
Red Teaming "A Narrow Path": An Analysis of Phase 0 Policies for Artificial Superintelligence Prevention

Sky Bell
Georgia Institute of Technology

With
ControlAI & Apart Research

Abstract

This report red teams ControlAI's "A Narrow Path" Phase 0 policies, designed to prevent Artificial Superintelligence (ASI) development for 20 years. We analyzed core policies including bans on ASI self-improvement, breakout capabilities, and deliberate ASI creation, alongside a proposed licensing system. Critical implementation challenges identified include the ambiguity in defining "superintelligence" and "forbidden capabilities", limitations of compute thresholds as reliable regulatory triggers, and the inherent difficulty in verifying the absence of dangerous capabilities in "black box" AI systems. Policy effectiveness is threatened by geopolitical competition fostering a "race to the bottom", the paradoxical risk of safety research inadvertently advancing dangerous capabilities, and the practical hurdles of enforcing international treaties against private actors. While acknowledging alternative perspectives that advocate for unrestricted AI development, our analysis focuses on strengthening ControlAI's framework for real-world legislative adoption and a feasible ASI moratorium.

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

● Introduction

This project undertakes a critical red-teaming analysis of ControlAI's "A Narrow Path" Phase 0 policies, which are actively being pushed to lawmakers in the UK and US through their Direct Institutional Plan (DIP). Our central research question is: How feasible and effective are ControlAI's Phase 0 policies in preventing Artificial Superintelligence (ASI) development for 20 years, and what are the most critical vulnerabilities that need to be addressed before implementation?

We specifically chose to analyze Phase 0 policies because they are the immediate, foundational step in ControlAI's comprehensive plan, deemed necessary to halt ASI development and provide a crucial 20-year window for humanity to develop control mechanisms. Their imminent real-world implementation makes our feedback directly actionable. This red-teaming exercise acknowledges the ongoing debate surrounding AI governance, including perspectives that advocate for unrestricted technological progress, while focusing on strengthening ControlAI's proposed policies.

Our threat model considers a range of actors:

State Actors: Nations driven by geopolitical competition and national security interests, potentially prioritizing AI dominance over safety.

Corporate Actors: Private companies motivated by competitive pressures and economic incentives, potentially leading to a "race to the bottom" in safety standards.

Rogue Researchers/Non-State Actors: Individuals or small groups operating outside established regulatory frameworks, potentially developing or misusing dangerous AI capabilities.

We are investigating several failure modes:

Circumvention: How determined actors could bypass or exploit weaknesses in the proposed policies.

Unintended Consequences: Negative second-order effects arising from policy implementation, such as an accelerated AI arms race or regulatory arbitrage.

Implementation Feasibility Gaps: Practical challenges in deploying and enforcing policies due to bureaucratic, technical, or resource limitations.

Our analysis directly contributes to strengthening ControlAI's policies by identifying specific weaknesses, implementation challenges, and gaps before they reach legislators' desks. By providing evidence-based critiques and actionable recommendations, we aim to enhance the robustness, adaptability, and global feasibility of "A Narrow Path," thereby increasing the likelihood of successfully preventing ASI extinction risks.

● Methods

Our systematic approach to policy critique during this sprint involved a multi-faceted analytical methodology:

1. **Historical Case Study Analysis:** We extensively examined historical precedents in high-stakes technology governance, primarily focusing on nuclear non-proliferation. This included:
 - **The Baruch Plan (1946):** Analyzed its failure to identify lessons regarding the necessity of perceived earnestness, equity, and avoiding maximalist solutions in international agreements.
 - **The Non-Proliferation Treaty (NPT) (1968):** Studied its success to understand the conditions necessary for effective international governance, such as shared risk understanding, strategic stability, and dominant power collusion.
 - **"Atoms for Peace" Initiative:** Explored its "grand bargain" concept and its unintended consequences, particularly regarding the acceleration of nuclear proliferation and the entrenchment of Cold War divisions, drawing parallels to AI's dual-use nature and the challenges of leveraging technological leads.
 - **Export Controls and Dual-Use Technology Regulation:** Investigated current US export controls on advanced computing Integrated Circuits (ICs) and AI model weights (ECCN 4E091) to understand mechanisms for controlling dual-use technologies and their enforcement challenges.
2. **Comparative Policy Framework Analysis:** We compared ControlAI's Phase 0 policies with existing and proposed AI regulatory initiatives, such as the EU AI Act and various US federal and state-level efforts. This allowed us to assess the unique scope and ambition of "A Narrow Path" against current regulatory trends and identify areas of overlap, divergence, and potential conflict.
3. **Threat Modeling Methodology:** We systematically considered various threat actors (state, corporate, rogue researchers) and their potential motivations (geopolitical dominance, economic gain, ideological aims) to identify how they might circumvent or exploit policy weaknesses. This involved anticipating "gaming" behaviors, black market development, and the strategic implications of unilateral actions.
4. **Implementation Feasibility Assessment Techniques:** Our analysis delved into the practical challenges of policy implementation, considering:
 - **Bureaucratic Complexities:** Assessed the capacity of government agencies, inter-agency coordination, and the impact of legacy IT systems and data fragmentation in the public sector.
 - **Technical Challenges:** Evaluated the feasibility of defining and measuring "superintelligence" and "forbidden capabilities", the limitations of compute thresholds, the "black box" problem in AI verification, and the readiness of hardware-enabled mechanisms (HEMs) for enforcement.

- **Legal Hurdles:** Examined potential constitutional challenges to development bans , complexities of licensing dual-use technologies and private sector actors , and compliance with data privacy and intellectual property laws.
5. **Evidence Gathering and Verification Methods:** Our analysis is rigorously supported by evidence drawn directly from the provided research notes. Claims are backed by:
- **Empirical Data:** References to polling data on public support for AI regulation , and projections on the rapid increase of models exceeding compute thresholds.
 - **Documented Case Studies:** Examples of past enforcement challenges, such as China acquiring restricted NVIDIA chips despite US sanctions.
 - **Credible Expert Analysis:** Citations from academic papers, policy reports, and statements by leading AI scientists and company executives.
 - **Direct Policy Documents:** References to ControlAI's "A Narrow Path" framework and its Direct Institutional Plan.

This systematic approach allowed for a comprehensive and evidence-based red-teaming of ControlAI's Phase 0 policies, identifying specific vulnerabilities and informing actionable recommendations.

● Results

Our red-teaming analysis of ControlAI's Phase 0 policies revealed critical findings across implementation feasibility, policy effectiveness, and evidence-based assessment:

Implementation Feasibility Gaps:

1. Ambiguity in Definitions and Measurement:

- **Challenge:** The core policies ban "superintelligence" and "dangerous AI capabilities" (e.g., automated AI research, hacking). However, there is a lack of scientific consensus on precise, measurable definitions for "superintelligence" or "general intelligence". Current evaluation tools are limited in scope and reliability.
- **Impact:** This vagueness creates significant loopholes, making objective legal enforcement exceedingly difficult. Regulators would struggle to definitively determine when an AI system crosses a prohibited threshold, leading to inconsistent application or developers claiming compliance without true adherence.

2. Limitations of Compute Thresholds:

- **Challenge:** Phase 0's licensing system likely relies on compute thresholds (e.g., 10^{25} FLOPs for systemic risk). However, model performance is influenced by factors beyond raw compute, such as data quality and architectural innovations.

- **Impact:** Over-reliance on static thresholds risks rapid policy obsolescence, as developers are incentivized to achieve dangerous capabilities with less compute, effectively "gaming" the system. This could lead to disproportionate scrutiny on computationally heavy but less risky models, while smaller, harmful models escape oversight. Projections show a rapid increase in models exceeding current thresholds (over 200 models $>10^{26}$ FLOPs by 2030), implying an unmanageable regulatory burden.
- 3. **Technical Verification Hurdles ("Black Box" Problem):**
 - **Challenge:** The requirement for companies to demonstrate "no forbidden capabilities" before deployment is technically challenging due to the "black box" nature of many advanced AI models. Proving a negative (that a system *cannot* exhibit certain behaviors) is inherently difficult.
 - **Impact:** This creates an unmanageable burden of proof for developers and regulators, potentially leading to lax enforcement or stifled innovation. Risks like data poisoning and adversarial attacks further complicate reliable safety verification.
- 4. **Underdeveloped Hardware-Enabled Mechanisms (HEMs):**
 - **Challenge:** Effective enforcement relies on technical mechanisms like detecting high-compute training runs and physical "kill-switches". While HEMs are proposed for verifiable reporting and enforcement, the hardware is currently "just a prototype maintained by one small team".
 - **Impact:** Relying on unproven or potentially vulnerable hardware for a global, 20-year moratorium introduces a critical single point of failure. Tamper-proofing, meter security, and resistance to circumvention are complex technical challenges that are not yet fully addressed.
- 5. **Public Sector Capacity and Coordination:**
 - **Challenge:** Public sector agencies frequently struggle with complex, fragmented administrative processes, high operational costs, inconsistent service quality, prolonged response times, and limited resource allocation. They face pervasive legacy IT systems, fragmented and poor-quality data, and a lack of specialized AI talent. Managing AI-driven change in large bureaucracies is difficult due to institutional resistance and risk aversion.
 - **Impact:** These foundational issues hinder effective AI adoption and regulation, creating bureaucratic confusion and slowing progress from pilot projects to full-scale implementation.

Policy Effectiveness Weaknesses:

1. **Circumvention by Malicious Actors:**
 - **Threat Model:** Rogue researchers or state actors seeking to gain an advantage.
 - **Weakness:** The "no AIs improving AIs" and "no AIs capable of breaking out of their environment" prohibitions are critical, but if circumvented, they could rapidly accelerate uncontrolled ASI

- development. The distributed nature of AI development and the potential for a compute black market make enforcement challenging.
- **Failure Mode:** Covert development in less regulated jurisdictions or through illicit channels could undermine the global moratorium.
2. **"Capability Externalities" and Safety Research Paradox:**
 - **Threat Model:** All actors engaged in AI development, including those focused on safety.
 - **Weakness:** Advancements in AI safety research (e.g., RLHF, interpretability) can inadvertently lead to gains in model performance and capabilities.
 - **Failure Mode:** This creates a strategic dilemma, disincentivizing open collaboration on safety among geopolitical rivals, as sharing research could inadvertently advance a competitor's strategic AI capabilities. This paradoxically fuels a covert "safety race" that leads to faster, more powerful AI development, potentially shortening the 20-year prevention window.
 3. **Geopolitical "Race to the Bottom":**
 - **Threat Model:** State and corporate actors.
 - **Weakness:** Intense global competition for AI dominance incentivizes rapid development over safety. Unilateral actions, like US export controls on compute, risk creating a "bifurcated AI ecosystem" and an "AI arms race".
 - **Failure Mode:** Without robust international consensus and verifiable commitments, any single actor's decision to prioritize speed over safety could undermine the collective effort to halt ASI development, leading to a proliferation of less-safe AI systems.
 4. **Legal Challenges to Development Bans:**
 - **Threat Model:** Corporate actors and individual researchers in democratic nations.
 - **Weakness:** A direct ban on AGI development could face significant legal challenges under First Amendment protections for free expression in the US, as "people, not technologies, have rights".
 - **Failure Mode:** Such bans might be legally challenged, overturned, or lead to development shifting to jurisdictions with less restrictive legal frameworks, undermining the policy's global reach.
 5. **Stifling Beneficial Innovation:**
 - **Threat Model:** Society at large.
 - **Weakness:** Overly broad bans or stringent licensing requirements, while intended to prevent catastrophe, could inadvertently stifle beneficial AI development in areas like healthcare and scientific discovery.
 - **Failure Mode:** This would represent a significant opportunity cost for humanity, potentially hindering solutions to other global challenges and undermining the long-term "Flourishing" phase of "A Narrow Path"

Evidence-Based Assessment:

1. Nuclear Non-Proliferation Analogies (Applicability and Limitations):

- **Evidence:** The Baruch Plan's failure highlights the need for earnestness, equity, and avoiding maximalist solutions in international agreements. The NPT's success stemmed from a shared understanding of risks, strategic stability (MAD), and dominant power collusion.
- **Assessment:** While valuable for conceptualizing governance of high-stakes dual-use technologies, direct application to AI is limited. AI lacks a clear "weapons-grade" equivalent, its development is distributed among private companies, and there's no equivalent "AI stability" or "stalemate" to incentivize a halt. The conditions for AI governance currently resemble the chaotic period after nuclear weapons were invented, not the later, more institutionalized phase.

2. Export Controls (Effectiveness and Pitfalls):

- **Evidence:** US export controls on advanced computing ICs and AI model weights (ECCN 4E091) demonstrate governmental recognition of compute as a control point. However, past instances show enforcement difficulties (e.g., China acquiring restricted NVIDIA chips through indirect channels).
- **Assessment:** While compute is "detectable, excludable, and quantifiable", its governance is fraught with challenges. Lack of global consensus, potential for black markets, and the risk of stifling beneficial innovation are significant concerns. Unilateral controls can lead to a "bifurcated AI ecosystem" and an "AI arms race," undermining global cooperation.

3. "If-Then Commitments" (Potential Solution):

- **Evidence:** "If-then commitments," where development pauses when "capability triggers" are met, are presented as a more adaptable approach. They avoid *ex ante* agreement on all risks and support "peaceful" AI development.
- **Assessment:** This capability-based approach could enhance credibility and mitigate racing dynamics if adopted universally. However, current commitments are voluntary, and their effectiveness hinges on eventual enforcement by an international regime, akin to the IAEA and NPT.

● Discussion and Conclusion

Synthesizing our analysis, it is highly improbable that the current Phase 0 policies, as written, would *actually* prevent ASI development for 20+ years without significant strengthening and adaptation. While ControlAI's framework is commendable for its proactive and comprehensive approach to existential risk, its ambition currently outstrips the technical, legal, and geopolitical realities of AI governance.

The most critical vulnerabilities are:

1. **Definitional and Verification Ambiguity:** The inability to precisely define "superintelligence" and "forbidden capabilities" and the "black box" nature of AI make robust verification and enforcement technically infeasible. This creates loopholes that determined actors could exploit.
2. **Static Thresholds and Rapid Obsolescence:** Over-reliance on static compute thresholds is a policy design flaw in a field characterized by rapid, unpredictable advancements and efficiency gains. This incentivizes "gaming" the system and renders regulations quickly outdated.
3. **Geopolitical Competition and Lack of Credible Commitment:** The intense global "race to the bottom" in AI development fundamentally undermines the premise of a coordinated halt. Unilateral actions (e.g., export controls) risk accelerating an AI arms race and creating a bifurcated ecosystem, rather than fostering the necessary global cooperation.
4. **Enforcement Against Non-State Actors:** The distributed, private-sector-led nature of AI development presents novel challenges for international treaty enforcement, as existing models are primarily designed for state actors.

These policies could be strengthened before implementation by:

- **Adopting Dynamic, Capability-Based Regulation:** Moving beyond static compute thresholds to incorporate continuous, multi-faceted capability evaluations and impact assessments. This would focus on an AI model's demonstrated abilities and potential for harm, ensuring regulatory relevance as AI evolves.
- **Investing in Verifiable Safety Mechanisms:** Prioritizing and funding AI safety research focused on "containment" and interpretability, including developing robust, tamper-proof hardware-enabled mechanisms (HEMs) for verifiable reporting and enforcement, moving them beyond prototype stages.
- **Fostering Strategic Diplomacy and Incentivized Cooperation:** Developing diplomatic strategies to build global consensus on AI governance, potentially leveraging "if-then commitments" as a framework for conditional development. This requires navigating geopolitical tensions and offering incentives for participation in international regulatory regimes, learning from the NPT's success in achieving "collusion among dominant powers".
- **Strengthening Public Sector Capacity:** Investing significantly in public sector expertise in AI governance, including technical and legal capabilities, and addressing foundational digital infrastructure challenges.

The broader implications for AI governance are profound. Given current AI development timelines, with dangerous systems potentially emerging within years, the urgency of implementing effective policies is paramount. However, rushing to implement an unrefined framework risks unintended consequences, such as an accelerated AI arms race or stifling beneficial innovation. Lessons from historical regulatory precedents, particularly nuclear non-proliferation, highlight that effective governance requires a shared understanding of risks, strategic stability, and adaptable mechanisms for compliance. The current AI governance landscape is in an immature, chaotic phase, underscoring the need for a nuanced approach that balances immediate risk mitigation with the long-term goal of safe, transformative AI.

● References

- Miotti, A., Bilge, T., Kasten, D., & Newport, J. (2024). *A Narrow Path: How to Secure Our Future*. Control AI. <https://www.narrowpath.co/> Apart Research.
- (2025). *Red Teaming A Narrow Path: Control AI Policy Sprint*. <https://apartresearch.com/sprints/red-teaming-a-narrow-path-control-ai-policy-sprint-2025-06-13-to-2025-06-13>
- Apart Research. (2025). *Red Teaming A Narrow Path*. <https://apartresearch.com/sprints/collaborations>
- Biocomm.ai. (2024). *A Narrow Path*. <https://blog.biocomm.ai/2024/10/02/a-narrow-path/>
- ControlAI. (n.d.). *About Us*. <https://controlai.com/about>
- ControlAI. (n.d.). *Direct Institutional Plan*. <https://controlai.com/dip>
- ControlAI. (n.d.). *How to Help*. <https://controlai.com/how-to-help>
- ControlAI. (n.d.). *Phase 0*. <https://www.narrowpath.co/phase-0>
- The Compendium.ai. (n.d.). *AI Governance*. <https://www.thecompendium.ai/ai-governance>
- The Next Frontier. (2025). *Lessons for AI Governance from Atoms for Peace: Introduction*. <https://www.thenextfrontier.blog/lessons-for-ai-governance-from-atoms-for-peace-introduction/>
- The Next Frontier. (2025). *Lessons for AI Governance from Atoms for Peace, Part 2: Chips for Peace?*. <https://www.thenextfrontier.blog/lessons-for-ai-governance-from-atoms-for-peace-part-2-chips-for-peace/>
- LessWrong. (2025). *What We Learned from Briefing 70 Lawmakers on the*

Threat.

<https://www.lesswrong.com/posts/Xwrajm92fdjd7cqnN/what-we-learned-from-briefing-70-lawmakers-on-the-threat>

LessWrong. (2025). *The EU is Asking for Feedback on Frontier AI Regulation.*
<https://www.lesswrong.com/posts/AbnfsnEEmHFmprGzm/the-eu-is-asking-for-feedback-on-frontier-ai-regulation-open>

LessWrong. (n.d.). *What is Compute Governance and How Does it Affect Overall AI?*. (<https://www.lesswrong.com/posts/j34HnuELsgj4GjBqH/what-is-compute-governance-and-how-does-it-affect-overall-ai>)

Miotti, A. (2025). *Keynote Talk with Andrea Miotti at Red Teaming A Narrow Path: Control AI Policy Sprint.*

YouTube. (<https://www.youtube.com/watch?v=VzRcYx6ldQY>)

Norton Rose Fulbright. (2025). *Practical Commentary Regarding Copyright and Generative AI Training.*
<https://www.nortonrosefulbright.com/en-ca/knowledge/publications/87200379/practical-commentary-regarding-copyright-and-generative-ai-training>

Opinio Juris. (2025). *Enforcing an AI Moratorium: Lessons from the ICC.*
<https://opiniojuris.org/2025/02/13/enforcing-an-ai-moratorium-lessons-from-the-icc/>

Purplesec. (n.d.). *AI in Cybersecurity.*
<https://purplesec.us/learn/ai-in-cybersecurity/>

RAND. (2025). *Insights from Nuclear History for AI Governance.*
https://www.rand.org/content/dam/rand/pubs/perspectives/PEA3600/PEA3652-1/RAND_PEA3652-1.pdf

ResearchGate. (2025). *International Agreements on AI Safety: Review and Recommendations for a Conditional AI Safety Treaty.*
https://www.researchgate.net/publication/390176935_International_Agreements_on_AI_Safety_Review_and_Recommendations_for_a_Conditional_AI_Safety_Treaty

The Federal Register. (2025). *Framework for Artificial Intelligence Diffusion.*
<https://www.federalregister.gov/documents/2025/01/15/2025-00636/framework-for-artificial-intelligence-diffusion>

The FIRE. (n.d.). *Artificial Intelligence, Free Speech, and the First Amendment.*
<https://www.thefire.org/research-learn/artificial-intelligence-free-speech-and-first-amendment> The Law & Economics Center. (2025).

US Export Controls on AI and Semiconductors: Two Divergent Visions.
<https://laweconcenter.org/resources/us-export-controls-on-ai-and-semiconductors-two-divergent-visions/>

The Tech Policy Press. (2025). *Bigger Might Not Be Better: The Limits of Regulating AI Through Compute Thresholds.*
<https://www.techpolicy.press/bigger-might-not-be-better-the-limits-of-regulating-ai-through-compute-thresholds/>

The Tech Policy Press. (2025). *DeepSeek and the Race to AGI: How Global AI Competition Puts Ethical Accountability at Risk.*
<https://www.techpolicy.press/deepseek-and-the-race-to-agi-how-global-ai-competition-puts-ethical-accountability-at-risk/>

UK Government. (n.d.). *Frontier AI Capabilities and Risks: Discussion Paper.*
<https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/future-risks-of-frontier-ai-annex-a>

University of Michigan Law School. (n.d.). *The International Legal Obligation to Mitigate AI Existential Risk.*
<https://repository.law.umich.edu/cgi/viewcontent.cgi?article=2169&context=mjil>

Waterfront Law. (n.d.). *Navigating AI and Copyright: What Every UK Business Needs to Know.*
<https://waterfront.law/navigating-ai-and-copyright-what-every-uk-business-needs-to-know/>

WSGR Data Advisor. (2025). *EU AI Office Clarifies Key Obligations for AI Models Becoming Applicable in August.*
<https://www.wsgrdataadvisor.com/2025/04/eu-ai-office-clarifies-key-obligations-for-ai-models-becoming-applicable-in-august/>

Zscaler. (n.d.). *AI Cybersecurity Regulations: What CISOs Need to Know.*
<https://www.zscaler.com/exorevolutionaries/insights/ai-cybersecurity-regulations-what-cisos-need-know>

Akamai. (n.d.). *AI Cybersecurity: How it's Impacting the Fight Against Cybercrime.*
<https://www.akamai.com/blog/security/ai-cybersecurity-how-impacting-fight-against-cybercrime>

arXiv. (2025). *International Agreements on AI Safety: Review and Recommendations for a Conditional AI Safety Treaty.*
<https://arxiv.org/html/2503.18956v1>

arXiv. (2025). *Risks of cooperation on AI safety.*

<https://arxiv.org/html/2504.12914v1>

arXiv. (2025). *The Strategic Landscape of AI Development*.

<https://arxiv.org/html/2505.04592v1>

Cato Institute. (n.d.). *Opportunity Costs of State and Local AI Regulation*.

<https://www.cato.org/policy-analysis/opportunity-costs-state-local-ai-regulation>

Cohere. (n.d.). *The Limits of Thresholds*.

<https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf>

CSIS. (n.d.). *AI Diffusion Framework: Securing US AI Leadership While Preempting Strategic Drift*.

<https://www.csis.org/analysis/ai-diffusion-framework-securing-us-ai-leadership-while-preempting-strategic-drift>

Diligent. (2025). *AI's impact on cybersecurity risk management — and what it means*.

<https://www.diligent.com/resources/blog/ai-cybersecurity-risk-management>

Epoch AI. (n.d.). *How Many AI Models Will Exceed Compute Thresholds?*

<https://epoch.ai/blog/model-counts-compute-thresholds>

F5. (n.d.). *Tackling AI challenges in public services with solutions designed for the complexity*.

<https://www.f5.com/company/blog/tackling-ai-challenges-in-public-services-with-solutions-designed-for-the-complexity>

Forbes Councils. (2024). *AI Regulation, Global Governance And Challenges*.

<https://www.forbes.com/councils/forbestechcouncil/2024/11/12/ai-regulation-global-governance-and-challenges/>

GovTribe. (n.d.). *Public Sector AI Adoption Challenges*.

<https://govtribe.com/topic-insights/public-sector-ai-adoption-challenges>

IIF. (2025). *Sustainable Finance Monitor*.

https://www.iif.com/portals/0/Files/content/Regulatory/32370132_iif_sfm_february_2025_-_final_version_for_release-199946066.pdf

King & Spalding. (2025). *AI Quarterly Update: Recent AI Legislation Efforts Signal Potential Challenges for State-Led Regulatory Approach*.

<https://www.kslaw.com/news-and-insights/ai-quarterly-update-recent-ai-legislation-efforts-signal-potential-challenges-for-state-led-regulatory-approach>

Law-AI. (n.d.). *The Role of Compute Thresholds for AI Governance*.

<https://law-ai.org/the-role-of-compute-thresholds-for-ai-governance/>

London VC Network. (n.d.). *AI is Powering Drug Discovery: Accelerating*

Innovation and How VCs are Scaling Breakthroughs.

<https://www.londonvcnetwork.com/post/ai-is-powering-drug-discovery-accelerating-innovation-and-how-vc-s-are-scaling-breakthroughs>

MDPI. (2025). *AI in Public Administration: Challenges and Opportunities.*

<https://www.mdpi.com/2076-3387/15/4/149>

NACD. (2025). *How AI Will Impact Cybersecurity Regulatory and Disclosure.*

<https://www.nacdonline.org/all-governance/governance-resources/governance-research/director-handbooks/DH/2025/ai-in-cybersecurity/how-ai-will-impact-cybersecurity-regulatory-and-disclosure-matters/>

Oxford Intersections: AI in Society. (n.d.). *Artificial Intelligence and the Future of Nuclear Weapons.*

<https://academic.oup.com/edited-volume/59762/chapter/511416833>

Quinn Emanuel. (2025). *Artificial Intelligence: EU Regulation and Competition Law Enforcement Addressing Emerging Challenges.*

<https://www.quinnemanuel.com/the-firm/publications/artificial-intelligence-eu-regulation-and-competition-law-enforcement-addressing-emerging-challenges/>

Tech Policy Press. (2025). *The Path to Effective AI in the Public Sector Starts with Content Management.*

<https://www.govexec.com/sponsors/2025/06/path-effective-ai-public-sector-starts-content-management/405923/?oref=featured-insights>

The ACLU. (n.d.). *Trump's Efforts to Dismantle AI Protections Explained.*

<https://www.aclu.org/news/privacy-technology/trumps-efforts-to-dismantle-ai-protections-explained> Wikipedia. (n.d.). *Existential risk from artificial intelligence.*
https://en.wikipedia.org/wiki/Existential_risk_from_artificial_intelligence

PNAS. (n.d.). *Existential risk narratives do not overshadow immediate societal threats of AI.* <https://www.pnas.org/doi/10.1073/pnas.2419055122>

arXiv. (n.d.). *Reconstructing and Evaluating Arguments Against the Existential Risk Perspective on AI.* <https://arxiv.org/abs/2501.04064>

Britannica. (n.d.). *AI Rules and Regulations.*

<https://www.britannica.com/money/ai-rules-and-regulations>

WeAreDevelopers. (n.d.). *EU AI Regulation: Artificial Intelligence Regulations.*

<https://www.wearedevelopers.com/magazine/eu-ai-regulation-artificial-intelligence-regulations>

Dataforce.ai. (n.d.). *Why AI Regulation Doesn't Have to Mean Less Innovation.*

<https://www.dataforce.ai/blog/why-ai-regulation-doesnt-have-mean-less-innovation>

LitsLink. (n.d.). *AI Regulation*. <https://litslink.com/blog/ai-regulation>

PMC. (n.d.). *Internal tensions and ethical issues in RLHF*.

<https://pmc.ncbi.nlm.nih.gov/articles/PMC12137480/>

Effective Altruism Forum. (n.d.). *What mistakes has the AI safety movement made?*.

<https://forum.effectivealtruism.org/posts/tEmQrfMs9qdBPrGKh/what-mistakes-has-the-ai-safety-movement-made>

Wikipedia. (n.d.). *Effective accelerationism*.

https://en.wikipedia.org/wiki/Effective_accelerationism

El País. (2024). *Accelerate or die: The controversial ideology that proposes the unlimited advance of artificial intelligence*.

<https://english.elpais.com/technology/2024-01-06/accelerate-or-die-the-controversial-ideology-that-proposes-the-unlimited-advance-of-artificial-intelligence.html>

- **Appendix**

Appendix A. Additional Threat Model Scenarios

- **Accidental Breakthroughs:** Unforeseen emergent capabilities in AI systems that were not intentionally designed to be superintelligent, but achieve it through complex interactions or self-optimization.
- **Open-Source Proliferation:** The rapid and uncontrolled spread of powerful open-source AI models that fall below regulatory thresholds but can be fine-tuned or combined to achieve dangerous capabilities by a wide range of actors.
- **"Shadow AI" Development:** Covert or decentralized AI development efforts that operate entirely outside of official regulatory oversight, potentially leveraging distributed computing resources or smaller, less detectable models.
- **Dual-Use Dilemma Exploitation:** Actors exploiting the beneficial applications of AI to mask or justify the development of capabilities that could be repurposed for harmful ends.

Appendix B. Geopolitical and International Coordination Gaps

Lessons from Nuclear Non-Proliferation (Baruch Plan, NPT) and their Applicability/Limitations to AI Governance

Analysts frequently draw parallels between nuclear governance and AI governance due to the dual-use nature of both technologies and their potential for catastrophic risks. Examining historical precedents offers critical insights.

The **Baruch Plan (1946)**, an unsuccessful attempt at international nuclear control, failed primarily because it was perceived by the Soviet Union as favoring US interests, lacking earnestness and equity, antagonizing rivals, and seeking an overly maximalist solution. This highlights the importance of perceived fairness and mutual benefit in international agreements.

In contrast, the **Non-Proliferation Treaty (NPT) (1968)** succeeded due to a growing shared understanding of existential risks (e.g., the "Nth country problem" and proliferation cascades), the establishment of strategic stability (Mutually Assured Destruction - MAD), the willingness of dominant powers (the US and USSR) to collude on mutually agreeable terms, and a clear understanding of what exactly was to be governed (nuclear material and weapons).

However, there are key divergences that limit the direct applicability of these nuclear precedents to AI governance:

- **"Weapons-Grade" Equivalent:** Unlike nuclear technology, AI lacks a clear, universally agreed-upon "weapons-grade" equivalent, making it difficult to define what constitutes a dangerous use that needs restriction.
- **Development Actors:** AI development is highly distributed and predominantly led by private companies with competitive incentives, unlike nuclear technology which requires massive state-backed infrastructure.
- **Pace of Change:** The AI landscape is unpredictable, with leadership positions and capabilities potentially shifting in months rather than decades, a pace far exceeding that of nuclear technology.
- **Strategic Stability:** There is no equivalent "AI stability" or "AI stalemate" yet; policymakers still perceive significant benefits in increasing AI deployment for strategic and economic advantages, disincentivizing a halt.

While nuclear non-proliferation offers valuable conceptual analogies, the fundamental differences in AI's "weapons-grade" equivalent, its rapid, private-sector-led development, and the absence of a shared understanding of "AI stability" render direct historical precedents largely inapt for Phase 0's immediate global halt. The NPT's success relied on conditions (shared risk understanding, strategic stability, dominant power collusion) that are currently absent in AI governance, indicating that the global AI governance landscape is in a chaotic, immature phase, making a rapid, universally accepted international agreement highly challenging.

Appendix C. International Coordination Framework Assessments

Achieving a global moratorium on Artificial Superintelligence (ASI) development, as envisioned by ControlAI's "A Narrow Path," fundamentally hinges on robust and effective international coordination. The framework explicitly states that for ASI development to be prevented globally, countries must ultimately agree to enforce international rules, likely through an international treaty, requiring individual nations to lead domestically first before building consensus for a treaty.

This section provides a granular analysis of potential international governance models, including the feasibility of a new international body, the role of existing organizations (e.g., UN, IAEA), and strategies for overcoming geopolitical divides and sovereignty concerns.

Existing International Efforts and Their Limitations: Numerous international efforts have emerged to address AI governance, including the OECD AI Principles (2019), the Global Partnership on AI (2020), UNESCO's Recommendation on the Ethics of Artificial Intelligence (2021), the G7 Hiroshima AI Process (2023), the UN High-Level Advisory Body on AI (2023), and the Council of Europe's Framework Convention on Artificial Intelligence (2024). AI Safety Institutes (AISIs) have also been established in the UK, US, EU, Japan, Singapore, France, and Canada, with the UK AISI publishing an International AI Safety Report with contributions from 33 countries. The UN High-Level Advisory Body on AI has emphasized that "AI governance regimes must also span the globe to be effective".

However, many researchers argue these efforts are insufficient, noting that while advanced AI is high-risk, it lacks the oversight of other high-risk sectors. Many initiatives rely on voluntary commitments or broadly-defined principles, lacking concrete measures for coordinated safety, research, development, and benefit-sharing.

Challenges to International Agreements:

1. **Lack of Consensus on "Alignment" and Values:** A significant hurdle is the lack of global agreement on what an "aligned" AI model looks like. Research indicates that even within countries, there is no consensus on alignment, and individuals have strong preferences for models. This difficulty is amplified across nations with diverse beliefs, traditions, and cultures, making it challenging for countries not adhering to democratic and liberal values to participate in such agreements.
2. **Technological Dominance and Geopolitical Competition:** Technological dominance is a key component of national security, and countries are increasingly reliant on developing domestic AI capabilities to avoid "international obsolescence". This dynamic, reminiscent of the Cold War pursuit of technological superiority, drives a "race to the bottom" in AI development, leaving "every country on its own on AI". This intense competition makes it hard for leading nations to credibly commit to prioritizing safety.
3. **Commitment Problems and Inadvertent Acceleration:** A "coalition of the cautious" attempting to leverage a technological lead for AI risk reduction faces commitment problems. The lead must be substantial enough to allow time for safety progress, but rapid scaling and model distillation create uncertainty about how long this advantage can be maintained. Furthermore, safety research itself can inadvertently contribute to capability gains ("capability externalities"), creating a dilemma where cooperation could advance a rival's strategic AI capabilities or expose sensitive national security information.
4. **Enforcement Against Non-State Actors:** Unlike traditional arms control treaties that primarily target state actors, AI development is largely

conducted by private companies driven by competitive incentives. A multilateral compute cap, for instance, would seek to regulate corporate entities and individual researchers, not just public entities. This introduces novel challenges in enforcement, particularly concerning complex cross-border elements where data centers, developers, and corporations may be physically dispersed. Companies may also refuse early access to their technology due to intellectual property concerns.

5. **Sovereignty Concerns:** AI safety discourse, while effective domestically, faces "distinct international political pitfalls surrounding notions of sovereignty, regional bias, and legitimacy" when attempting to establish multilateral treaties.

Potential Governance Models and Mechanisms:

1. **New International Body (e.g., Network of AISIs):** Proposals suggest a treaty establishing a compute threshold above which development requires rigorous oversight, mandating complementary audits of models, information security, and governance practices. This could be overseen by an international network of AI Safety Institutes (AISIs) with authority to pause development if risks are unacceptable. Such a body would need to balance immediately implementable measures with a flexible structure to adapt to evolving AI.
2. **Role of Existing Organizations (UN, IAEA):**
 - **United Nations:** The UN has created a High-Level Advisory Body on AI to advance "globally coordinated AI governance".
 - **International Atomic Energy Agency (IAEA) as a Model:** The IAEA, established following Eisenhower's "Atoms for Peace" initiative, institutionalized nuclear non-proliferation. It demonstrates a model for enforcing international obligations, including on non-state legal persons, and its principle of complementarity could be embedded in AI regulation enforcement. However, the analogy has shortcomings: AI lacks a clear "weapons-grade" equivalent, reliably verifying AI safety is difficult, and intense global competition makes credible commitment to safety challenging. The NPT's success relied on dominant power collusion and a clear understanding of what to govern, conditions currently lacking for AI.
3. **"If-Then Commitments":** This approach involves allowing continued AI development *unless* pre-agreed "capability triggers" are flagged during evaluations. This model could address the difficulty in achieving *ex ante* agreement on all potential risks or defining a "weapons-grade" equivalent for AI upfront. It supports "peaceful" AI by avoiding unnecessary restrictions on innovation, offering a "grand bargain" where developers maintain freedom to innovate while accepting predefined limits when specific risk thresholds are reached. If-then commitments could also help mitigate the "racing" dynamic if all leading developers commit to the same capability triggers. While currently voluntary, their eventual enforcement by an international regime could be a path forward.

4. **Hardware-Enabled Mechanisms (HEMs):** HEMs can support verifiable reporting of key AI training properties (e.g., compute used, location) and enable policy enforcement through methods like offline licensing or geolocking of AI accelerators. These mechanisms could underpin international agreements restricting access to advanced AI hardware to actors adhering to responsible AI development standards. However, their current developmental stage and challenges in tamper-proofing and reliability are significant concerns.

Strategies for Overcoming Challenges: Effective AI nonproliferation requires a robust science of capability evaluation to reliably detect dangerous AI thresholds and robust verification mechanisms. Strategies include:

- **Building Scientific Consensus:** Fostering a shared understanding of AI risks and how to measure progress.
- **Standardization, Auditing, Verification, and Incentivization:** Implementing these processes as part of international agreements.
- **Strategic Diplomacy:** Navigating geopolitical tensions, building trust, and potentially offering incentives for participation in international regulatory regimes, while being mindful of avoiding the pitfalls of the Baruch Plan. This involves seeking "collusion among dominant powers" as seen with the NPT.
- **Focus on Capability-Based Nonproliferation:** Moving beyond hardware restrictions to control dangerous AI capabilities, preventing both horizontal proliferation (spread to actors without safeguards) and vertical proliferation (unchecked escalation of capabilities).
- **Mandatory, Transparent Evaluations:** Incorporating these from the outset, balancing openness and security, and designing adaptive institutions to avoid repeating historical mistakes.

The path to effective international coordination for AI governance is complex, requiring a delicate balance between national interests, technological advancement, and the shared imperative of mitigating existential risks.

Appendix D. Considering Alternate Perspectives: The Case for AI Accelerationism and Critiques of Stringent Regulation

While ControlAI's "A Narrow Path" advocates for a cautious, preventative approach to ASI, it is important to acknowledge a prominent counter-narrative, particularly from proponents of "effective accelerationism" (e/acc) and other techno-optimist viewpoints. This perspective posits that unrestricted technological progress, especially driven by artificial intelligence, is the ultimate solution to universal human problems such as poverty, war, and climate change. Proponents of e/acc view Artificial General Intelligence (AGI) as an inevitable destiny and fundamental for humanity's advancement, advocating for its rapid, unhindered development and deployment. They often express opposition to AI regulation and government intervention in markets, believing that AGI will be most aligned with human interests if multiple AGIs compete against each other in a free marketplace. Some critics of stringent AI regulation argue that it risks stifling innovation and

progress, making companies less competitive internationally. They contend that overly burdensome regulations can create significant economic barriers to entry for new companies, potentially strengthening the hand of incumbents. Furthermore, there is a concern that regulations, given the rapid evolution of AI, can quickly become outdated, failing to keep pace with technological advancements.

This viewpoint also often critiques the "existential risk" narrative itself. Some scholars argue that by focusing on future, speculative capabilities of AI systems, the existential risk discourse may inadvertently distract from more immediate and pressing societal harms of AI, such as bias, misinformation, and job displacement. They suggest that such narratives might align with the interests of technology companies by diverting attention from current, demonstrable issues.

However, proponents of regulation offer counter-arguments. They contend that regulation does not inherently stifle innovation; rather, it can provide clarity and establish ethical guardrails, allowing developers to operate with greater confidence and fostering public trust. By building compliant and ethically sound AI systems from the outset, organizations are better equipped to scale globally and enter regulated markets, potentially turning regulation into a competitive advantage. Furthermore, research suggests that existential risk narratives do not consistently distract from concerns about immediate harms; in fact, public concern for immediate harms often remains consistently higher. The argument is also made that precisely because technologies cannot be uninvented once released, proactive regulation is necessary *now* to prevent irreversible negative outcomes.

While these alternate perspectives highlight important considerations regarding innovation, economic impact, and the focus of AI risk discourse, our red-teaming analysis remains centered on strengthening ControlAI's "A Narrow Path" framework. This is due to its direct relevance to current legislative efforts and the urgent imperative, as articulated by a significant body of experts, to address the potential for uncontrolled ASI to pose an existential threat to humanity.

Appendix E. Extended Policy Comparison Tables

Effective AI governance requires understanding the diverse approaches being adopted or proposed across different jurisdictions. ControlAI's "A Narrow Path" framework, with its focus on preventing Artificial Superintelligence (ASI) development, stands alongside various national and international initiatives. This section provides a detailed comparative analysis of ControlAI's Phase 0 policies against key regulatory frameworks, highlighting their differences in scope, thresholds, enforcement mechanisms, and risk classifications.

Table 1: Comparative Analysis of Key AI Policy Frameworks

Feature / Framework	ControlAI: "A Narrow Path" (Phase 0)	EU AI Act	US Executive Order 14110 (Biden-era)	US Export Controls (BIS)	California SB 1047 (Vetoed)
Overall Goal/Focus	Prevent ASI development for 20 years; keep humanity in control; halt uncontrolled AI development. ¹	Lay foundations for AI regulation in EU; classify AI risks; ensure safety, transparency, and human oversight. ⁵	Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. ⁷	Regulate global diffusion of advanced computing ICs and AI model weights to protect US national security and foreign policy interests. ⁸	Mandate safety testing and disclosure obligations for AI developers. ⁹
Key Prohibitions/ Bans	Ban deliberate ASI development; prohibit dangerous AI capabilities (automated AI research, hacking); No AIs improving AIs; No AIs capable of breaking out of their environment; No unbounded AIs; Limit general intelligence of AI systems. ¹	Prohibited uses (e.g., predictive systems for criminal offenses, social scoring, manipulative AI). ⁵	N/A (focused on standards and safeguards, not outright bans).	Restrict export of advanced computing ICs (ECCNs 3A090.a, 4A090.a) worldwide; New controls on model weights (ECCN 4E091) for models trained with $\geq 10^{26}$ operations, worldwide. ⁸	N/A (focused on testing and disclosure).

Licensing/Regulatory Triggers	Comprehensive licensing system for advanced AI development. ¹ Safety cases required before training ($>10^{25}$ FLOP for training licenses). ¹⁰	GPAI presumption: $>10^{22}$ FLOP (rebuttable). ¹¹ Systemic risk GPAI: $>10^{25}$ FLOP of cumulative compute. ¹¹ Significant modifications ($>1/3$ original compute or $>3 \times 10^{21}$ FLOP) can trigger new obligations. ¹¹	Introduced a compute threshold for reporting requirements and evaluations (revoked). ⁷	Worldwide license requirements for specified ICs and model weights. ⁸	Defined frontier AI models with a compute threshold. ⁷
Scope of Application	Any found system directly improving other found systems (for "no AIs improving AIs"). ¹⁰ Applies to ASI development and precursor technologies. ¹	GPAI models; "high-risk" AI systems; "limited-risk" AI systems (e.g., chatbots, deepfakes); applies to providers and deployers. ⁵	Broadly applies to AI development and use across federal government and private sector. ⁷	Dual-use AI models and ICs; applies worldwide, including Foreign Direct Product Rule (FDPR). ⁸	Frontier AI models and their developers. ⁹

Safety/Transparency Requirements	Require companies to demonstrate AI will not use forbidden capabilities <i>before</i> it is run. ¹	Risk management system, quality controls, record-keeping, transparency obligations, human oversight for high-risk systems. ⁵ Public disclosures for consumers. ⁹ Transparency for GPAI (e.g., training data summaries). ⁵ Disclosure for AI-generated content (deepfakes). ⁵ Cybersecurity protection, serious incident reporting for systemic risk GPAI. ⁹	Robust public transparency, internal oversight (chief AI officers), regular testing (civil rights, accuracy). ¹³	License Exception AIA requires certification that ICs won't be used to train ECCN 4E091 models for entities outside low-risk destinations. ⁸ Red flag guidance for IaaS providers. ⁸	Mandated safety testing and disclosure obligations. ⁹
---	---	--	---	--	--

Enforcement Mechanisms/ Oversight	Direct Institutional Plan (DIP) to convince policymakers; international treaty ultimately. ¹ Relies on technical systems (high-compute training runs detection) and mechanisms (physical kill-switches). ¹⁵	AI Office can demand info/model access, enforce risk mitigations/recalls, impose fines (up to 3% global turnover or €15M). ¹¹ Enforcement remains a challenge. ¹⁶	Federal agencies, NIST frameworks. ⁹	License applications (presumption of denial for high-risk, approval for low-risk). ⁸ License exceptions (AIA, ACM, LPP, NAC, ACA). ⁸ Data Center Validated End-Users (DC VEUs). ⁸	State-level enforcement. ⁹
Risk Classification Approach	Focus on "existential risk" from ASI. ¹⁷	Four categories: Unacceptable, High, Limited, Minimal Risk. ⁶	Focus on "systemic risk" and "critical infrastructure" impacts. ⁷	Based on national security and foreign policy interests, targeting "dual-use" potential. ⁸	Defined "frontier AI models" with a compute threshold. ⁷
Status/Implementation	Policies actively being pushed to lawmakers in UK and US. ¹⁷ Draft bills authored for UK and US. ¹⁴	Adopted June 2024; obligations for GPAI models becoming applicable in August. ⁵	Introduced Oct 2023; revoked by Trump administration. ⁷	Implemented (Oct 2022 for chips, Jan 2025 for model weights). ⁸	Vetoed by Governor Newsom. ⁹

Key Observations from the Comparison:

- **Scope of Ambition:** ControlAI's "A Narrow Path" stands out for its explicit focus on preventing *existential risk* from ASI, a more ambitious and preventative goal compared to the broader risk management or national security focus of other frameworks.
- **Compute Thresholds as a Common Denominator:** Compute thresholds are a recurring mechanism across multiple frameworks (EU AI Act, US EO 14110, US Export Controls, SB 1047) for identifying models requiring greater scrutiny. However, the specific FLOP thresholds and their implications vary significantly, and their limitations as a sole proxy for risk are widely acknowledged.¹⁸
- **Enforcement Challenges:** All frameworks face significant challenges in enforcement, particularly concerning the "black box" nature of AI, the rapid pace of technological change, and the global, distributed nature of AI development involving private actors.¹⁵
- **Fragmented US Landscape:** The US regulatory landscape is highly fragmented, with federal, state, and executive actions often lacking a cohesive, long-term strategy, as evidenced by the revocation of EO 14110 and the veto of SB 1047.⁹ This contrasts with the more centralized, though still challenging, approach of the EU AI Act.
- **Dual-Use Dilemma:** The US export controls explicitly grapple with the dual-use nature of AI, attempting to restrict harmful proliferation while enabling beneficial applications. This mirrors a core challenge for ControlAI's broader nonproliferation goals.⁸
- **Emphasis on Transparency and Safety Cases:** There is a general trend across frameworks towards requiring greater transparency, risk management systems, and safety demonstrations from AI developers, particularly for high-risk or systemic models.

This comparative analysis underscores the complexity of AI governance and highlights the need for ControlAI's framework to address not only its internal consistency but also its interoperability and resilience within a diverse and often conflicting global regulatory environment.

Appendix F. LLM Prompts Used

- "Based on the provided research notes, identify and categorize the key strengths of ControlAI's 'A Narrow Path' Phase 0 framework."
- "From the research notes, extract and detail the technical and operational weaknesses of ControlAI's Phase 0 policies, focusing on challenges in defining superintelligence, limitations of compute thresholds, difficulties in proving 'no forbidden capabilities,' and risks of 'capability externalities.' Provide relevant citations."
- "Analyze the legal and regulatory implementation challenges for ControlAI's Phase 0 policies in the UK and US, covering fragmented

landscapes, legal implications of banning AGI, challenges in licensing dual-use tech and private actors, and data privacy/copyright issues. Cite all relevant snippets."

- "Examine the geopolitical and international coordination gaps in ControlAI's Phase 0, drawing lessons from nuclear non-proliferation (Baruch Plan, NPT), challenges in leveraging technological leads, commitment problems, and feasibility of international treaties against non-state actors. Include citations."
- "Based on the identified weaknesses and gaps, formulate actionable recommendations for legal professionals to strengthen ControlAI's Phase 0 policies. Ensure recommendations are specific and cite supporting research."
- "Based on the identified weaknesses and gaps, formulate actionable recommendations for cybersecurity professionals to strengthen ControlAI's Phase 0 policies. Ensure recommendations are specific and cite supporting research."
- "Based on the identified weaknesses and gaps, formulate actionable recommendations for public sector professionals to strengthen ControlAI's Phase 0 policies. Ensure recommendations are specific and cite supporting research."
- "Draft an executive summary and introduction for a white paper on ControlAI's 'A Narrow Path' Phase 0, incorporating the key findings, strengths, weaknesses, and the purpose of the analysis. Ensure all claims are cited from the provided research notes."
- "Create a table summarizing ControlAI's Phase 0 policy mechanisms and their specific objectives, using information from the provided research notes."
- "Review and refine the entire white paper for coherence, flow, accuracy, and adherence to the specified evaluation criteria (Implementation Feasibility, Policy Effectiveness, Evidence-Based Reasoning). Ensure all citations are correctly formatted."
- "Reformat the provided paper content into the specified template, ensuring all sections are filled, content is maintained, and citations are preserved. Add placeholder author names and affiliations. Ensure the abstract is under 250 words."
- "Integrate a section discussing the alternate view, specifically AI accelerationism and critiques of stringent AI regulation, into the white paper. Ensure it is balanced, cites relevant snippets, and is placed appropriately within the report, likely in the Discussion and Conclusion. Update the abstract and introduction to reflect this inclusion."
- "Create the 'International Coordination Framework Assessments' section for the Appendix, providing a granular analysis of potential international governance models, including the feasibility of a new international body, the role of existing organizations (e.g., UN, IAEA), and strategies for overcoming geopolitical divides and sovereignty concerns. Ensure all information is cited from the provided research notes."