
Red Teaming A Narrow Path: ControlAI Policy Sprint - AI⁴: Allow It!

Anthony Bailey
anthony@PauseAI.info
(but opinion is my own)

With
ControlAI & Apart Research

Abstract

A Narrow Path currently includes a condition (A): “No AIs improving AIs” that underlies various parts of the document.

It makes no exception for what I will abbreviate as AI⁴: AI Assisting In AI Alignment Improvement. It should, because sufficiently many in AI safety while acknowledging its unique hazards still see value in exploring “AI helping with our alignment homework.”

Specifically I am Red Teaming Phase 0 policy 3 (“Prohibit the development and use of AIs that improve other AIs”) arguing no policy that bans AI⁴ will garner sufficient support to be agreed or enforced.

A variety of Deep Research implementations suggested 70-90% of those expressing relevant opinions in forums associated with AIXR concern would oppose such a ban.

If AI assistance in AI alignment improvements is to be allowed, that needs to be made more clear, and consequences to licensing and enforcement considered.

● Introduction

One of the fundamental safety conditions used throughout A Narrow Path Phase 0 is introduced as:

(A) “No AIs improving AIs”

It is used and refined throughout the document, particularly in policy 3: “Prohibit the development and use of AIs that improve other AIs”

Many folk agree that AI-assisted AI research & development is a likely way for AGI to swiftly become superintelligence. The likelihood of an unsafe intelligence explosion through these means keeps us up at night. A policy to mitigate this risk is vital. No argument from me here.

But: consistently across the document it is implied that the policy intends to include what is pejoratively (but fairly) called “AI helping with our alignment homework.” (Throughout this document I will use a novel abbreviation **AI⁴** derived from “AI assisting in AI alignment improvement”).

My fundamental concern is that banning AI⁴ in particular most likely does **net harm** to our survival chances: that there is a rough consensus that this path should *not* yet be closed.

{Interlude: If as an individual you disagree, I get it. But please read e.g. [Can we safely automate alignment research?](#), a 2025 essay from Joe Carlsmith. It is illuminating: focused on building an “alignment MVP” - AI systems assisting with alignment research at or above top human expert level - as a potentially achievable waystation that can unlock solving full superalignment. He suggests that automating relatively safe narrow “empirical” alignment research can help humans develop better evaluation methods for “conceptual” alignment research. He holds no illusions about the “extremely serious” challenges such as potential AI scheming (which we’d need to prevent, detect, or carefully work around), and resource constraints (time, data, investment) but advocates doing the necessary work.}

In the context of this sprint I am red teaming the application of policies from the plan and identifying failure modes. The sprint calls for identifying threat models and actors. My slant here is that unless policy explicitly allows AI⁴, it will receive very little support even from within the AI safety communities most concerned with AIXR, and therefore it will simply not be accepted as sensible by lawmakers and implemented: this is the main failure mode. A related concern is that if it were implemented, enforcement would be hard because so many people believe that AI⁴ is the best path: they would continue sufficient work in the dark that we should instead encourage in the light.

I chose this as my solo red-teaming effort based on personal experience of this consensus on the value of AI⁴ in the field, but I did go on to attempt to gather quantitative evidence. Deep Research attempts across four environments all find

70-90% of those expressing relevant opinions in forums associated with AIXR concern would oppose a policy that banned AI⁴.

Details (methods and results) derivable from specific requests in the LLM appendix: I will summarize here given time. concern would oppose a policy that banned AI⁴.

I also offer examples from history where broad-strokes regulation generated resistance and unintended consequences and failed where lighter regulation carefully implemented ultimately succeeded.

If AI assistance in AI alignment improvements is to be allowed, I am relieved - but that needs to be made more clear, and knock-on consequences to licensing and enforcement considered.

• Methods

This was a one-person effort against the clock, but I tried!

I present quantitative analysis of expressed opinions sourced from Deep Research affordances, and historical precedents for lighter touch regulation succeeding where heavier regulation failed. I've also considered new threat models caused by my proposed change.

• Results

Historical precedents. I would much rather present cases where successful regulation resulted from adding a particular exclusion to the broad regulation. I ran out of time to find them. Instead, a more obvious pattern in these examples.

Concern	Heavy failure	Lighter success
Effect of alcohol on health	Prohibition: complete ban rejected through widespread bootlegging and organized crime.	Post-repeal licensing, taxation, and age restrictions achieved goals more effectively.
Sulphur emissions	Initial rigid technology mandates faced industry resistance and high costs.	1990 Clean Air Act's sulfur dioxide cap-and-trade system faster and cheaper than projected.
China population	One-child policy strictly enforced created demographic crisis and human rights issues.	Two-three child relaxation maintained population goals while reducing social costs.

Consensus in the field. I refined a deep research request to establish whether folk on “our side” of AI safety would accept policies that disallowed AI⁴. I then tried it in four different contexts.

Details of implementation in linked reports in references.

Many of these reports go into detail re their assumptions and methodologies, and present other breakdowns of the data. I encourage reading!

Source	Key finding
OpenAI GPT-4o in research mode	In sum, we estimate roughly 70% of active alignment-focused contributors (categories 1–3 combined) would <i>oppose a strict ban</i> on using AI to improve AI, whereas around 10% would actively <i>support a ban</i> . The remaining ~20% are uncertain or have nuanced views. The true proportions could plausibly vary by ± 10 percentage points, but it appears a clear majority leans against an outright ban on AI ⁴ ,
Gemini Pro 2.5(?) in research mode	Aggregate Opposition Category: Would Oppose Ban Estimated Percentage: 85% Confidence Interval: $\pm 15\%$
Claude Opus 4 research (fresh session)	While researchers express varying levels of caution about recursive improvement risks, virtually none advocate prohibiting AI ⁴ even for safety research. The proposed policy would likely face strong opposition from 75-85% of active Alignment Forum contributors, with the remainder mostly uncertain rather than supportive.
Claude Opus 4 research (in session where request was first created.)	With 85-95% opposition to strict AI⁴ bans , the AI alignment community demonstrates remarkable

	consensus that prohibiting AI-assisted AI development would be counterproductive for safety. This position spans from those who see it as already essential (25-30%) to those viewing it as a valuable option (15-20%), with virtually no support (<5%) for restrictive policies. The community instead emphasizes developing safety frameworks that can harness AI assistance while managing recursive improvement risks.
--	--

New threat models. Making them explicit to reduce likelihood folk think I'm unaware: current consensus above assumed to incorporate knowledge of same.

Labs can “smuggle” capability work as alignment research work.

Accidental discovery of capability mechanisms through alignment research work.

Greater chance of scheming if models perform conceptual alignment research work.

- **Discussion and Conclusion**

“Would the current Phase 0 policies, as written, actually prevent ASI development for 20+ years? What are the most critical vulnerabilities? How could these policies be strengthened before implementation?”

My summary is that as written policy 3 would sufficiently lack support that it would not make it into law in the first place. Non-existent / unenthusiastically enforced policies are clearly critically vulnerable. Exceptions for AI⁴ make it more likely policy is passed and enforced.

- **References**

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

Carlsmith, J. (2025). Can We Safely Automate Alignment Research? <https://joecarlsmith.com/2025/04/30/can-we-safely-automate-alignment-research>

Deep Research reports:

- Report produced by Claude Opus 4 at the end of [a session](#) where the common research request text used in other cases was created and refined: <https://claude.ai/public/artifacts/cb5380e6-c918-4105-96c3-580324f84b8c>
- Claude Opus 4 when given request text without prior session context:

<https://claude.ai/public/artifacts/2c4360bf-aede-4647-872b-5b2b80b1eb77>

- Gemini Deep Research when given request text:
<https://docs.google.com/document/d/1gRwz8eUugjjhD4dcFW5Seb0vjbV-gT3-3GJE4WTeRG0/edit?usp=sharing>
- GPT-4o in research mode when given request text:
<https://chatgpt.com/share/684c9993-d970-8007-a68a-9df45668232c>

- **Appendix**

- Prompts used while using LLMs

Some further detail in the linked Deep Research reports in the references section above.

This is the deep research request text I used to derive my results:

Research Request: Quantifying AI Alignment Community Opposition to Strict Ban on AI-Assisted AI Development

Core Question: What proportion of active Alignment Forum contributors would oppose a policy that prohibits using AI systems (created through mathematical optimization) to directly improve other AI systems, where "direct improvement" includes:

- Generating training data
- Hyperparameter search
- Architecture design
- Any key step in AI development beyond trivial coding assistance

Key Context:

- "AI^4" = AI-Assisted AI Alignment/Advancement
- Policy would ban practices currently used at frontier labs
- Only permitted AI use: typo correction, writing small functions, general information search
- Ban applies to all AI improvement work, including safety/alignment research

Research Objectives:

1. **Quantify positions from:**
 - Alignment Forum posts/comments

- LessWrong AI safety content
- Public statements by alignment researchers
- Safety team publications
- 2. **Position categories:**
 - **Essential Now** (AI⁴ already necessary for competitive alignment work)
 - **Becoming Essential** (AI⁴ will be necessary within 2-3 years)
 - **Valuable Option** (AI⁴ helpful but not essential)
 - **Neutral/Unclear**
 - **Support Ban** (AI⁴ too risky even for alignment)
- 3. **Key context to capture:**
 - Is the researcher already using AI⁴ in their work?
 - Do they believe major labs depend on AI⁴?
 - Timeline beliefs correlation?

Deliverables:

- Percentage breakdown with confidence intervals
- Estimated "would oppose" aggregate (categories 1-3)
- Representative quotes
- Methodological limitations