
Red Teaming A Narrow Path: ControlAI Policy Sprint

Abeer Sharma
University of Hong
Kong

With
ControlAI & Apart Research

Abstract

This report critically examines three Phase 0 AI governance proposals from *A Narrow Path*, aimed at preventing artificial superintelligence (ASI) development for 20 years. It evaluates Policy 2 (Prohibit AIs capable of breaking out of their environment), Policy 5 (Licensing regime & general intelligence restrictions), and Policy 6 (International treaty on AI development redlines). Using threat modeling, historical analogies, and implementation feasibility analyses, the report identifies key vulnerabilities including ambiguous terms (e.g., "environment"), challenges regulating open-source or covert AI projects, and rigid licensing thresholds that risk stifling innovation. Historical regulatory precedents (Clean Air Act, FDA, nuclear licensing, WMD treaties) highlight crucial gaps such as the absence of explicit insurance or compensation frameworks and risks of regulatory capture. Recommendations include clarifying policy definitions, extending coverage to human-enabled breaches, employing dynamic licensing criteria, integrating incentive structures (insurance requirements, liability funds), and adopting polycentric international coordination. These enhancements aim to

fortify policy effectiveness, maintaining the strategic goal of halting uncontrolled ASI development.

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

Abstract	1
Introduction	3
Objectives & Context	3
Hypotheses & Threat Model	3
Significance of These Policies	4
Methods	4
Policy and Textual Analysis	4
Threat Modeling Exercises	4
Historical & Policy Precedent Comparison	4
Implementation Feasibility Analysis	5
Integration of Stakeholder Feedback	5
Observations	5
Policy 2 – Prohibit AIs Capable of Breaking Out of Their Environment	5
Summary	5
Identified Gaps and Ambiguities	5
Effectiveness Weaknesses	6
Policy 5 – Licensing Regime and Restrictions on AI General Intelligence	6
Summary	6
Identified Weaknesses and Challenges	6
Effectiveness and Precedent Reflections:	7
Policy 6 – International Treaty on AI Development Redlines	9
Summary	9
Identified Weaknesses and Challenges	9
Discussion and Conclusion	
Cross-Cutting Vulnerabilities	10
Policy 2: Fragility of Containment-Based Controls	10
Policy 5: Rigid Licensing and Innovation Friction	11
Policy 6: Overextension of Arms Control Paradigms	11
Recommendations Across Policies	11
Final Reflection	12
References	12

Introduction

Objectives & Context

This analysis rigorously stress-tests three critical Phase 0 proposals outlined by A Narrow Path, a policy framework seeking to mitigate existential risks by halting artificial superintelligence (ASI) development for approximately 20 years. The policies examined are:

Policy 2: Prohibiting AI systems from breaking out of containment.

- **Policy 5:** Implementing stringent licensing regimes with capability restrictions.
- **Policy 6:** Establishing an international treaty to enforce common AI development redlines.

These policies constitute foundational layers of a "defense in depth" strategy designed to prevent the emergence of rogue superintelligence. This report identifies vulnerabilities and proposes improvements, preserving the overarching intent of each policy while addressing critical weaknesses.

Hypotheses & Threat Model

The hypothesis driving this analysis is that each policy contains latent vulnerabilities exploitable by malicious actors or susceptible to inadvertent human errors. Specifically:

- **Policy 2 (Prohibit AIs capable of breaking out of their environment):** Ambiguity in defining the AI's "environment" could allow breaches via human-mediated or officially sanctioned pathways.
- **Policy 5 (Licensing regime & general intelligence restrictions):** Rigid licensing thresholds may inadvertently incentivize unregulated AI development, overlooking incremental innovations, especially within open-source communities.
- **Policy 6 (International treaty on AI development redlines):** Excessive dependence on universal international agreement could heighten geopolitical friction and increase risks of partial compliance or outright defiance by key nations.

The threat model thus considers malicious actors (states, unethical corporations, rogue researchers), accidental breaches (human oversight), and the AI systems themselves as potential threat actors.

Significance of These Policies

The selected policies address different critical aspects of ASI risk:

- **Policy 2:** Directly targets technical escape and replication threats.
- **Policy 5:** Establishes institutional mechanisms to enforce capability restrictions.
- **Policy 6:** Aims to prevent international regulatory gaps that could enable competitive "races to the bottom."

Each policy's failure risks undermining the broader strategic safety objectives, making rigorous scrutiny essential.

Methods

Policy and Textual Analysis

Detailed textual analysis was conducted on Policies 2, 5, and 6 to identify ambiguities, particularly around critical terms such as "environment" (Policy 2) and "general intelligence" (Policy 5), alongside enforcement specifics.

Threat Modeling Exercises

Structured threat modeling identified potential exploit scenarios. This included technical breaches (AI-driven hacking), social engineering (AI manipulating human gatekeepers), and scenarios spanning intentional non-compliance to accidental breaches.

Historical & Policy Precedent Comparison

Comparative analysis evaluated these AI policies against analogous historical regulatory frameworks, including:

- **Computer Fraud and Abuse Act** (cybersecurity regulations).
- **Clean Air Act** (environmental protection).
- **FDA drug approval processes** (pharmaceutical safety).
- **FAA certification** (aviation safety).
- **Nuclear licensing and the NPT treaty** (non-proliferation of high-risk technologies).

This comparison highlighted both beneficial regulatory practices and pitfalls such as regulatory capture and enforcement evasion.

Implementation Feasibility Analysis

Practical enforcement challenges were analyzed, focusing on technical tools, institutional capacity, and international coordination mechanisms. Special attention was given to the feasibility of monitoring global compute usage and handling open-source AI developments.

Integration of Stakeholder Feedback

The analysis incorporated specific concerns raised externally, including ambiguity in policy definitions, risks of innovation suppression, and limitations of current enforcement proposals. This feedback ensured comprehensive critique coverage.

By combining textual policy analysis, threat modeling, historical comparisons, and stakeholder feedback, this methodology provides a robust examination of potential policy weaknesses and recommended improvements.

Observations

Policy 2 – Prohibit AIs Capable of Breaking Out of Their Environment

Summary

Policy 2 prohibits the development or operation of AI systems capable of achieving "unauthorized access" beyond their authorized computing environments, extending existing anti-hacking laws such as the Computer Fraud and Abuse Act to AI agents. The policy criminalizes intentional creation and negligent failures that enable unauthorized AI access, emphasizing preventive measures through rigorous testing and monitoring.

Identified Gaps and Ambiguities

Ambiguity of “Environment”: The term lacks clarity, especially regarding human-enabled breaches. While the policy states that unauthorized access includes misuse via intent or social engineering, it remains ambiguous whether more subtle or sanctioned forms of human cooperation—such as AI persuasion through permitted interfaces—are adequately covered.

Thought experiments like Yudkowsky’s AI-Box scenario suggest enforcement challenges where formal containment exists but human compliance is compromised.

Open-Source and Independent Agents: Policy effectiveness is limited against open-source models. Globally accessible models, exemplified by the LLaMA leak in 2023, can circumvent national regulations, complicating accountability and enforcement.

Accidental vs. Intentional Breakouts: Differentiating inadvertent and deliberate breaches poses enforcement difficulties. Drawing analogies from regulatory precedents like the Clean Air Act highlights the challenge of effectively monitoring latent AI capabilities compared to tangible pollutants.

Lack of Liability or Compensation Mechanisms: The absence of liability insurance or compensation schemes means the framework misses an opportunity to complement its punitive enforcement tools (e.g. fines or shutdowns) with market-aligned incentives for safety. Incorporating such mechanisms would align with precedents in sectors like transportation and finance, where insurability acts as both a risk signal and a compensatory fallback.

Effectiveness Weaknesses

Enforcement Practicality: Detection and attribution of unauthorized AI activities present significant practical challenges, especially with sophisticated or covert breaches, necessitating advanced surveillance and detection technologies.

Comparison with Existing Regulatory Analogies: Unlike environmental (Clean Air Act) or pharmaceutical (FDA) frameworks, the intangible and dynamic nature of AI complicates effective monitoring, enforcement, and prevention.

Social Engineering Risks: The current framework inadequately addresses scenarios where AI might leverage permitted channels or human cooperation, underscoring a critical weakness in relying solely on technical containment measures.

Policy 5 – Licensing Regime and Restrictions on AI General Intelligence

Summary

Policy 5 proposes a stringent licensing system regulated nationally, including Training Licenses (TL), Compute Licenses (CL), and Application Licenses (AL), aimed at limiting AI model capabilities and managing risks through controlled compute thresholds and regulated usage.

Identified Weaknesses and Challenges

Rigid Compute Thresholds: Reliance on fixed thresholds, even if nominally adjustable, risks rapid obsolescence due to bureaucratic inertia or political capture., potentially creating loopholes or unnecessarily restricting legitimate innovation.

Application Licensing Burdens: The AL framework introduces approval requirements for deploying AI applications that significantly enhance model capabilities. While the policy includes an automatic approval pathway for applications that do not trigger capability increases, and references benchmarks and a prohibited capabilities list, the enforcement process and updating mechanisms remain under-specified. Without clearer definitions or procedural safeguards, the AL system risks inconsistent application across jurisdictions and may deter benign, iterative innovations..

Compute License Practicality: Implementation complexities, such as real-time monitoring of compute usage and hardware tracking (including geopolitical concerns around GPS), present significant practical challenges, raising doubts about effective enforcement.

Regulatory Capture Risks: The powerful role envisioned for national AI regulators risks industry influence and regulatory capture, similar to historical examples like the FAA’s oversight failures.

Open-Source and Global Equity Concerns: Centralized licensing poses equity issues, potentially stifling global open-source innovation and creating resistance from nations prioritizing technological sovereignty and open development.

Effectiveness and Precedent Reflections:

We can examine some analogous frameworks to gauge how effective Policy 5 might be, drawing a comparative table below to illustrate:

Domain & Regulator	What is Licensed / Regulated	Purpose (Risk Mitigated)	Observations for AI Analogy
Pharmaceuticals (FDA/EMA)	Approval of new drugs (clinical trial phases, manufacturing quality)	Ensure safety & efficacy of medicines before public use. Prevent catastrophic side-effects.	Slow (10+ years R&D), costly process. High compliance; innovation is constrained but public trusts drug safety. AI analogy: similar pre-deployment review could catch alignment issues, but AI development

			cycles are faster than drug trials.
Civil Aviation (FAA)	Certification of aircraft designs, pilot licensing, maintenance procedures	Ensure aircraft are airworthy and pilots competent – minimize crash risk.	Very successful: air travel is extremely safe. However, incidents like 737 MAX show capture risks. For AI: licensing core models akin to aircraft type cert could improve safety, but AI “pilots” (operators) also matter.
Nuclear Energy (NRC)	Plant construction & operation licenses, nuclear material handling	Prevent reactor accidents & proliferation of nuclear materials for weapons.	High safety record in West, but very stringent – nuclear power growth stunted partly by regulatory burden. For AI: extreme safety focus could similarly slow progress (intended by Narrow Path). Also requires international treaties (NPT) to work globally.
Automobile Safety (Regulations & Insurance)	Vehicle safety standards (crash tests), driver licensing, mandatory insurance	Reduce accidents and compensate victims. Insurance spreads risk and incentivizes safer behavior.	Auto industry innovated (seatbelts, airbags) under regulation but not stifled. Insurance model crucial for compensation. AI: analogous insurance for AI dev/deployers could complement licensing, ensuring accountability for failures (currently missing in Policy 5).

Financial System (Banking Licenses & Capital Requirements)	Licenses for banks, stress tests, capital reserves, “know-your-customer” rules	Prevent bank failures from hurting economy, avoid illicit finance flows.	Generally effective in stabilizing finance, though crises still occur. KYC in finance parallels KYC for compute – enforcement has been mixed (money laundering still occurs via weak links). For AI: expect similar – determined actors may route around strict licensed compute to “shadow AI labs” in less regulated sectors or countries.
---	--	--	--

(Sources: FDA regulations; FAA/ICAO safety reports; NRC guidelines; Automotive insurance laws; Basel III Accords for banking.)

Table 1 – Licensing Regimes vs Historical Analogues

Policy 6 – International Treaty on AI Development Redlines

Summary

Policy 6 proposes an international treaty requiring signatory states to adopt domestic AI regulations mirroring Phase 0 measures, including compute thresholds, licensing systems, and prohibitions on training or deploying high-risk AI capabilities (e.g. self-improving or self-replicating systems). Signatories must also prohibit the use of AI models originating from non-member states, creating an exclusionary enforcement mechanism. The treaty is framed as a global harmonization effort, drawing analogies to arms control frameworks such as the Nuclear Non-Proliferation Treaty (NPT) and the Chemical Weapons Convention (CWC), and anticipates new institutional structures such as the International AI Safety Commission (IASC) and GUARD (Global Unit for AI Research and Development).

Identified Weaknesses and Challenges

Geopolitical Unrealism: The proposal assumes a level of international consensus that is unlikely given existing rivalries among major powers. Strategic competition in AI makes mutual limitation politically unattractive to key players like the US, China, and India. Expecting universal adoption may delay implementation or reduce participation to a small bloc of like-minded states.

Verification Constraints: Unlike nuclear weapons, AI development is intangible and distributed. Tools for detecting large training runs (e.g. telemetry, watermarking) are still experimental, and covert development using distributed compute or model obfuscation could evade detection. Verification capacity currently falls short of treaty ambitions.

Ambiguity in Model Origin and Enforcement: Determining whether an AI model originated from a non-signatory state is technically difficult, especially in open-source environments where weights, architectures, or derivative models can circulate freely. Enforcement of use-bans may be infeasible in practice.

Risk of Authoritarian Co-option: Broad enforcement powers under an international AI treaty could be misused by authoritarian governments to justify repression, surveillance, or information control under the pretext of treaty compliance. Without strong institutional safeguards, the regime may entrench illiberal norms.

Institutional Power Asymmetries: Proposed structures like the IASC may replicate dynamics seen in bodies like the UN Security Council, raising concerns about permanent dominance by a few powerful states. This could alienate middle powers and Global South countries who lack veto power or meaningful participation.

Exclusionary Trade Measures: The treaty's requirement to ban AI systems from non-signatory states may provoke digital trade disputes or tech fragmentation. If major producers remain outside the agreement, enforcing exclusions could undermine global scientific cooperation or provoke retaliatory measures.

Discussion and Conclusion

Cross-Cutting Vulnerabilities

While the three Phase 0 proposals aim to erect critical safety barriers against uncontrolled ASI development, they share structural shortcomings that limit their enforceability, legitimacy, and resilience. Chief among these are ambiguities in scope, reliance on rigid technical triggers, insufficient attention to post-incident responsibility, and overconfidence in top-down regulatory coordination. None of the policies adequately account for adversarial behavior, human fallibility, or the decentralization of AI development—especially in open-source environments.

Policy 2: Fragility of Containment-Based Controls

Policy 2's focus on preventing AI "breakouts" hinges on an underdefined notion of "environment." The policy does not clearly cover cases where an AI manipulates its human

overseers or leverages permitted interfaces to cause downstream harm. Its emphasis on unauthorized access fails to anticipate more subtle failure modes involving persuasion, social engineering, or accidental misuse. Moreover, enforcement depends on detecting violations whose signatures may be neither obvious nor attributable—especially with models running across distributed systems or user interfaces. Without clearer definitions and independent red-teaming protocols, containment measures risk being both overbroad and underinclusive.

Policy 5: Rigid Licensing and Innovation Friction

Policy 5 offers a formal structure for regulating model development, but its reliance on compute thresholds—even if nominally adjustable—risks rigidity in practice. While the policy allows for periodic recalibration by national or international regulators, it does not detail how such updates will be made responsive to rapid algorithmic improvements or insulated from bureaucratic lag and political capture. These features are ill-suited to a domain characterized by rapid efficiency gains and incremental capability escalation. Compute monitoring proposals—such as GPS tracking or hardware serialization—raise serious feasibility and sovereignty concerns, particularly for nations wary of surveillance risks. Although the framework includes internal tools like safety case submissions and near-miss reporting, it lacks complementary market-aligned mechanisms—such as liability insurance or compensation funds—that provide proactive incentives for responsible conduct. And while some regulatory analogies (e.g. pharmaceuticals or aviation) offer partial lessons, AI’s intangibility and adaptability present unique enforcement gaps those regimes were not designed to address.

Policy 6: Overextension of Arms Control Paradigms

Policy 6 correctly identifies the international dimension of ASI risk but leans too heavily on comparisons to nuclear non-proliferation treaties. Unlike fissile material, AI models can be duplicated, fine-tuned, or deployed covertly without centralized oversight. Verification tools remain underdeveloped and speculative, with no mature mechanisms proposed for reliably monitoring global compute flows or tracing model provenance. While the proposal includes preliminary ideas like the Multi-Threshold System and global compute tracking to aid verification, these mechanisms remain technically speculative and untested in open, distributed environments. The treaty’s exclusion clause risks economic fragmentation and retaliation, particularly if major AI-producing states opt out. Meanwhile, the proposed institutions (e.g. IASC) mirror existing multilateral bodies known for power asymmetries and gridlock, undermining their prospective legitimacy. There is also insufficient recognition of how treaty enforcement tools might be co-opted by authoritarian regimes under the guise of AI safety.

Recommendations Across Policies

To address these challenges, the following cross-cutting reforms are proposed:

- **Clarify Ambiguous Terms:** Policies should explicitly address human-in-the-loop failures and socially engineered breakouts, not just technical unauthorized access.

- **Adopt Dynamic Risk Metrics:** Replace fixed compute thresholds with adaptable criteria reflecting current model scaling trends, algorithmic efficiency, and contextual deployment risk.
- **Limit Bureaucratic Overreach:** Streamline licensing layers—especially for application development—to avoid stifling low-risk innovation while maintaining scrutiny over major capability escalations.
- **Incentivize Responsibility:** Introduce mandatory insurance schemes and compensation funds to provide downstream accountability for harms, following models from transportation and finance.
- **Abandon Overreliance on Monolithic Treaties:** Shift from universal treaty ambitions to modular, interoperable agreements among aligned states or actors, coupled with voluntary norms and market-based incentives.
- **Design for Polycentric Governance:** Encourage overlapping regulatory efforts from regional blocs, technical consortia, and civil society actors to reduce reliance on centralized enforcement bodies vulnerable to capture.
- **Safeguard Civil Liberties:** Ensure any surveillance or enforcement infrastructure is paired with due process guarantees, transparency mechanisms, and strong human rights protections.

Final Reflection

A Narrow Path presents an ambitious architecture for ASI risk mitigation. But ambition alone does not guarantee effectiveness. Preventing catastrophic AI outcomes requires governance frameworks that are not only robust and enforceable, but also adaptable to technical uncertainty, geopolitical fragmentation, and the inherent unpredictability of powerful general-purpose technologies. These policies will need to evolve toward more incentive-compatible, decentralized, and rights-respecting designs if they are to meet that challenge.

References

A Narrow Path – Phase 0: Safety. ControlAI (2025). (Sections on Policies 2, 5, 6) [narrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.co](https://narrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.conarrowpath.co)
narrowpath.conarrowpath.conarrowpath.co. Accessed via *narrowpath.co*.

Yudkowsky, E. (2002). *AI-Box Experiment*. (Described in: Wikipedia – AI capability control) – Demonstration that a human “gatekeeper” can be persuaded by a hypothetical AI to release it [en.wikipedia.org.en.wikipedia.org](https://en.wikipedia.org/wiki/Wikipedia:AI_capability_control).

United States Congress (1986). *Computer Fraud and Abuse Act* (CFAA). (Referenced in *A*

Narrow Path as model for unauthorized access definition)narrowpath.co.

Clean Air Act (42 U.S.C. §7401 et seq.) & EPA regulations. (Referenced precedent for banning harmful outputs regardless of intent)narrowpath.co.

U.S. Food & Drug Administration (FDA) – *Drug Approval Process & Good Manufacturing Practice (GMP) Regulations*. (Referenced as precedent for strict development protocols)narrowpath.co.

Volkswagen Emissions Scandal (2015) – Investigations by EPA and others. *Washington Post* tech report on defeat device[washingtonpost.com](https://www.washingtonpost.com). Demonstrates industry evasion of testing and importance of independent checks.

Lior, Noam. “Insuring AI: The Role of Insurance in Artificial Intelligence Regulation.” *Harvard J. of Law & Tech.* 35.2 (2022): 453-538. (Discusses compulsory insurance and compensation funds for AI/robot-caused harms)jolt.law.harvard.edujolt.law.harvard.edu.

European Parliament. *Report with Recommendations to the Commission on Civil Law Rules on Robotics*. (2017). (Proposed a compensation fund and insurance for robots)jolt.law.harvard.edujolt.law.harvard.edu.

Blumenthal, R. & Hawley, J. *Letter to Meta on LLaMA model leak*. (June 2023) blumenthal.senate.gov. U.S. Senate – notes the BitTorrent leak of LLaMA model and issues of open availability.

New Yorker (2020). “How Boeing and the FAA Created the 737 MAX Catastrophe.” (House report summary)[newyorker.com](https://www.newyorker.com). Example of regulatory capture undermining safety oversight.

Arms Control Association. “The Biological Weapons Convention at a Glance.” (2024)armscontrol.org. Notes known treaty violations (Soviet BW program) and verification limits.

United Nations. *Treaty on the Non-Proliferation of Nuclear Weapons* (NPT, 1968, entered 1970). (Historical arms control treaty limiting spread of a dangerous technology – partial success, lessons for AI treaty).

Organisation for the Prohibition of Chemical Weapons (OPCW). *Chemical Weapons Convention* (1993, entered 1997). (Nearly universal ban on an entire class of weapons; verification regime could inspire AI treaty verification mechanisms).

LessWrong forum discussions (various). e.g., “I played the AI box game as Gatekeeper — and lost.” (first-hand accounts reinforcing how humans might let an AI out given social pressure).