of nearly every policy described.
 - i. Policy 5, which described licensing regimes, was more detailed than the rest by far. Each of the other policies could likely be expanded to that length without adding filler.
3. Evidence-Based Assessment:
- a. There were numerous claims of fact unsupported by evidence. For example, in the section titled “No AIs Improving AIs,” the paper claims that “AIs competent enough to...enact improvements on themselves...can quickly enable runaway feedback loops that can bring the AI system... to levels of competence and risk far beyond those intended.” Evidence would assist this claim in seeming less bold.
 - b. Likewise, this proposal claims that “AI systems with the ability to break out of their intended boundaries can quickly evade human control and monitoring,” and then provides the examples of self replication and bootstrapping. This is fine but it would be better supported with reputable studies or examples.
 - c. Throughout the proposal, the reasoning for the choices of specific numbers are often not justified. For example, it is not clear why the authors chose “20 years” rather than 15 years or 25 years, which would both be plausible alternatives as “the minimum time frame to construct our defenses, formulate our response, and navigate the uncertainties to gain a clearer understanding of the threat and how to manage it.”

● Discussion and Conclusion

In essence, a more effective version of this proposal would include claims supported by evidence and stylistically stronger language. A professional tone and evidence-based argumentation can be more convincing to government entities, and strong evidence in particular can make it more

difficult for opposing political forces to argue against this proposal's ideas. Historically, political arguments claiming a (real or imagined) lack of scientific support for regulation have been used to argue against regulating at all -- even preventing regulations that would require the production of such evidence (in the form of documentation and research). This historical failure of tobacco and oil regulation is already one that risks repetition in AI policy (Casper, 2025), and this proposal risks falling into the same trap in its current form.

Particular evidence to bolster this proposal can include references and research from the references below, as described in the annotation document. These are only a few examples, however; the proposal could certainly use more substantial and varied evidence. We expect that with these additions and further line editing for clarity and tone, this proposal can be made even stronger.

As it stands, the proposal is unlikely to achieve the aims of Phase 0 by preventing 20 years of ASI development. In its current form, it might contribute to shaping general priorities in audiences already receptive to this policy direction. However, due to the time pressures and specialized duties of those who work in policy, successful proposals must be very clear about who is responsible for each part of each policy. These time pressures also mean that policymakers (or their staff) must quickly assess the quality, completeness, and evidential support for any proposal they consider—often on the basis of established writing styles and structural elements. By falling short of these expectations and heuristics, the current proposal is likely to fail high-stakes “sniff tests” necessary for practical consideration. Our recommendations throughout the attached Phase 0 Annotations document are targeted to address this challenge.

● References

Casper, Stephen, David Kreuger, and Dylan Hadfield-Menell. “Pitfalls of Evidence-Based AI Policy.” *arXiv preprint arXiv:2502.09618* (2025).

Kierans, Aidan, Kaley Rittichier, and Utku Sonsayar. "Catastrophic Liability: Managing Systemic Risks in Frontier AI Development." *arXiv preprint arXiv:2505.00616* (2025).

Stegenga, Jacob. *Medical nihilism*. Oxford University Press, 2018.

Weil, Gabriel. "Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence." *Available at SSRN 4694006* (2024).

Zweiger et al. "Self-Adapting Language Models." *arXiv preprint arXiv:2506.10943* (2025).

● Appendix

No language models were used in the writing of this response. Our detailed responses to the Narrow Path proposal can be found in the supplementary documents attached to our submission. They can also be accessed with these links:

Detailed annotation of the Introduction: [W Narrow Path Introduction Line Editing.docx](#)

Detailed annotation of Phase 0: [☰ Narrow Path Phase 0 Annotations](#)