
Red Teaming A Narrow Path: Treaty Enforcement in China (ControlAI Policy Sprint)

Jack Stennett

With
ControlAI & Apart Research

Abstract

This report red-teams A Narrow Path’s international treaty proposal, using China as a stress test. It identifies four key failure modes: lack of independent regulators, compute-centric loopholes, local regulatory capture, and covert circumvention. Drawing on real-world examples—from China’s fintech crackdowns to model-weight leaks and GPU laundering—it argues that current treaty assumptions about regulatory independence and enforceability are unrealistic in non-democracies.

The report recommends reframing “independence” as “verifiable capability,” enabling international inspections, and introducing tools like model provenance ledgers, sub-national sanctions, and chip export incentives.

While no treaty can fully override geopolitical mistrust or arms-race dynamics, well-designed safeguards can reduce the risk of catastrophic defection—especially if they reflect the political realities on the ground.

Keywords: AI policy, treaty enforcement, superintelligence risk, regulatory capture, geopolitical compliance

Introduction

This paper is a red-teaming policy paper focused on the international implementation of *A Narrow Path*, using China as the stress-test case.

The primary target is Policy Proposal 5, *A licensing regime and restrictions on the general and intelligence of AI systems*, and Policy Proposal 6, *International Treaty on Red-Lines*; with secondary references to the challenges of implementing a policy proposal with both aspirational and technical components in China.

Treaties are the central tool for countries to coordinate and prevent the spread of existential risk. All other proposals in *A Narrow Path* depend on this cross-border enforcement. This is especially critical because: a) frontier AI labs operate internationally, and ASI could plausibly emerge in either the US or China; and b) both state and private actors in China and the US face classic arms-race incentives, which discourage unilateral restraint and create pressure to race, free-ride, or defect.

The aim of this red-teaming policy paper is to:

- Provide evidence-based, game-theoretic critique of the treaty model and identify specific flaws that may risk ASI emerging
- Describe a few realistic scenarios where these flaws may result in the emergence of ASI
- Offer alternative components to the policy proposal that mitigate these flaws

I suggest three hypotheses that the Narrow Path team should consider:

- Independent-regulator assumption is infeasible in non-democracies; instead capture and Party control are structural.
- Real-world constraints in China: regulatory capture, principal-agent gaps, and geopolitical incentives will bias actors toward partial or total non-compliance.
- Failure scenarios and threat actors can involve state-led actors, corporate frontier labs or some collusion between the two.

I'll outline key failure modes using evidence from Chinese governance, along with incentive levers drawn from nuclear regulation, biosafety, finance, and anti-corruption regimes. For each, I sketch the institutional resources and technical expertise likely needed for implementation.

Methods

The analysis follows a three-step structure: identifying key flaws in the treaty design, modelling how these flaws could lead to failure, and proposing realistic mechanisms to mitigate those risks. I began by isolating four core assumptions embedded in the proposal. For each, I examined how the assumption might break down in the Chinese context, using historical precedents drawn from nuclear non-proliferation, export controls, and dual-use technology regulation. I used simplified threat models, including payoff matrices that contrast the incentives of a chip-dependent China with a more self-sufficient one, and influence-mapping

diagrams showing how power flows between national regulators, provincial authorities, and frontier labs—highlighting channels for misreporting or capture.

To assess the viability of alternative mechanisms, I drew from comparative policy frameworks in biosafety, finance, and sanction regimes. For each failure mode, I outlined plausible enforcement tools or incentive levers that could shift behaviour, including technical mechanisms and institutional models. These were not treated as exhaustive case studies, but rather as empirical probes to test where the treaty’s assumptions run up against real-world governance and implementation constraints.

Results

“Independent regulators” are unlikely in non-democracies

The Narrow Path text calls for national AI regulators to have “adequate independence from political decision-making”, while many other parts of the text rely on this principle. This principle is sensible in theory, but unrealistic in systems like China’s, where regulators are structurally subordinate to the ruling party (Economy, 2018).

How could it go wrong: *China signs the treaty, and regulators initially comply with international norms; however, they quietly prioritise national goals, allowing increasingly lax oversight of frontier labs. While overseas regulators monitor compute, capabilities, and objectives in line with the other provisions of the proposal, Chinese regulators gradually allow state priorities or industrial policy goals to overtake adherence to safety regimes, leading to the emergence of dangerous models.*

This is not speculative. Official capture of regulators is likely in any sector in China, especially one which is so potentially powerful, and integrated with more powerful government sectors.

In China, AI has been explicitly labelled a core strategic technology and a national priority by top-level CCP bodies ([Concordia, 2023](#)), making it inseparable from national development goals. Regulatory documents such as the CAC’s Generative AI Measures ([China Law Translate, 2023](#)) place political reliability and content control at the heart of governance. Case studies from China’s fintech sector, including the crackdown on Jack Ma and Ant Group ([Reuters, 2024](#)), or sudden crackdown on private education in 2021, show how regulatory action tends to follow political priorities rather than fixed legal mandates.

Because of this, the treaty’s call for “independence” needs to be rethought. A more viable model would require each signatory to establish a competent national authority (see the [Cartagena Protocol, 2000](#)), with verifiable qualifications, technical capacity, and openness to international inspection. The language should shift from “independent” to “verifiably capable.” International inspections, similar to those used by the International Atomic Energy Agency, are needed to play the on-the-ground accountability role that domestic separation of powers cannot. This would avoid triggering sovereignty concerns while still demanding a measurable level of professionalism and transparency. Compliance can also be encouraged through incentives for chip exports, both in the terms of positive and negative

incentives, where access to high-performance hardware is offered conditionally on meeting treaty obligations. The U.S. Commerce Department’s licensing regime for advanced chips, which selectively relaxed or tightened based on partner country behaviour, offers one precedent for such a strategy ([Bis.gov, 2023](#)).

Compute-centric regulation and strategic subversion

One of the core pillars of the Narrow Path proposal is the idea that regulating compute use by setting clear FLOP thresholds for training can meaningfully constrain dangerous AI development. While the appeal is clear (it’s measurable and easy to enforce in theory), relying on this as the only hard metric, especially alongside imperfectly aligned local regulators in politically centralised or decentralised systems, risks focusing too narrowly on reported compute use while ignoring actual capability or intent. This creates a dangerous loophole: advanced models can be trained in ways that technically follow the rules but still undermine their purpose.

How could it go wrong: *A lab independently trains multiple “sub-threshold” model shards using local GPUs ([Yang et al, 2024](#)). Once it passes a basic licensing audit, it merges these into a highly capable frontier model. Local regulators, focused on surface-level compliance rather than underlying objectives or capabilities, overlook the emergent risks and fail to catch the capability leap.*

There is already some evidence that compute-based thresholds are becoming less effective at tracking capabilities. New Claude-Instant/GPT-3.5-Turbo models both demonstrate that smaller, cheaper models can now match the performance of larger systems, and Meta’s LLaMA-2 open weights have been repeatedly fine-tuned by third parties to push beyond their initial specifications ([Lerman and Ladish, 2023](#)). Local fine-tuning, especially on private or grey-market GPUs, is extremely hard to monitor in practice, and algorithmic advances like sparse activation ([Song et al. 2024](#)) can allow labs to do more with less. In this kind of environment, local, imperfectly aligned actors have strong incentives to claim compliance while gaming the spirit of the rules.

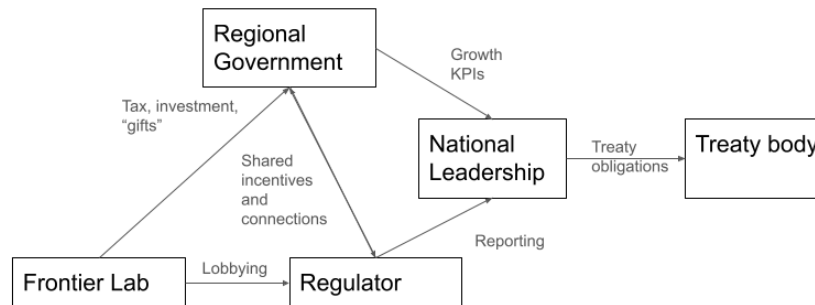
One option is the creation of an independent, international inspector corps, similar to the IAEA, empowered to conduct surprise audits, red-team capability tests, and hardware inventory checks. While such inspections would be politically sensitive, even limited access or voluntary participation could provide useful signals. Another tool is a provenance ledger (e.g. [Tran et al. 2022](#)) that records model weights and updates, enabling treaty bodies to detect clandestine fine-tunes or unlicensed upgrades. By tying labs, provinces, and national authorities to each model release, this system helps track alignment not just in compute use, but in actual behaviour.

Local regulatory capture and principal–agent collusion

Even if national-level signatories to the Narrow Path text support its aims, local regulatory capture remains a problem. The text assumes national regulators will act in the public interest and have the power to oversee frontier labs. But in practice, enforcement is often delegated to local actors whose incentives don’t align. In countries like China, where provincial governments compete for GDP growth and prestige, local regulators may prioritise industrial success over long-term

safety (Concordia, 2023). This creates a classic principal–agent problem: the central government delegates enforcement, but local authorities have reasons to ignore or overlook violations.

Figure 1 – Principal-Agent Dynamics in Chinese AI Treaty Enforcement



How it could go wrong: *A frontier lab seen as a “regional champion” partners closely with local authorities, provides its own evaluation data, helps draft safety standards, and then receives a rubber-stamp licence. As long as its activities stay within formal thresholds, local regulators have little reason to interfere, and may actively protect it from scrutiny.*

There is ample precedent for this kind of regulatory capture. In China, key AI security standards have been drafted by employees of companies like Alibaba and Tencent under the umbrella of MIIT industry groups ([Carnegie Endowment, 2023](#)). The connections between industry and regulators, and the importance of flagship tech firms to local economies, makes capture more likely than less profitable areas.

To mitigate this, treaty mechanisms need to account for local incentives. One option is to build a sub-national sanction ladder. This could target provinces or special zones that are found to be colluding with firms or ignoring treaty obligations. A relevant example comes from the Global Magnitsky Act ([State.gov, 2023](#)); in 2020, the XPCC, a province-scale conglomerate in Xinjiang, was sanctioned without affecting central government ties. Similar tools have been used against individual oligarchs following Russia’s 2022 invasion of Ukraine. These cases suggest that targeted financial and technological pressure below the national level can shift incentives for local actors. This precedent could help design enforcement mechanisms that avoid full diplomatic fallout while still applying meaningful pressure.

Another option is to raise the cost of covert development. For example, labs could be required to post large surety bonds, which they lose if caught breaking the rules. Combined with international inspectors and export controls, this creates a layered system where the risk of being caught comes with real financial and reputational penalties.

Treaty circumvention and covert development are hard to prevent

The text assumes that rule-breaking can be easily detected or stopped, but it overlooks how easy it is to hide or move digital models and hardware—for example, it claims that “regulators should undertake frequent testing of the procedures that AI developers would employ to respond to dangers,” and proposes that “in the case of an emergency, a compute provider and/or an AI developer can be called upon to shutdown the model,” which presumes a level of observability and control that may not be realistic given the ease of covert fine-tuning, grey-market GPU use, and digital weight transfer.

How it could go wrong: *A politically-shielded lab routes compute purchases through shell companies and third countries, trains a near-capable model offshore, and presents it as a benign research tool. Treaty bodies detect nothing until capabilities leak or the model is deployed.*

Advanced models and compute hardware are relatively easy to conceal and move, and China has repeatedly circumvented U.S. export controls via routes through the Middle East and other parts of Asia. In one recent case, Chinese AI developers reportedly smuggled hard drives containing training data into Malaysia and rented foreign servers to train models beyond the reach of export restrictions ([FT, 2024](#)).

To make matters more difficult for compute governance, the incentive to comply with treaty provisions is not fixed. It shifts depending on a state’s strategic position. A country with limited domestic hardware or algorithmic expertise is more likely to follow the rules to preserve access and reputation. In contrast, a more self-sufficient actor (such as China in the near future) may see little cost to defection, especially if enforcement is weak or delayed.

This basic dilemma can be illustrated with a simple payoff grid:

Figure 2 – Payoff grid in chip-dependent and self-sufficient countries

	Weaker State (Chip-dependent)	Stronger State (Self-sufficient)
Comply	+ Access to chips + Legitimacy + Global goodwill	+ Status quo stability – Missed strategic edge
Defect	– Risk of sanctions – Loss of hardware access	+ Strategic edge – Risk of penalties (less strategically significant)

A treaty might work well while a country depends on foreign chips, but fall apart once it becomes self-sufficient. That’s why enforcement needs to keep working even after a state no longer relies on outside hardware. These continued incentives in the case of changing dynamics are neglected in the policy proposal.

Discussion and Conclusion

This paper identified four key risks to international treaty implementation under A Narrow Path, using China as a stress test. These include: (1) the unrealistic assumption of politically independent regulators; (2) the growing gap between compute use and actual capability; (3) local regulatory capture and misaligned incentives at the provincial level; and (4) the ease of clandestine development and offshore training. In each case, I proposed mitigation strategies such as verifiability standards, sub-national sanctions, behavioural audits, and model provenance tracking.

My overall takeaway is that deep value divergence between the US and China makes a fully robust treaty unlikely. No framework will survive open hostility or a renewed arms race. But small adjustments, such as shifting from “independence” to “verifiable capacity,” adding behavioural monitoring, and building in sub-national accountability, can improve compliance odds, even in adversarial contexts.

I hope this red-teaming encourages more attention to domestic political realities, especially in China. Treaties will fail if they don’t account for who enforces the rules, under what incentives, and with what tools.

References

Advanced Computing and Semiconductor Manufacturing Items Controls to PRC. (n.d.). Retrieved 14 June 2025, from <https://www.bis.doc.gov/index.php/policy-guidance/advanced-computing-and-semiconductor-manufacturing-items-controls-to-prc>

AI Regulation is Coming- What is the Likely Outcome? (n.d.). Retrieved 13 June 2025, from <https://www.csis.org/blogs/strategic-technologies-blog/ai-regulation-coming-what-likely-outcome>

Chinese billionaire Jack Ma sees AI future for Ant Group, in rare appearance | Reuters. (n.d.). Retrieved 14 June 2025, from <https://www.reuters.com/technology/artificial-intelligence/chinese-billionaire-jack-ma-sees-ai-future-ant-group-rare-appearance-2024-12-09/>

Gao, Q., Huang, Y., Sui, Y., & Zheng, Y. (2023). The Chinese Communist Party and regulatory transparency in China’s food industry. PNAS Nexus, 2(3), pgad028. <https://doi.org/10.1093/pnasnexus/pgad028>

Global Magnitsky Human Rights Accountability Act Annual Report. (2024, February 23). Federal Register. <https://www.federalregister.gov/documents/2024/02/23/2024-03532/global-magnitsky-human-rights-accountability-act-annual-report>

GPT-4o Mini v/s Claude 3 Haiku v/s GPT-3.5 Turbo: A Comparison. (n.d.). Retrieved 14 June 2025, from

<https://www.vellum.ai/blog/gpt-4o-mini-v-s-claude-3-haiku-v-s-gpt-3-5-turbo-a-comparison>

Lermen, S., & Ladish, J. (2023). LoRA Fine-tuning Efficiently Undoes Safety Training from Llama 2-Chat 70B — LessWrong.

<https://www.lesswrong.com/posts/qmQFHCgCyEEjuy5a7/lora-fine-tuning-efficiently-undoes-safety-training-from>

Maeda, R., & Sieg, L. (2012, July 5). Japan's atomic disaster due to 'collusion:' panel report. Reuters.

<https://www.reuters.com/article/world/japans-atomic-disaster-due-to-collusion-panel-report-idUSBRE8640K4/>

Marquis, C. (2023, July 17). Denise van der Kamp on how China's environmental businesses are regulated. The China Project.

<https://thechinaproject.com/2023/07/17/denise-van-der-kamp-on-how-chinas-environmental-businesses-are-regulated/>

Model Merging in LLMs, MLLMs, and Beyond: Methods, Theories, Applications and Opportunities. (n.d.). Retrieved 14 June 2025, from

<https://arxiv.org/html/2408.07666v1>

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

Nuclear Proliferation Case Studies—World Nuclear Association. (n.d.-a). Retrieved 13 June 2025, from

<https://world-nuclear.org/information-library/appendices/nuclear-proliferation-case-studies>

Nuclear Proliferation Case Studies—World Nuclear Association. (n.d.-b). Retrieved 13 June 2025, from

<https://world-nuclear.org/information-library/appendices/nuclear-proliferation-case-studies>

Park, S., & Albert, K. (n.d.). A Researcher's Guide to Some Legal Risks of Security Research.

Russian Biological Facility Build-up Underscores Need for Enhanced Transparency. (2024, October 28). The Nuclear Threat Initiative.

<https://www.nti.org/risky-business/russian-biological-facility-build-up-underscores-need-for-enhanced-transparency/>

Secretariat of the Convention on Biological Diversity (Ed.). (2000). Cartagena protocol on biosafety to the Convention on Biological Diversity: Text and annexes. Secretariat of the Convention on Biological Diversity.

Sheehan, M. (n.d.). China's AI Regulations and How They Get Made.

Song, J., Huang, K., Yin, X., Yang, B., & Gao, W. (2024). Achieving Sparse Activation in Small Language Models (No. arXiv:2406.06562). arXiv.

<https://doi.org/10.48550/arXiv.2406.06562>

Tran, N. K., Sabir, B., Babar, M. A., Cui, N., Abolhasan, M., & Lipman, J. (2022). ProML: A Decentralised Platform for Provenance Management of Machine Learning Software Systems (No. arXiv:2206.10110). arXiv. <https://doi.org/10.48550/arXiv.2206.10110>

Translate, C. L. (2023, July 13). Interim Measures for the Management of Generative Artificial Intelligence Services. China Law Translate. <https://www.chinalawtranslate.com/generative-ai-interim/>

Tsai, C. (n.d.). Regulating China's Power Sector: Creating an Independent Regulator without Autonomy* | The China Quarterly. Cambridge Core. <https://doi.org/10.1017/S0305741014000381>

Appendix

Prompts used while using LLMs

“Summarise failure modes in China for international AI treaty enforcement.”

“Explain why ‘independent regulators’ are unrealistic in Chinese governance.”

“Describe how compute thresholds might fail to capture model capability.”

“Give historical examples of regulatory capture in China.”

“Propose mechanisms to monitor AI development beyond FLOP limits.”

“What international inspection models could apply to AI governance?”

“Outline a 2x2 matrix of treaty incentives by power and compliance.”

“Find sources for model smuggling, GPU grey markets, and offshore labs.”

“Simplify and rephrase policy critique for use in a red-teaming report.”

“List alternative enforcement tools to support AI treaty compliance.”