
Red Teaming Policy 5 of A Narrow Path: Evaluating the Threat Resilience of AI Licensing Regimes

Desmond Gatling
Western Governors
University

Wyatt Snyder
Hope College

With
ControlAI & Apart Research

Abstract

This report presents a red teaming analysis of Policy 5 from A Narrow Path, ControlAI's proposal to delay Artificial Superintelligence (ASI) development through national AI licensing. Using a simplified PASTA threat modeling approach and comparative case studies (FDA, INCB, and California SB 1047), we identified two critical failure modes: regulatory capture and lack of whistleblower protections.

We developed a custom policy CVSS framework to assess cumulative risk exposure across each case. Due to time constraints, we used ChatGPT-assisted simulation to complete the results section and illustrate potential findings from our scoring method.

Our analysis suggests that, as written, Policy 5 is vulnerable to institutional influence and lacks sufficient safeguards to ensure enforcement. We recommend clearer accountability structures, built-in whistleblower protections, and stronger international coordination to make the policy more resilient.

Introduction

Our submission will focus on Policy 5 of the Phase 0 policy framework outlined in A Narrow Path. Policy 5 focuses on “developing Licensing Regimes and placing restrictions on the general intelligence of AI systems” (Miotti et al., 30). Particularly this section of ControlAI’s policy framework proposes three licenses depending on the nature of AI development.

- A. Training License (TL) -All AI developers seeking to train frontier AI models above the compute thresholds of 10^{25} FLOP must apply for a TL and have their application approved prior to training the proposed model.
- B. Compute License (CL) - All providers of cloud computing services and data centers operating above a threshold of 10^{17} FLOP/s must obtain a license to operate these and comply with specific know-your-customer regulations as well as physical GPU tracking requirements.
- C. Application License (AL) -Any developer seeking to use a model that has received an approved TL and will be expecting to make major changes, increases, or improvements to the capabilities of the model as part of a new application will need to apply and be granted an AL (Miotti et al., 30).

This policy contributes to the ultimate goal of preventing the development of Artificial Superintelligence within the next 20 years by “limit[ing] the expected general intelligence of AI systems” (Miotti et al., 9). It allows for regulating bodies to preserve the benefits of AI development while mitigating risks from larger and more volatile models (Miotti et al., 30).

Developing licensing regimes similar to those outlined in Policy 5 of A Narrow Path is critical for preventing ASI development because it creates a clear global understanding of how to responsibly develop AI technologies. It also enables straightforward protocols for enforcing compliance internationally among national and independent actors.

In order to evaluate the efficacy of Policy 5, we have chosen to utilize a combination of the Process for Attack Simulation and Threat Analysis (PASTA) threat modeling framework and historical and political comparative analysis methods. By using these methods we identified vulnerabilities within potential state and corporate AI developers, proposed two failure modes to investigate throughout our analysis, and identified ways to strengthen policies that make up ControlAI’s Direct Institutional Plan.

The PASTA threat modeling framework is “a risk-focused 7-step threat modeling methodology” (Sreekumar) that is well suited for assessing AI regulations vulnerabilities because it links regulatory requirements to concrete threat scenarios. We used portions of this threat modeling framework to analyze Policy 5 alongside three case studies that consist of historical and contemporary examples of international regulatory bodies and regional efforts to regulate and license powerful emerging technologies.

In our study, we analyzed two key failure modes: **regulatory capture** and the **failure to protect or recognize whistleblowers**. Regulatory capture occurs when a regulatory body becomes influenced by and acts in favor of the private interests it is meant to oversee, instead of serving public interests. ControlAI's existing framework already consisted of a robust rationale section that explained the merits of its regulatory and licensing regimes. However, the policy framework made little mention of industry influence over regulators. We identified regulatory capture as a failure mode, and took actions to address this vulnerability that was previously unmentioned. Additionally, we identified failure in protecting and or recognizing whistleblowers as a second fail mode. Whistleblower protections were also notably missing from Policy 5's original framework, and our study intends to examine ways to include these safeguards to encourage reporting of violations.

Methods

Our study consisted of a combination of analytical and research approaches. We started by using a simplified PASTA threat modeling framework to guide our analysis.

For the scope of our project, we used portions of the modified PASTA threat modeling framework to define the objective of ControlAI's policy framework as well as developing a red teaming objective that was used to test for vulnerabilities. The objective of ControlAI's policy framework is to delay or prevent the development of Artificial Superintelligence (ASI) for at least 20 years through regulatory frameworks that limit the general intelligence of AI systems (Miotti et al., 9). Our analysis uses a red teaming approach to identify vulnerabilities within Policy 5 that can be exploited by state corporate actors(GPTprompt).

How PASTA Works (7 Stages, Simplified)

1. **Define Objectives** – Understand business goals and security requirements.
2. **Define Technical Scope** – Outline systems, apps, and data in scope.
3. **Decompose the System** – Break down architecture (data flows, trust boundaries).
4. **Analyze Threats** – Identify threats (using sources like threat intelligence or past incidents).
5. **Vulnerability Analysis** – Find weaknesses attackers could exploit.
6. **Attack Modeling** – Simulate attack paths (e.g., how an attacker might breach a user account).
7. **Risk Analysis & Mitigation** – Prioritize risks based on impact/likelihood and apply countermeasures.

(GPTprompt)

From there, we drafted a Common Vulnerability Scoring System in order to determine a scale assessing relative risks associated with each individual case ([CVSS source](#)). For the purpose of our study, we will refer to this as the red team CVSS matrix.

We then created three case studies to analyze alongside policy 5. Each case study reflected an institution, bureaucratic, and technologic example of regulations and/or licensing at an international scale.

Finally, each case study and Policy 5 were evaluated with our red team CVSS matrix to highlight potential risks that may have triggered our two fail modes (regulatory capture, and failure to protect whistleblowers).

Based on the cumulative tally of our scoring system, we are able to determine a risk assessment value to Policy 5 and to each case study.

● Results

This section is intended to outline our findings from our Comparative Analysis between Policy 5 and each of the three case studies. Each item would be assessed using our red team CVSS matrix. As an example, we used an AI generated simulation to illustrate an example of the findings we intended to find by running our red team matrix exercise.

([CVSS Results Matrix](#))

● Discussion and Conclusion

Synthesize your analysis to answer the core question: Would the current Phase 0 policies, as written, actually prevent ASI development for 20+ years?

Based on the results of our project thus far, we believe do not believe that Phase 0 policies as written actually will prevent ASI development for 20+

years.

Specifically with regards to our focus, which was policy 5 and related to regulatory and licensing regimes. For example, we recently witnessed how California SB 1047 was vetoed by Governor Gavin Newsome after garnering support in the California State legislature. This bill was important for raising awareness and concern for regulatory considerations, however it faced many political challenges and was not ultimately passed.

While it was a milestone for the AI Safety community, it highlighted a major obstacle in preventing ASI development. This lack of consensus for where to draw the threshold on limiting the power and potential of AI development leads to a climate where influential developers can sway public opinion by citing how AI regulations will stifle innovation and lead to a loss of a country's competitive advantage. Without a campaign to garner support for limiting the development of powerful and potentially harmful ASI's we fear that private interests will continue to exploit opportunities for regulatory capture of regulating bodies. In this example, it was California SB 1047. And if future regulatory and licensing regimes fail to consider this potential risk, it will just be a larger international regulatory body that is compromised by private interests.

What are the most critical vulnerabilities? How could these policies be strengthened before implementation?

Issues with the current proposal are especially prominent in dealing with potential regulatory capture issues. Given the variety of factors pushing private companies towards rapid AI development, as well as the risk analysis done in the present study, it is reasonable to expect private companies, and compliant nations, to take actions circumventing established legislation related to licensing. Both the PASTA framework and historical precedents such as the Toshiba incident suggest that these risks must be further addressed before confidence in humanity's existential security can be established.

Consider the broader implications for AI governance: How do your findings relate to the urgency of implementing these policies given current AI

development timelines? What lessons from historical regulatory precedents are most applicable?

● **References**

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

Sreekumar, N. (2023, May 1). *Types of Threat Modeling Methodology*. Practical DevSecOps. <https://www.practical-devsecops.com/types-of-threat-modeling-methodology/>

White & Case LLP. (2024). AI watch: Global regulatory tracker - United States. <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states>

Wikipedia contributors. (n.d.). *Baruch Plan*. In *Wikipedia, The Free Encyclopedia*. Retrieved June 14, 2025, from https://en.wikipedia.org/wiki/Baruch_Plan

Wikipedia contributors. (n.d.). *International Drug Control Conventions*. In *Wikipedia, The Free Encyclopedia*. Retrieved June 14, 2025, from https://en.wikipedia.org/wiki/International_drug_control_conventions

Wikipedia contributors. (n.d.). *Toshiba–Kongsberg scandal*. In *Wikipedia, The Free Encyclopedia*. Retrieved June 14, 2025, from https://en.wikipedia.org/wiki/Toshiba–Kongsberg_scandal

● **Appendix**

Add any extra content you wish here! Unrestricted.

- Additional threat model scenarios
- Prompts used while using LLMs
- Detailed historical precedent analysis
- Extended policy comparison tables
- Technical implementation details for licensing systems
- International coordination framework assessments
- Supplementary evidence and data sources

Policy Comparison Tables

Tables comparing NarrowPath’s Licensing Policy with other existing and proposed policies (GPTPrompt)

Criteria	FDA (Drugs)	INCB (Narcotics)	Policy 5 (AI Licensing)
Regulatory Body	Centralized national agency	Multinational treaty body	National AI regulator (domestic)

Scope	Drug development & approval	Narcotic drug production, trade, use	Advanced AI development, deployment, compute
Licensing Tiers	Preclinical → Phase I-III → Post-market	Production quotas, authorized import/export	Training → Compute → Application
Enforcement Power	Strong domestic authority, limited globally	Weak, depends on national compliance	TBD — depends on implementation & funding
Successes	Life-saving drug approvals, consumer protections	International norms around drug scheduling	(Intended) Delay or prevention of ASI
Failures / Risks	Regulatory capture, data manipulation, uneven enforcement	Politicization, uneven compliance, race dynamics	Potential for shell labs, coercion, vague thresholds
Key Historical Lesson	Licensing regimes are manipulable without strong transparency & public accountability	International consensus is fragile and vulnerable to geopolitical pressure	AI policy must be resilient to both domestic capture and global race conditions



Comparison Table: AI Governance vs. Historical Regulatory Models

Category	Policy 5 – A Narrow Path (AI Licensing)	FDA (Drug Regulation)	INCB (Narcotics Control)
Jurisdiction	National (UK, US, other sovereign states)	U.S. federal government	International (UN treaty-based)
Primary Objective	Prevent or delay the emergence of Artificial Superintelligence (ASI)	Ensure safety and efficacy of pharmaceuticals	Limit production and use of narcotics to medical and scientific purposes
Regulatory Entity	Proposed national AI regulator (not yet defined in structure)	Food and Drug Administration (FDA)	International Narcotics Control Board (INCB)
Licensing Framework	3-tiered: Training → Compute → Application	Phased: Preclinical → Clinical Trials → Post-Market	Country-based quotas and approval for production/import/export
Scope of Application	Applies to the most capable AI systems (general intelligence threshold)	Applies to all pharmaceuticals within the U.S.	Applies to narcotics under international treaty
Monitoring Mechanism	Continuous AI research monitoring (details TBD)	Clinical trials, manufacturing inspections, post-market surveillance	National self-reporting, compliance monitoring, trend analysis
Whistleblower Protections	Conceptually important but undefined; vulnerabilities identified in proposal	Limited formal protections; ad hoc mechanisms (e.g., FDA whistleblowers)	None institutionalized; dependent on state mechanisms

Enforcement Power	Enforcement assumed at national level; variable by capacity	Strong domestic authority to approve, restrict, or recall drugs	No direct enforcement power; depends on treaty compliance and national action
Transparency Requirements	Implicit via licensing; details not specified	Mandatory disclosures during drug trials and post-market reviews	Varies by nation; limited transparency in reporting enforcement or violations
International Coordination	Encouraged but not yet implemented; presumed cooperation between states	None (U.S.-only, but FDA model sometimes emulated globally)	Built on global treaty structure with nearly universal participation
Potential Weaknesses	Regulatory capture, ambiguous thresholds, uneven enforcement, coercion, race dynamics	Influence by pharmaceutical lobby, fast-track loopholes, poor post-market responsiveness	Political weaponization, unequal enforcement, circumvention by powerful states
Strengths	Ambitious structural model with global aspirations; proactive framing of frontier risk	Structured pipeline; real authority; consumer safety impact	Long-term institutional memory; precedent for global standards in high-risk domains

Updated Comparison Table: Policy 5 vs. California SB 1047

(Using the same metrics as your institutional licensing comparison)

Criteria	Policy 5 – A Narrow Path	California SB 1047 (2024)
Regulatory Body	National AI regulator (not fully defined; could be independent or government-affiliated)	<i>California Frontier Model Division</i> under the State CIO
Scope	Applies to most capable AI systems; aims to prevent development of ASI	Applies to “covered models” using $\geq 10^{26}$ FLOPs or with high-risk capabilities

Licensing Tiers	3 tiers: Training → Compute → Application	No explicit tiers, but phased safety obligations: Pre-deployment evals → Red teaming → Post-deployment monitoring
Enforcement Power	Left to national governments; variable by country; assumed capacity to enforce	State enforcement authority: legal penalties, deployment restrictions, civil action
Successes <i>(projected)</i>	Creates structural framework for long-term ASI prevention	Operationalizes many best practices: compute-based thresholds, red teaming, transparency, whistleblower protections
Failures / Risks	Regulatory ambiguity, potential capture, global coordination challenges	State-level limits, potential loopholes (e.g. open-source), possible corporate resistance
Key Historical Lesson	Global strategies can be derailed by national incentives and vague enforcement (cf. INCB/FDA)	Strong local policies can serve as templates, but risk limited impact without national or international scale