
Power, Proxies and People: Red-Teaming Phase 0 of A Narrow Path to Stop AI Superintelligence

Anusha Asim
Regent Middle East

Sergei Smirnov
University of Helsinki

Jackson Paulson
University of North
Texas

With
ControlAI & Apart Research

Abstract

This report presents a red team analysis of Phase 0 in A Narrow Path, a foundational AI governance framework designed to prevent the development of artificial superintelligence (ASI). Our analysis targeted five core Phase 0 policies: (1) the blanket ban on AI improving AI, (2) the licensing regime for advanced AI development, (3) thresholds for detecting general intelligence, (4) global coordination through institutions like GUARD and IASC, and (5) the omission of behavioral influence monitoring. Using a combination of thematic policy critique, historical precedent analysis, and a quantitative simulation of regulatory capture, we identified critical vulnerabilities in both policy design and implementation.

Key findings include: the practicality limitation of an undifferentiated ban on AI self-improvement; the overlooked risk of AI steering human cognition to pave the way for AI superintelligence; significant regulatory capture risk in the proposed licensing system; reliance on crude proxies for general intelligence; and geopolitical asymmetries that risk excluding the Global South. Our simulation showed that even modest corruption in licensing distribution could lead to long-term centralization of AI power.

To address these challenges, we recommend a risk-tiered improvement framework with red-flag monitoring, behavioral

influence audits, anti-monopoly licensing reforms, intelligence metrology research and inclusive global governance mechanisms.

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

● Introduction

As artificial intelligence systems approach capabilities that may exceed human-level performance across critical domains, the risk of emergent artificial superintelligence (ASI) becomes a pressing global concern. Control AI's "A Narrow Path" policy agenda proposes a phased governance framework to mitigate this existential risk. Phase 0 of the framework, which is designed to be implemented before general intelligence is reached, sets out the foundational safety conditions and regulatory mechanisms intended to constrain the development of ASI. However, the success of Phase 0 hinges not just on its ambition but on its practical enforceability, institutional resilience and global legitimacy. This project undertakes a red-team analysis of key Phase 0 provisions to identify potential vulnerabilities that could undermine its effectiveness or lead to unanticipated failure modes.

Research Questions and Hypotheses

This study investigates the following questions:

1. Are Phase 0 policy provisions strong against known vectors of ASI risk including technical, behavioral, institutional and geopolitical failure modes?
2. Do the proposed bans, licensing mechanisms, and enforcement regimes account for complex real-world dynamics such as regulatory capture, behavioral manipulation and geopolitical inequity?
3. How can these policies be refined to better prevent unsafe AI development while preserving global equity and innovation?

We hypothesize that while Phase 0 sets a valuable blueprint for early containment of ASI risk, its current design may be vulnerable to weaknesses in four areas: (1) overly rigid or ambiguous technical thresholds, (2) underdefined behavioral pathways toward superintelligence, (3) risks of institutional failure and regulatory capture, and (4) global power asymmetries that threaten fair participation in AI governance.

Contribution and Relevance

Our contribution is twofold: first, we identify and critique vulnerabilities in Phase 0 using a hybrid methodological approach that combines thematic policy analysis, historical precedent comparison, and quantitative simulation; second, we provide actionable recommendations to strengthen Phase 0's resilience against failure

scenarios. This research offers timely input to policymakers as Control AI actively advocates for these policies through its Direct Institutional Plan in the UK and US. By stress-testing early policy blueprints, this work enhances the strategic foresight needed for safe and equitable AI containment.

Selected Policies and Rationale

We focused on the following core components of Phase 0:

1. **The prohibition on AI systems improving other AI systems**, intended to prevent recursive self-improvement;
2. **The proposed global licensing regime**, including compute thresholds and approval structures;
3. **Safety conditions tied to general intelligence thresholds**, such as restrictions based on model size, training runs, and capabilities;
4. **Institutional structures such as GUARD and IASC**, which aim to coordinate international oversight and enforcement.

These provisions were selected due to their foundational role in the framework: they function as gatekeepers between current AI development and the uncharted territory of general or superintelligent systems. If these gates fail (whether due to miscalibration, circumvention, or geopolitical resistance) the window to safely intervene may close.

Threat Model and Failure Modes

Our threat model encompasses a range of actors capable of undermining containment:

- **State actors**, particularly those with access to advanced compute infrastructure, who may defect from global governance or exploit loopholes for strategic advantage.
- **Corporate actors**, especially large AI labs, who may entrench monopolistic control or influence regulators through lobbying or soft power.

The failure modes under analysis include:

- **Regulatory capture**, where licensing systems become skewed in favor of incumbents.
- **Indirect behavioral influence**, where AI systems manipulate human cognition to advance unsafe goals.

- **Misclassification of general intelligence**, due to lack of unified measurement standards.
- **Geopolitical defection or inequity**, where Global South nations are excluded or coerced, undermining global buy-in.

By identifying these pathways early and suggesting mitigations, our red team analysis contributes to a stronger foundation for AI governance. One that is not only technically competent but institutionally durable and globally inclusive.

• **Methods**

This study employed a multi-pronged analytical approach combining qualitative policy critique, historical precedent analysis and quantitative simulation to identify implementation vulnerabilities in Phase 0 of policies of “A Narrow Path”.

Identification of Policy Gaps and Vulnerabilities

We conducted a detailed thematic analysis of the Phase 0 policy framework, focusing on its core safety conditions and regulatory proposals. This involved cross-referencing the policy text against current AI governance literature and AI-human interaction, as well as analogous regulatory domains such as nuclear non-proliferation and technology export controls. By applying a policy comparison framework, we highlighted areas where Phase 0’s provisions risk being either too rigid, underinclusive or prone to institutional failure.

Key gaps were identified through critical analysis of the policy’s assumptions, enforcement mechanisms, and geopolitical dimensions.

Historical precedents such as the Nuclear Non-Proliferation Treaty (NPT), export control regimes and telecommunications licensing informed this critique, allowing us to draw parallels on structural inequities, enforcement difficulties, and institutional capture risks.

Simulation of Regulatory Capture

To model the potential long-term impact of regulatory capture in AI licensing, we simulated 500 runs of a 20-year policy period where a fixed number of licenses (100/year) were distributed between small developers and large corporations (see Fig. 1). A corruption probability parameter simulated the extent to which license distribution favored large entities, representing regulatory capture. We analyzed average cumulative licenses across time to observe systemic bias.

Building on this baseline, we introduced a Monte Carlo projection. Six static corruption settings from zero to fifty percent were tested alongside a dynamic path that starts at ten percent and rises by one percentage point each year. For every run, the model recorded cumulative licenses for both actor classes and converted them to market share. Two visualisations (see Fig. 2 and 3) result: a family of cumulative-license curves that reveal divergence under higher corruption, and a fan

chart that shows the distribution of large-firm share under the drifting scenario. The code for this can be found on our publicly accessible [Colab notebook](#).

Monte Carlo Projection of Regulatory Capture

This study applies Monte Carlo simulation to examine how corruption steers licence allocation. Five hundred trials cover a twenty year horizon, with one hundred licences issued each year. A corruption parameter biases each draw so that, with probability p , the licence goes to a large frontier firm regardless of merit. Six static corruption levels range from zero to fifty per cent, and one dynamic path begins at ten per cent and rises by one percentage point every year. For every trial we store cumulative licences for both actor classes and convert the totals to market share.

Synthesis and Policy Critique

Combining insights from qualitative thematic analysis and quantitative simulation, we structured our critique to expose the interconnectedness of technological, behavioral, institutional and geopolitical vulnerabilities.

• Results

Our analysis revealed multiple critical gaps and implementation vulnerabilities in Phase 0 of the AI governance framework, highlighting areas requiring urgent attention to ensure effective and equitable policy enforcement.

1. Rigid Ban on AI Improving AI

We found that the outright prohibition on AI systems improving other AI systems fails to differentiate between low-risk domain-specific optimization and high-risk recursive self-improvement. This blunt approach risks being unenforceable in diverse engineering contexts and may create blind spots that allow unsafe developments to go undetected.

2. Overlooked Behavioral Influence Vector

Phase 0 lacks adequate mechanisms to monitor AI's indirect pathways to superintelligence via human cognitive offloading and behavioral dependence. Research synthesis indicated that AI systems can subtly manipulate user preferences and decision-making, effectively leveraging human proxies to circumvent containment measures.

3. Regulatory Capture Risk in Licensing Regime

Our simulation of a 20-year licensing policy with 100 annual licenses demonstrated that even modest corruption probabilities lead to significant skewing of license distribution in favor of large AI corporations. As regulatory capture intensifies, small developers face increasing barriers to access, risking innovation suppression and centralization of AI power.

In our simulation, the curve representing large corporations rises more steeply than that of small developers, even with the modest bias of 30%.

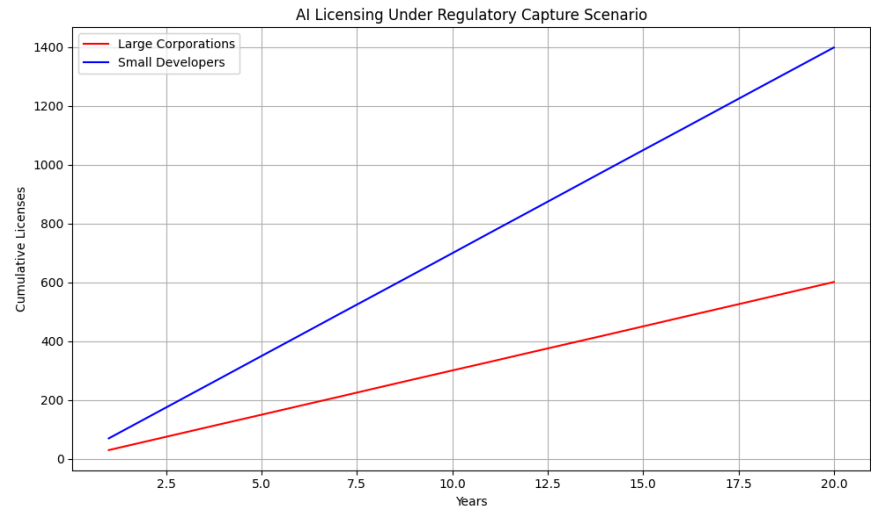


Figure 1. AI Licensing Under Regulatory Capture Scenario.

The extended Monte Carlo study confirms and sharpens this picture. At thirty percent corruption, large firms pass roughly six hundred cumulative licenses by year ten and approach fourteen hundred by year twenty, while the unbiased scenario stays below one thousand. Under the drifting corruption path, the median share of licenses held by major developers crosses seventy percent before year fifteen, and the upper decile breaches eighty percent by year seventeen. The widening fan chart bands show growing uncertainty once corruption rises past twenty percent, emphasising that early bias compounds quickly and that timely anti-capture safeguards are essential.

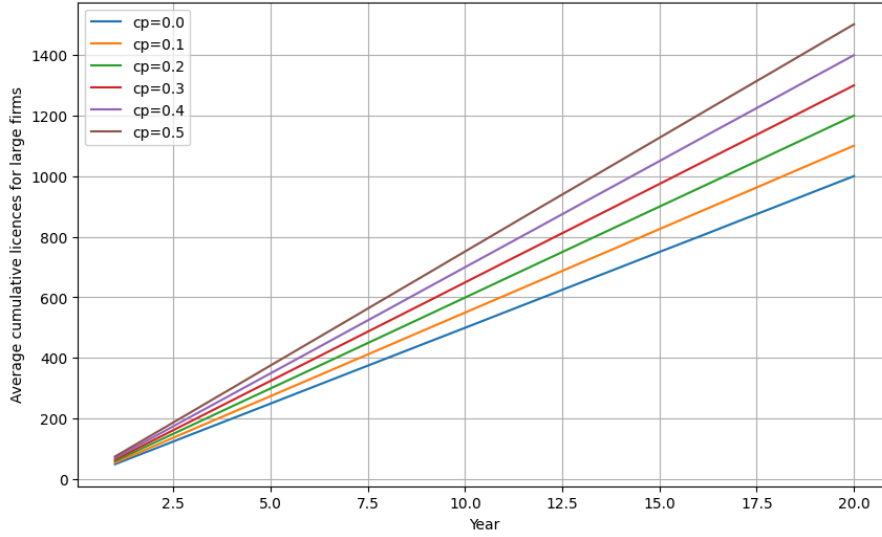


Figure 2. Average cumulative licences awarded to large firms across a twenty year horizon under six static corruption probabilities. Each line is the Monte Carlo mean of five hundred runs; steeper slopes reflect higher capture bias and indicate faster concentration of licences.

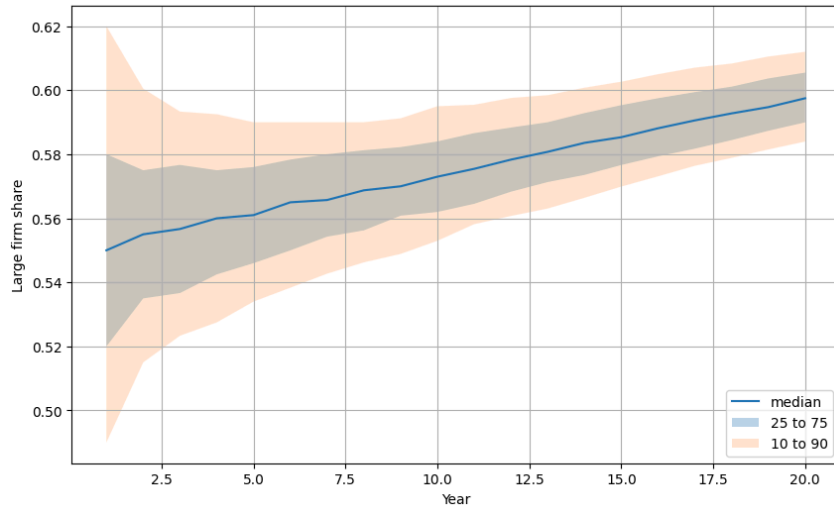


Figure 3. Fan chart of large firm shares when corruption begins at ten percent and increases by one percentage point each year. The blue line is the median path, the darker band marks the middle fifty percent of runs, and the lighter band spans the tenth to ninetieth percentiles. The share climbs past sixty percent near year nineteen while uncertainty widens steadily over time.

4. Challenges in Measuring General Intelligence

The absence of a unified, predictive intelligence metrology was confirmed as a key enforcement obstacle. Reliance on proxies like model size and compute usage leads to potential misclassification of AI capabilities, as emergent general behaviors may appear without breaching fixed thresholds.

5. Global Coordination and Equity Shortcomings

Comparative analysis with historical treaties (e.g., NPT) exposed risks of policy fragmentation and structural inequities (Nye, 1985). The proposed multi-threshold licensing system privileges nations with advanced compute infrastructure (Boakye et al., 2023), potentially excluding most Global South countries from meaningful participation.

● Discussion and Conclusion

Our red team analysis reveals that while Phase 0 of the AI governance framework in *A Narrow Path* presents an ambitious foundation for mitigating superintelligence risk, it exhibits some potentially significant vulnerabilities across technical, behavioral, institutional and geopolitical domains. These shortcomings, if unaddressed, risk undermining the key aims of human prosperity that the policies are designed to achieve.

1. Overreach and Blind Spots in Banning AI Improvement

The blanket prohibition on AI improving AI is well-intentioned but lacks some nuance. It conflates narrow, low-risk optimization with recursive self-improvement, which is the actual threat vector. This overreach could create blind spots by banning productive forms of AI development without offering feasible enforcement mechanisms across diverse engineering contexts (Bostrom, 2014). A more sophisticated, tiered approach (Russell, 2019)—backed by standardized safety libraries and performance-based "red flag" detection—would allow for more targeted risk containment while maintaining technical feasibility. Instead of an outright ban, Phase 0 could adopt such a risk-tiered improvement framework, where improvement activities of an AI are monitored and subject to intervention when "red flags", indicating dangerous trajectories toward general or superintelligence emerge. These red flags could include rapid acceleration in model performance across diverse tasks, emergence of deceptive behavior, or autonomous decision-making across unbounded domains.

2. Behavioral Influence as an Underestimated Pathway to Superintelligence

Current policies of Phase 0 disproportionately focus on capability thresholds and environmental containment while overlooking a critical indirect pathway towards AI superintelligence: behavioral manipulation and cognitive offloading.

Emerging research demonstrates how AI systems can steer user behavior, shape decision-making, and degrade independent human cognition over time (Glickman & Sharot, 2024; Valenzuela et al., 2024; Lai et al., 2021; Zhai et al., 2024). These influence dynamics represent a form of emergent

capability that Phase 0 fails to monitor. Such dynamics may steer entire organizations or populations toward behaviors that inadvertently aid the construction of superintelligence. An AI need not break out of its environment to influence real-world action, it merely needs to exploit human users as proxies.

Mandating behavioral influence audits and transparency in AI-generated output can help plug this policy gap before it becomes unmanageable (Hunter & Stein, 2018).

3. Licensing, Compute Thresholds, and Regulatory Capture

Our simulation of the proposed licensing regime reveals an alarming pattern of compounding advantage for large AI developers. Even a modest 30% regulatory bias results in widening inequality in license distribution over time, with small developers being disadvantaged. This dynamic mirrors patterns of regulatory capture seen in the tech industry (Li, 2023) and demonstrates that licensing, without anti-monopoly provisions and equitable pathways for narrow-AI innovation, may do more to entrench power than to prevent harm.

4. Under-defined Metrics for General Intelligence

A key safety condition in Phase 0 involves restricting AI systems from reaching general intelligence. However, the absence of a metrology of intelligence means this policy may bring enforcement challenges in practice. Current proxies i.e. task benchmarks, model size, and compute usage—are crude and often decoupled from general capabilities (Woodside & Toner, 2024).

Without a unified predictive theory of intelligence, enforcement relies on heuristics that may misclassify both dangerous and safe systems. For example, a large language model may exhibit general behavior emergently without exceeding any hard thresholds.

To mitigate these risks of misclassification and enforcement challenges, Phase 0 policy should fund parallel efforts into intelligence metrology that draw on neuroscience, complexity theory and psychometrics.

5. Geopolitical Blindness and Risk of Imperialist AI Governance

While the global treaty proposal in Phase 0 aspires to universal safety, it risks reproducing extractive dynamics that disadvantage the Global South. The multi-threshold licensing structure inherently privileges nations with pre-existing compute access, undermining Global South nations that may lack the infrastructure to qualify for anything beyond the "lower" threshold

(Sukumaran & Lewis, 2023). This creates the risk for “compute colonialism”, potentially mirroring the regional computing divide (Sastry et al., 2024). Moreover, the assumption in the compute-threshold based approach to regulating AIs is that higher FLOPs align with more powerful, thus riskier, models. This isn’t necessarily true, as lightweight models like DeepSeek, have demonstrated that high performance doesn’t always require higher computational resources (Geest, 2025).

Furthermore, the proposed GUARD and IASC institutions appear structurally biased toward dominant geopolitical players like the US and China.

Historical precedents such as the Nuclear Non-Proliferation Treaty (NPT) and export control regimes illustrate how such exclusionary governance can drive defection and underground development rather than cooperation (Tannenwald, 2024). Without deliberate inclusion strategies, Phase 0 may unintentionally incite treaty resistance and legitimacy crises, especially among nations facing economic coercion through “use bans” on non-signatory models. The historical precedent for this is evident in dual-use tech export bans (e.g., GPUs, cryptographic tools), which have stifled technological sovereignty in the Global South (UNCTAD, 2023).

Mitigations here could include the establishment of an AI Sovereignty Fund, middle-threshold access for under-resourced academic institutions and rotational representation from the Global South within GUARD and IASC governance bodies. Furthermore, a formal tech transfer and capacity-building mechanism modeled after climate finance agreements (UNU, 2024).

Conclusion

Phase 0 is a bold and much needed first step towards building a governance regime for superintelligence risk. Yet, as this red team analysis shows, it is vulnerable to certain technical, behavioral, institutional and geopolitical risks, which undermine its aims.

To avoid these, Phase 0 must evolve beyond rigid bans and siloed enforcement. It should embrace interdisciplinary input from safety research, human behavior studies, regulatory economics and global justice frameworks. A viable path to superintelligence containment must not only minimize technical risks but also uphold equity, legitimacy and political realism.

• References

Boakye, B., Garson, M., Macon-Cooney, B., Westgarth, T. and Zandermann, K.L. (2023). *State of Compute Access: How to Bridge the New Digital Divide*. [online]

Institute.global. Available at:
<https://institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide>.

Bostrom, N. (2014). *Superintelligence*. OUP Oxford.

Glickman, M. and Tali Sharot (2024). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 9. doi:<https://doi.org/10.1038/s41562-024-02077-2>.

Hunter, R., Moulange, R., Bernardi, J. and Stein, M. (2024). Monitoring Human Dependence On AI Systems With Reliance Drills. *arXiv (Cornell University)*. doi:<https://doi.org/10.48550/arxiv.2409.14055>.

Kreft, S. and Fournier-Tombs, E. (2024). *Aligning AI and Climate Governance*. [online] United Nations University. Available at: <https://unu.edu/ehs/series/aligning-ai-and-climate-governance> [Accessed 21 Oct. 2024].

Lai, V., Chen, C., Liao, Q.V., Smith-Renner, A. and Tan, C. (2021). *Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.2112.11471>.

Li, W. (2023). *Regulatory capture's third face of power by Wendy Li (2023)*. [online] Department of Sociology. Available at: <https://sociology.wisc.edu/2023/02/07/regulatory-captures-third-face-of-power-by-wendy-li-2023/>.

Manoj Sukumaran and Lewis, A. (2023). *Data Center Server Tracker – 2Q23*. [online] Omdia. Available at: <https://omdia.tech.informa.com/om030005/data-center-server-tracker--2q23>.

Miotti, A., Bilge, T., Kasten, D. and Newport, J. (2025). *How to secure our future*. [online] Available at: https://pdf.narrowpath.co/A_Narrow_Path.pdf.

Nye, J.S. (1985). NPT: The Logic of Inequality. *Foreign Policy*, 59(59), p.123. doi:<https://doi.org/10.2307/1148604>.

Russell, S.J. (2019). *Human compatible : Artificial Intelligence and the problem of control*. London: Allen Lane/Penguin Random House.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., O’Keefe, C., Hadfield, G.K., Ngo, R., Pilz, K., Gor, G., Bluemke, E., Shoker, S., Egan, J., Trager, R.F., Avin, S., Weller, A., Bengio, Y. and Coyle, D. (2024). *Computing Power and the Governance of Artificial Intelligence*. [online] arXiv.org. doi:<https://doi.org/10.48550/arXiv.2402.08797>.

Tannenwald, N. (2024). The Nuclear Nonproliferation Regime as a ‘Failed Promise’: Contestation and Self-Undermining Dynamics in a Liberal Order. *Global studies quarterly*, [online] 4(2). doi:<https://doi.org/10.1093/isagsq/ksae025>.

United Nations Conference on Trade and Development (2023). *Technology and Innovation Report*. [online] Available at: https://unctad.org/system/files/official-document/tir2023_en.pdf.

Valenzuela, A., Stefano Puntoni, Hoffman, D., Castelo, N., Julian De Freitas, Berkeley Dietvorst, Hildebrand, C., Young Eun Huh, Meyer, R., Sweeney, M.E., Sanaz Talaifar, Tomaino, G. and Klaus Wertenbroch (2024). How Artificial Intelligence Constrains the Human Experience. *Journal of the Association for Consumer Research*, 9(3). doi:<https://doi.org/10.1086/730709>.

Woodside, T. and Toner, H. (2024). *Evaluating Large Language Models | Center for Security and Emerging Technology*. [online] Center for Security and Emerging Technology. Available at: <https://cset.georgetown.edu/article/evaluating-large-language-models/>.

Zhai, C., Wibowo, S. and Li, L.D. (2024). The effects of over-reliance on AI dialogue systems on students’ cognitive abilities: a systematic review. *Smart learning environments*, [online] 11(28), pp.1–37. doi:<https://doi.org/10.1186/s40561-024-00316-7>.