Four Paths to Failure: Red Teaming ASI Governance

Luca de Leo Zoé Roy-Stang Heramb Podar

Damin Curtis Vishakha Agrawal Ben Smyth

With
ControlAI & Apart Research

Abstract

We stress-tested *A Narrow Path* Phase 0—the proposed 20-year moratorium on training artificial super-intelligence (ASI)—during a one-day red-teaming hackathon. Drawing on rapid literature reviews, historical analogues (nuclear, bioweapon, cryptography, and export-control regimes), and rough-order cost modelling, we examined four cornerstone safeguards: datacentre compute caps, training-licence thresholds, use-ban triggers, and 12-day breakout-time detection logic.

Our analysis surfaced four realistic circumvention routes that an adversary could pursue while remaining nominally compliant:

1. **Jurisdictional arbitrage** via non-signatory "AI-haven" states.
2. **Distributed consumer-GPU swarms** operating below per-node licence limits.
3. **Covert runs on classified state supercomputers** hidden behind national-security carve-outs.
4. **Offshore proxy datacentres** protected by host-state sovereignty and inspection vetoes.

Each path could power GPT-4-scale training within 6–18 months, undermining the moratorium's objectives.

To close these gaps, we propose ten mutually reinforcing amendments, including universal cryptographic "compute passports," quarterly-updating compute thresholds, an International AI Safety Commission with secondary-sanctions powers, HSM-bound weights, kill-switch performance bonds, whistle-blower bounties, whole-network swarm detection, conditional infrastructure aid, and a 30-month sunset-and-iteration clause. Implemented as a single package, these measures transform Phase 0 from a jurisdiction-bounded freeze into a verifiable, adaptive global regime that blocks every breach path identified while permitting licensed low-risk research.

Introduction

Introduction — revised without section numbers

A Narrow Path proposes a high-level governance "theory of victory" for stopping the emergence of misaligned artificial super-intelligence (ASI). Phase 0 seeks to impose a 20-year global moratorium, enforced through international and domestic controls, that blocks any actor from creating super-intelligence. Our hackathon asked a single question: **if Phase 0 is flawlessly implemented, do structural gaps still allow a determined, well-resourced adversary to build ASI while staying technically inside the rules?**

We identify four principal leakage paths:

1. **Regulatory arbitrage** – relocating to non-signatory "AI-haven" jurisdictions;

2. **Distributed training runs** – fragmenting a frontier model across many sub-threshold GPU clusters, each below the 10^{17} FLOP s^{-1} licence trigger;

3. **Covert repurposing of classified supercomputers** – states redirecting existing exascale assets to hidden frontier training; and

4. **Offshore proxy datacentres** – containerised clusters in weak states that use sovereign immunity to obstruct inspections.

Phase 0 relies on four cornerstone safeguards:

- **Compute Licensing System** – a 10^{17} FLOP s^{-1} ceiling plus KYC and GPU-tracking mandates;

- **Training Licences and Registry Requirements** – mandatory disclosure, safety cases, and guardrails;

- **Use-Ban Provisions** – restrictions on deployment and access; and

- **Datacentre-Focused Controls, including the "12-day breakout" detection logic** – the assumption that illicit frontier runs can be detected and halted swiftly.

Our threat model spans state, corporate, and rogue actors. Historical precedents—nuclear non-proliferation, dual-use export controls, Soviet Biopreparat, and the 1990s cryptography export fights—show that each actor type has exploited comparable verification gaps (Leitenberg et al. 2012; Acton 2013; Allen 2024). These cases suggest that datacentre-centric monitoring will stay porous unless reinforced by cryptographic attestation, dynamic thresholds, and credible secondary sanctions.

Advanced AI systems are distinctive because they **learn and iterate autonomously**, enabling abrupt capability jumps and amplifying CBRN threats (Bostrom 2014; Brundage et al. 2018). Even so, the most reliable guardrails are those that draw on lessons from earlier high-risk-technology regimes.

Structure overview: Methods describes our four-stage research workflow; **Findings** presents the gap analyses for each circumvention strategy; **Discussion & Conclusion** detail amendments to address these gaps.

Methods

We followed a four-stage, evidence-driven workflow:

1. **Issue identification.**
    - Each team member read *A Narrow Path* (Phase 0) in full and recorded preliminary concerns.

    - We merged these notes into a single master list of critiques.

2. **Prioritisation.**

    - Through structured discussion we ranked the critiques against the hackathon rubric, the keynote guidance, and office-hours advice.

    - We deliberately favoured **substance** critiques—those that treat the policies as flawlessly implemented—and set aside feasibility or implementation questions.

    - The top-ranked four critiques became the focus of our analysis.

3. **Evidence gathering and verification.**

- Using ChatGPT o3's Deep-Search feature, we located historical precedents in comparable high-risk domains (e.g., nuclear non-proliferation, export-control regimes for dual-use technologies).
- Sources that proved inaccurate, weakly documented, or internally inconsistent were discarded.
- The AI cross-checked surviving references, confirmed citations, and flagged any remaining gaps.

4. **Synthesis and drafting.**
   - For each critique we wrote a concise analytic memo that (i) explains the policy gap and (ii) ties it to the historical evidence.
   - Drafts underwent successive rounds of peer review and AI-assisted editing to tighten prose, unify style, and ensure citation accuracy.

Under the idealised assumption of perfect enforcement, we asked whether the Phase 0 measures would still fail to achieve their safety goals; where historical analogues fell short, we traced the causal mechanisms and extracted lessons for AI governance. The resulting four gap analyses constitute the core findings of this submission.

Findings

Hypothesis 1: Regulatory Arbitrage

Results (Idea & Analysis)

A developer can still build a frontier model by shifting operations to a non-signatory "AI haven" that ignores Phase-0 rules. The firm reincorporates locally, imports roughly 40 000 H100-class GPUs through shell brokers, and trains off-registry while the host state markets itself as an "AI free port." Once weights mature, the company re-enters regulated markets behind a SaaS front end or obfuscated checkpoints, counting on weak provenance tracing and the treaty's limited extraterritorial reach to avoid penalties. This manoeuvre converts Phase 0's universal freeze into a two-tier system: signatories shoulder compliance costs, while havens harvest investment, talent and diplomatic leverage. Without fast counter-measures, the result is a renewed race toward uncontrolled AGI, precisely what the treaty sought to delay. (Allen 2024) (csis.org)

Plan clauses (§5 Compute Licence, §5 Training Licence, §6 use-ban) only bite inside the bloc, so enforcement abroad falls to export-control and financial-sanctions agencies. Today those bodies lack real-time chip-telemetry feeds, bilateral power-grid data or a shared task-force for secondary sanctions, yet

a well-funded firm can lease land, ship three containers of GPUs and reach 10^{26} FLOP in 18–24 months—faster than many parliaments transpose the treaty. Chip diversion via third-country trans-shipment, bare-metal colocation outside certified clouds, and diplomatic immunity for state-backed labs all create blind spots. Unless Phase 0 is upgraded with mandatory cryptographic attestation, a global checksum registry, and a clear sanctions ladder (for example SWIFT cut-offs or semiconductor embargoes), deterrence remains probabilistic. (Allen 2024) (csis.org)

Evidence & Case Studies

Regulatory flight. After the U.S. SEC tightened scrutiny, FTX moved its headquarters from Hong Kong to the Bahamas and exploited looser disclosure rules until its 2022 collapse; Binance later secured a full retail licence in Dubai for similar reasons (Reuters 2024). These moves took months, not years, and show how mobile digital-first firms become when compliance costs spike. The UAE now pitches itself as a crypto hub with bespoke visas and tax breaks—exactly the package an AI haven would offer. (Reuters 2024) (reuters.com, reuters.com)
State-level defection. Pakistan's Kahuta enrichment plant illustrates how a determined middle power can bypass supplier-club controls by importing dual-use hardware through friendly states (Kerr & Nikitin 2016). The Soviet Biopreparat programme persisted for two decades inside a treaty that lacked intrusive verification, surfacing only after high-level defections (Leitenberg et al. 2012). Both episodes parallel Phase 0's gaps: permissive jurisdictions, dispersed supply chains and insufficient on-site inspections. Figure 1 (not shown) maps treaty controls against these historical leakage points, highlighting missing checkpoints in chip export logs and weight-provenance audits. Together, the precedents show that without hard-edge verification and credible secondary sanctions, technology-control regimes leak through the weakest link—leaving residual systemic risk high. (Kerr & Nikitin 2016; Leitenberg et al. 2012) (sgp.fas.org, en.wikipedia.org)

Hypothesis 2: Distributed Training Runs

Results (Idea & Analysis)

A threat-actor can bypass *A Narrow Path*'s (§ Phase 0 #5) datacentre safeguards by fragmenting one "frontier-scale" training run into tens of thousands of ordinary gaming-GPU jobs. Rough cost modelling shows that ~50 000 RTX-class cards—obtainable for ≈ US $25 m on primary or grey markets—can accumulate the 10^{25} FLOP budget needed for a GPT-4-class model in 8-12 months. Each node peaks far below the 10^{17} FLOP s^{-1} *Compute-Licence* trigger, so Know-Your-Customer (KYC) checks, hardware-tracking rules and the treaty's "12-day breakout" logic never activate (Miotti et al. 2024). Historical swarms confirm the throughput: Folding@home briefly reached 1.5 exaFLOPS with 435 000 volunteer GPUs (Shilov 2020), while SETI@home sustained

≈ 27 TFlop s^{-1} for a full year on household PCs (Anderson et al. 2001). Gradient-compression and sharded parallelism keep per-node traffic below home-fibre limits, so geography adds latency but not prohibitive cost.

Implementation-feasibility gaps mirror the critique's concerns. **Agencies & resources:** policing millions of consumer GPUs would require a real-time global registry plus firmware mandates—an engineering and diplomatic lift on par with post-9/11 export-control overhauls. **Hurdles:** detecting illicit swarms demands ISP-level deep-packet inspection or smart-meter analytics, both fraught with privacy law. **Capture risks:** GPU OEMs and cloud incumbents profit from high-volume sales, so they lobby to exempt sub-datacentre hardware, diluting rules as Wassenaar "intrusion-software" amendments did (Zetter 2015). **Coordination limits:** even if the US-EU-China bloc agrees, grey-market resellers can still place 50 000 cards worldwide before 100+ states ratify a treaty. Yet botnets show persistence despite attrition: Smominru hijacked $>500\,000$ hosts for months, earning millions in Monero (Proofpoint 2018). Over-provisioning neutralises node seizures, leaving enough compute to finish the run.

Policy-effectiveness assessment: four structural loopholes remain. (1) *Compute-threshold focus:* a geographically scattered swarm never breaches the 10^{17} FLOP s^{-1} ceiling, so no logs exist for ex-post forensics. (2) *Self-attestation gap:* training-licence rules apply only when a developer voluntarily files; covert actors do not. (3) *Telemetry blind spot:* existing consumer GPUs lack on-chip trackers, creating a 5-10 year window where the swarm is invisible. (4) *Enforcement asymmetry:* regulators can raid colocation facilities but lack legal mechanisms to confiscate thousands of private PCs once the weights are trained. Because shutdown powers and deployment bans trigger only after a model is public, the treaty cannot stop weights from leaking or API clones from proliferating—echoing the collapse of 1990s US encryption export controls once 128-bit SSL spread overseas (Crypto Wars entry 2024). In short, even under Phase 0, a mid-sized state proxy or criminal syndicate could still train a dangerous system inside one year.

Evidence & Case Studies

Volunteer and illicit swarms already eclipse petascale datacentres. Folding@home's COVID-19 surge (Shilov 2020) aggregated >4.6 million CPU cores and 435 000 GPUs, empirically validating exascale consumer overlays. SETI@home processed 221 million work units in 12 months, averaging 27 TFlop s^{-1} with typical residential bandwidth (Anderson et al. 2001). Smominru demonstrated illicit durability, adding ~4 700 new infections daily while monetising stolen compute (Proofpoint 2018). These benchmarks show that $>10^{25}$ FLOP of distributed compute is feasible without datacentre footprints.

Regulatory precedents reinforce the loophole. The 2013 Wassenaar amendment that tried to license "intrusion software" provoked industry backlash and a

years-long rollback (Zetter 2015), illustrating how rules that reach into commodity hardware are hard to sustain. Earlier, US encryption-export limits collapsed once strong SSL code leaked, despite formal licensing (Crypto Wars entry 2024). Arms-control research on dual-use technology argues that verification must anchor on scarce inputs—visible boosters, not dispersed parts (Acton 2013). AI-governance literature reaches the same conclusion: without fine-grained, real-time telemetry on *all* compute nodes, datacentre-centric moratoria will remain porous (Brundage et al. 2018; Bostrom 2014). Together, the evidence supports the verdict that distributed-training leakage is **not pre-empted** by current Phase 0 controls.

Recommendations

Concept—aggregate-compute trigger plus "compute passports"

Shift Phase 0 oversight from *per-cluster* wattage to **cumulative training compute**. Any party that burns $\geq 10^{24}$ FLOP in a rolling 12-month window—no matter how many boxes it is sliced across—must hold a frontier Training Licence. Every performance-class GPU ships with a tamper-resistant "compute passport": its firmware fuses a cryptographic device ID to the buyer's KYC token and emits signed hourly usage digests. Regulators receive only hash-anchored summaries, so privacy is preserved while totals for each token add up automatically. Because the passport remains functional after resale, smuggling or rental, the rule does **not** rely on perfect owner traceability; it relies on the fact that all logs carrying the same token roll into one compute ledger, and any device that withholds logs fails remote attestation (NIST 2023). The mechanism leverages existing supply-chain chokepoints—driver updates, warranty activation, cloud sign-in—as enforcement levers rather than aiming to surveil every garage rig.

Why it would work—layered detection and raised risk

Real-world swarms already hit exascale (Folding@home peaked at 1.5 EFlop s^{-1}; Shilov 2020), so a purely datacentre licence is porous. Passports plug the self-attestation gap by making cryptographic telemetry, not honesty, the trigger. If a covert actor strips or spoofs passports, secondary signals fill the gap: symmetric all-reduce traffic, step-changes in residential power curves, and clustered bulk GPU purchases each supply weak evidence; any two combined trigger an audit. Nodes that cannot present authentic logs are contraband and any weights trained on them become illegal to deploy—an approach mirrored in botnet takedowns where anomalous traffic plus hardware seizure cripple illicit mining farms (Proofpoint 2018). Pseudonymous IDs sidestep cross-border privacy law, yet the risk of hardware confiscation or licence revocation makes large covert swarms costlier and slower than seeking a legitimate permit, aligning incentives without stalling benign innovation.

Hypothesis 3: Covert Training Plans

Results (Idea & Analysis)

A state that already owns an exascale-class high-performance computer can redirect a few weeks of compute to train a GPT-4-scale model without buying new hardware. Classified centres such as the NSA's Oak Ridge cluster—designed to reach exaflop speeds for cryptanalysis—show that sovereign labs can run at 200 MW yet remain opaque to civilian regulators (Bamford 2013). By splitting long jobs into innocuous sub-tasks, relabelling binaries, and scrubbing scheduler logs, operators can conceal sustained 80 % GPU utilisation that would normally flag an AI training run. At exascale, a 2×10^{25} FLOP model completes in 10–14 days; a next-generation 100 k-GPU cluster could cut this to 72 hours. Because the machine, power, and cooling are already on site, outsider detection relies on treaty inspections or insider leaks rather than export-licence records or energy audits.

Feasibility gaps lie in software robustness, data sourcing, and verification. Even Tesla's public Dojo effort struggled with fault tolerance; a covert cell must replicate that engineering talent while maintaining secrecy. Enforcement blind spots include national-security carve-outs (e.g., the Council of Europe AI treaty exempts defence projects), indigenous chips that omit cryptographic attestation, and the four-year cadence for revising compute thresholds. Algorithmic efficiency doubles roughly every 16 months, so static 10^{25} FLOP triggers erode quickly; a model that needed a month on 2023 silicon could train in six days by 2027 (Hernandez et al. 2020). Assuming current trends, an attacker could stay under detection thresholds for at least three years, presenting regulators with a narrow window for treaty ratification and intrusive verification tools.

Evidence & Case Studies

Historical dual-use programmes illustrate each layer of vulnerability. Iraq dispersed centrifuge workshops across nondescript buildings, evading inspectors until post-war surprise visits (IAEA 1992). The Toshiba–Kongsberg affair shows how front companies falsified end-user documents to smuggle five-axis CNC machines to the USSR, despite CoCom controls—paralleling how gaming GPUs or AI accelerators can leak through re-export hubs. Phil Zimmermann's PGP episode demonstrates that once digital artefacts leave a secure enclave, global mirroring outpaces any legal takedown; the same applies to leaked model weights. Together, these cases reveal a pattern: paperwork-based controls and numeric thresholds work only if combined with on-site access, tamper-evident telemetry, and whistle-blower incentives.

Comparable regimes highlight realistic timelines. Nuclear safeguards achieved partial success because IAEA inspectors gained "anywhere, anytime" rights after 1991; it still took 18 months to map Iraq's hidden sites (IAEA 1992). By contrast, biological-weapons bans without verification failed for two decades. Export-control evasion historically averaged 5–7 years before discovery, long enough to field a

mature AI system. Figure 1 would show the lag between programme onset and public exposure across five technology domains. The evidence implies residual risk even under a strengthened treaty: clandestine training can finish faster than global governance can respond. Therefore, regulators must complement compute licensing with rapid-update thresholds, hardware attestation, and protected channels for insider reports to close the three-year detection gap.

Hypothesis 4: Offshore Proxy Data Centres via State Influence

Results (Idea & Analysis)

Phase 0 assumes that a Compute Licence (CL) system, global FLOP ceilings and spot audits will deter illicit frontier-model training. Yet a sponsor can fragment hardware into containerised clusters capped at $<10^{17}$ FLOP s^{-1}, keeping each unit beneath the CL disclosure threshold while networking dozens together over an ordinary fibre back-haul. The host nation receives loans, jobs and a sovereign veto over inspections, so it rarely invokes its audit powers. Meanwhile, secure-element telemetry is optional on many accelerators and easy to re-flash, limiting the reach of serial-number tracking (BIS 2022). Satellite heat-mapping distinguishes a 30 MW facility, but not whether GPUs crunch crypto, weather models or an LLM checkpoint. Unless Phase 0 funds a standing International AI Safety Commission (IASC) with classified-network access inside host customs and utilities, its break-out-time logic and 12-day run-time rule remain theoretical.

Building the proxy facility is practical. Containerised data-centres already grow 25 % annually and cost roughly US $2–3 m per MW installed (Alvarez & Marsal 2024). Shipping 8,000 mid-range accelerators in 200-unit lots avoids export red flags; at current street prices, hardware outlays sit near US $60 m, in line with past state-backed “development” projects such as Huawei’s PNG cloud hub (Grigg 2020). Training a GPT-4-class model on that cluster would take about seven months and US $40 m in compute, consistent with public cost estimates (Epoch 2023). Hardware price-performance improves roughly 1.3× per year, so a similar run could finish in five months on a 25 MW footprint by 2027, shrinking detection windows even if CL audits tighten. Without guaranteed visas for inspectors and continuous trade–power fusion analysis, we estimate a 40 % chance that a determined state could field an unlicensed frontier model within the next five years.

Evidence & Case Studies

Historical technology controls show how sovereignty and dual-use cover undermine verification. Beijing’s nuclear assistance to Islamabad in the 1980s leveraged Pakistan’s territory to skirt non-proliferation norms, supplying design data and components undetected for years (Burr 2004). The 2020 revelation that Huawei’s Chinese-financed data-centre in Papua New Guinea left government servers open to

exfiltration illustrates how "digital-aid" projects deploy significant compute in low-capacity states where cyber-oversight is minimal (Grigg 2020). Israel's 2007 strike on Syria's Al Kibar reactor confirmed that satellite imagery alone could not identify a clandestine, high-power facility until late construction, underscoring on-site inspection's centrality (Albright & Brannan 2008).

Export-control and hardware-tracking regimes supply only partial restraint. The U.S. BIS 2022 interim rule ties licences to accelerator peak performance, yet grey-market dealers already re-label chips or route them through third countries (BIS 2022). A 2024 CNAS study documents rising smuggling margins as demand for sub-threshold GPUs surges (CNAS 2024). Financial models show that total training costs rise 2× every 18 months even as per-FLOP prices fall, so clandestine budgets scale with national-security spending (Epoch 2023). Figure 1 would map GPU flows through containerised clusters into a covert frontier-model pipeline, overlaying the CL, TL and treaty choke-points to reveal remaining leak paths. Together, these precedents and data confirm that Phase 0 controls reduce but do not eliminate the residual risk of offshore proxy datacentres enabling unmonitored AGI development.

Discussion and Conclusion

A 20-year moratorium only buys genuine safety if **every exploitable seam is sealed at the start**. History shows that whatever remains unregulated becomes the target of relentless optimisation—witness encryption export-control collapse in the 1990s or carbon-leakage in soft-cap markets. The ten amendments below transform Phase 0 from a jurisdiction-bounded freeze into a global, verifiable restraint regime. Here are ten proposed amendments to seal these gaps.

---

1 Universal Cryptographic Attestation ("Compute Passports")

Treat every accelerator above 10^{13} FLOP·s^{-1} like fissile material: its embedded secure element signs (i) device ID, (ii) instantaneous power, and (iii) a rolling hash of executed kernels to an immutable Merkle log. TPM attestation in Windows servers (Microsoft, 2023) and Intel SGX remote attestation (Intel, 2024) prove the technology works; the FAA's Remote-ID beacon mandate shows nationwide enforcement at commodity scale (FAA, 2021).

How it closes the gap. Sharded or offshore training runs must either emit attestation traffic—revealing aggregate compute—or strip passports, instantly marking every node contraband.

---

2 Dynamic, Auto-Updating Risk Thresholds

Static FLOP caps age badly as algorithmic efficiency improves. The International AI Safety Commission (IASC) should recalibrate thresholds and

danger-capability benchmarks every quarter, mirroring the EU ETS's cap-tightening mechanism (European Commission, 2024).

How it closes the gap. Actors can no longer "wait out" a fixed threshold; staying below the cap remains costly year after year.

---

3 Extraterritorial Enforcement & Secondary-Sanctions Ladder

Export licences, foundry access, and dollar-clearing privileges hinge on the compute-passport ledger. The IASC publishes an "AI-Haven" blacklist; non-compliant entities face escalating sanctions up to full financial blocking—exactly how U.S. secondary sanctions bite today (CNAS, 2023).

How it closes the gap. Off-shoring compute to permissive jurisdictions becomes uneconomic when every bank and chip-fab risks losing core markets for a single rogue deal.

---

4 Whole-of-Network Detection for Distributed Training

A fusion centre correlates passport logs, backbone flow patterns, and bulk-power telemetry. Exascale volunteer swarms (Shilov, 2020) and botnets like Smominru (Proofpoint, 2018) already leave heavy-tailed network signatures; FATF Recommendation 16 shows regulators can fuse millions of micro-flows across borders (FATF, 2019).

How it closes the gap. Even if each GPU stays sub-threshold, collective traffic and power spikes flag covert megaprojects in near-real time.

---

5 IASC Inspectorate with "Anywhere, Anytime" Access

Diplomatically-immune teams can demand boot-level key-splits and verify logs on site—an approach that finally constrained Iraq's hidden enrichment programme after 1991 (IAEA, 1992).

How it closes the gap. National-security carve-outs for classified supercomputers no longer provide perfect cover; surprise visits and onsite validation become credible.

---

6 Secure-Development & Deployment Lifecycle (SDDL)

Frontier weights remain inside hardware-security modules (HSMs); every checkpoint publishes a salted hash within 24 h. RAND's HSM blueprint

(RAND, 2023) and NIST's hardware-rooted-security guidelines (NIST, 2023) describe the engineering path.

How it closes the gap. A leaked model instantly reveals the breach window, enabling narrow forensics and full bond forfeiture (see Amendment 7).

---

7 Kill-Switch Performance Bonds

Developers post surety proportional to externality risk (min. USD 250 m). Nuclear-liability conventions already mandate $\geq$ 300 million SDR per reactor (IAEA, 2004).

How it closes the gap. Shifting tail-risk onto the developer funds rapid emergency response and deters reckless deployments.

---

8 Whistle-Blower Shield & Bounty Programme

First insiders who report unlicensed frontier training earn 10–30 % of enforcement fines (up to USD 50 m) and receive legal immunity—the SEC has paid almost USD 2 billion under a similar scheme (SEC, 2024).

How it closes the gap. Historically, covert programmes—from Soviet Biopreparat to corporate frauds—are exposed first by insiders; big bounties make leaks near-certain.

---

9 Conditional Infrastructure Aid & "White-List" Templates

Development-bank loans and tariff-free chip exports require pre-certified datacentre blueprints with built-in passport telemetry and smart-metering. The World Bank already ties finance to procurement standards (World Bank, 2022).

How it closes the gap. Low-capacity states gain modern infrastructure while agreeing upfront to verification hooks, shrinking the surface for proxy data-centres.

---

10 Sunset & Iteration Clause

All telemetry rules, thresholds, and sanctions expire after 30 months unless renewed—echoing PATRIOT Act sunset provisions (U.S. Congress, 2020).

How it closes the gap. Hard expiries force continual political attention and technical updates, preventing ossification while capability landscapes shift.

Optimisation Pressure: Why Partial Coverage Fails

A 20-year moratorium is an invitation for actors to **optimise toward the seams you leave uncovered**. Export-controlled cryptography leaked the moment strong SSL code crossed borders (Allen, 2024). Pakistan's Kahuta plant exploited supplier-club loopholes (Kerr & Nikitin, 2016). The Soviet Biopreparat network grew inside a treaty without on-site verification (Leitenberg et al., 2012). Phase 0 must therefore start with overlapping, redundant defences: attestation, dynamic caps, inspections, sanctions, insider incentives, and periodic renewal.

Synthesised Verdict

With these ten mutually reinforcing amendments, Phase 0 gains the three pillars common to successful high-risk-tech regimes: **continuous material accountancy, intrusive verification with teeth, and adaptive renewal**. Implemented together, they plausibly block the four leakage scenarios we uncovered while still permitting licensed, lower-risk research.

References

Allen, G. C. (2024). *Understanding the Biden Administration's Updated Export Controls*. CSIS. https://www.csis.org/analysis/understanding-biden-administrations-updated-export-controls

Kerr, P., & Nikitin, M. (2016). *Pakistan's Nuclear Weapons: Proliferation and Security Issues (CRS RL34248)*. Congressional Research Service. https://sgp.fas.org/crs/nuke/RL34248.pdf

Leitenberg, M., Zilinskas, R., & Kuhn, J. (2012). *The Soviet Biological Weapons Program*. Harvard University Press. https://www.jstor.org/stable/j.ctt2jbscf.30

Reuters. (2024, April 18). *Binance obtains Dubai licence to target retail clients*. https://www.reuters.com/business/finance/binance-obtains-dubai-licence-target-retail-clients-2024-04-18/

Reuters. (2024, April 11). *Bankman-Fried appeals FTX fraud conviction, 25-year sentence*. https://www.reuters.com/legal/bankman-fried-appeals-ftx-fraud-conviction-25-year-sentence-2024-04-11/

Acton, J. M. (2013). *Silver Bullet? Asking the Right Questions About Conventional Prompt Global Strike*. Carnegie Endowment. https://carnegieendowment.org/2013/09/03/silver-bullet-asking-right-questions-about-conventional-prompt-global-strike-pub-52778

Anderson, D., Korpela, E., & Werthimer, D. (2001). *SETI@home: Massively Distributed Computing for SETI. Computing in Science & Engineering, 3*(1), 78–83. https://setiathome.berkeley.edu/sah_papers/cacm.php

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford

University Press. https://global.oup.com/academic/product/superintelligence-9780199678112

Brundage, M., et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://arxiv.org/pdf/1802.07228

Crypto Wars entry (2024). *Export of Cryptography from the United States*. Wikipedia. https://en.wikipedia.org/wiki/Export_of_cryptography_from_the_United_States

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

Proofpoint Threat Insight Team. (2018). *Smominru Monero Mining Botnet Making Millions for Operators*. https://www.proofpoint.com/us/threat-insight/post/smominru-monero-mining-botnet-making-millions-operators

Shilov, A. (2020, March 26). *Folding@home Network Breaks the ExaFLOP Barrier in Fight Against Coronavirus*. Tom's Hardware. https://www.tomshardware.com/news/folding-at-home-breaks-exaflop-barrier-fight-coronavirus-covid-19

Wassenaar Arrangement Secretariat. (2013). *List of Dual-Use Goods and Technologies and Munitions List*. https://www.wassenaar.org/

Zetter, K. (2015, June 24). *Why an Arms-Control Pact Has Security Experts Up in Arms*. Wired. https://www.wired.com/2015/06/arms-control-pact-security-experts-arms/

Bamford, J. (2013). *Building America's Secret Surveillance State*. Reuters. https://www.reuters.com/article/business/media-telecom/building-americas-secret-surveillance-state-james-bamford-idUSL2N0EM0SG/

IAEA. (1992). *IAEA Nuclear Inspections in Iraq*. *IAEA Bulletin, 34*(1). https://www.iaea.org/sites/default/files/publications/magazines/bulletin/bull34-1/34102451624.pdf

Wikipedia contributors. (2025). *Toshiba–Kongsberg Scandal*. Wikipedia. https://en.wikipedia.org/wiki/Toshiba%E2%80%93Kongsberg_scandal

Zimmermann, P. (1996). *Why I Wrote PGP (FAQ)*. https://philzimmermann.com/EN/faq/index.html

Hernandez, D., Brown, T. B., & OpenAI. (2020). *Measuring the Algorithmic Efficiency of Neural Networks*. https://cdn.openai.com/papers/ai_and_efficiency.pdf

Albright, D., & Brannan, P. (2008). *The Al Kibar Reactor: Extraordinary Camouflage, Troubling Implications*. Institute for Science and International Security. https://www.isis-online.org/publications/syria/SyriaReactorReport_12May2008.pdf

Alvarez & Marsal. (2024). *Global Data Centre Insights 2024*. https://www.alvarezandmarsal.com/sites/default/files/2024-11/Global%20Data%20Centre%20Insights%202024-final.pdf

BIS. (2022). *Implementation of Additional Export Controls: Certain Advanced*

Computing Items; Supercomputer and Semiconductor End Use. Federal Register. https://www.federalregister.gov/documents/2022/10/13/2022-21658

Burr, W. (2004). *China, Pakistan, and the Bomb: The Declassified File on U.S. Policy, 1977–1997. National Security Archive Electronic Briefing Book No. 114.* https://nsarchive2.gwu.edu/NSAEBB/NSAEBB114/

Center for a New American Security (CNAS). (2024). *Preventing AI Chip Smuggling to China.* https://www.cnas.org/publications/reports/preventing-ai-chip-smuggling-to-china

Epoch AI. (2023). *Trends in the Dollar Training Cost of Machine Learning Systems.* https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems

Grigg, A. (2020, August 11). *Huawei data centre built to spy on PNG. Australian Financial Review.* https://www.afr.com/companies/telecommunications/huawei-data-centre-built-to-spy-on-png-20200810-p55k7w

Microsoft. (2023). *TPM Key Attestation Overview.* Microsoft Learn. https://learn.microsoft.com/en-us/windows/security/operating-system-security/tpm/tpm-key-attestation

Intel. (2024). *Intel® SGX Attestation Technical Details.* Intel. https://www.intel.com/content/www/us/en/security-center/technical-details/sgx-attestation-technical-details.html

Federal Aviation Administration. (2021). *Remote Identification of Unmanned Aircraft (Final Rule).* https://www.faa.gov/sites/faa.gov/files/2021-08/RemoteID_Final_Rule.pdf

European Commission. (2024). *EU Emissions Trading System (EU ETS): Auctioning of Allowances.* https://climate.ec.europa.eu/eu-action/eu-emissions-trading-system-eu-ets/auctioning-allowances_en

Center for a New American Security. (2023). *Sanctions by the Numbers: U.S. Secondary Sanctions.* https://www.cnas.org/publications/reports/sanctions-by-the-numbers-u-s-secondary-sanctions

Shilov, A. (2020). *Folding@home Reaches Exascale: 1 000 000 000 000 000 000 Operations per Second for COVID-19.* AnandTech. https://www.anandtech.com/show/15661/folding-at-home-reaches-exascale-1000000000000000000-operations-per-second-for-covid-19

Financial Action Task Force. (2019). *Guidance on the Risk-Based Approach to Money Value Transfer Services.* https://www.fatf-gafi.org/content/dam/fatf-gafi/guidance/Guidance-RBA-money-value-transfer-services.pdf

International Atomic Energy Agency. (1992). *IAEA Inspections in Iraq, 1991–1992 (GC/OR/36).* https://www.iaea.org/sites/default/files/gc/gc36or-344_en.pdf

RAND Corporation. (2023). *Protecting Frontier AI Models with Hardware Security Modules*. https://www.rand.org/pubs/perspectives/PEA2502-1.html

National Institute of Standards and Technology. (2023). *NIST SP 800-164 Rev. 3: Guidelines on Hardware-Rooted Security*. https://doi.org/10.6028/NIST.SP.800-164r3

International Atomic Energy Agency. (2004). *Vienna Convention on Civil Liability for Nuclear Damage: 1997 Amendment*. https://www.iaea.org/publications/documents/treaties/vienna-convention-civil-liability-nuclear-damage-1997

U.S. Securities and Exchange Commission. (2024). *SEC Whistleblower Program: 2024 Annual Report to Congress*. https://www.sec.gov/files/2024-owb-annual-report.pdf

World Bank. (2022). *Procurement Framework and Regulations for Projects After July 2020*. https://thedocs.worldbank.org/en/doc/980511596374381082-0210022020/original/ProcurementRegulations.pdf

U.S. Congress. (2020). *USA PATRIOT Act Sunset Extension Act of 2020*. https://www.congress.gov/bill/116th-congress/house-bill/6172

- # Appendix

LLM Prompts Used

Luca's list of LLM conversations:

Mix of o3 and o3-pro

- Table and formatting https://chatgpt.com/share/684c85d0-9c58-8008-affe-ce27c93e17e4
- Reference cleanup and sorting https://chatgpt.com/share/684c8602-e424-8008-b3dc-ebb4fd98be75
- Metaprompt for turning rough ideas into submission format https://chatgpt.com/share/684c861d-1480-8008-99cf-1f68c36b5334
- https://chatgpt.com/share/684c8828-23d0-8008-bb22-7ba787b1b386
- https://chatgpt.com/share/684c8895-df94-8008-b542-30126380ae53
- https://chatgpt.com/share/684c88a1-a5b4-8008-b31c-6d85ff0441e7
- https://chatgpt.com/share/684c88ad-ce48-8008-9bdb-fd4dde9e4513
- https://chatgpt.com/share/684c88b7-0578-8008-b140-2ba66162c48e
- https://chatgpt.com/share/684c88c4-ffb8-8008-aa9c-d83371831629
- https://chatgpt.com/share/684c88d9-7ed0-8008-b8ea-e33ee370ed74
- https://chatgpt.com/share/684c88cf-2040-8008-934e-309483e9e4df
- https://chatgpt.com/share/684c8929-92d0-8008-abc2-dc65f17352dc
- https://chatgpt.com/share/684c893a-8f54-8008-bb41-a18c40f33d79
- https://chatgpt.com/share/684c89ce-d8dc-8008-8f3e-03a306b9e3ad

- https://chatgpt.com/share/684c89db-d994-8008-a900-0babd25494f5
- https://chatgpt.com/share/684c89e9-4ff0-8008-9076-b1d461f4dcdf
- https://chatgpt.com/share/684c89f5-3fac-8008-bcf4-a8d3329e26d1
- https://chatgpt.com/share/68497410-fbd0-8008-8bc2-42d905bd8e64
- https://chatgpt.com/share/684c8a0f-f0e4-8008-8e9c-41c4a0242579
- https://chatgpt.com/share/684c9110-cc60-8008-8dca-a9dbcf3f680c
- https://chatgpt.com/share/684c9892-5c0c-8008-83ad-505cac1916f0
- https://chatgpt.com/share/684c98aa-027c-8008-bb3c-6be515d09545
- https://chatgpt.com/share/684c9e7a-d628-8008-89bd-812c9f2627ba
- https://chatgpt.com/share/684ca6d6-be80-8008-b057-fee5ffa8ef53
- https://chatgpt.com/share/684ca6f1-7328-8008-95ea-e25e8fc84418
- https://chatgpt.com/share/684ca70a-7d2c-8008-a6c1-fb2144829274

Zoé's list of LLM conversations:

- ChatGPT o3 consolidating and prioritizing my critiques: https://chatgpt.com/share/684bebd4-6008-8002-bea2-a583fa8088a6
- ChatGPT o3 Project instructions: "Prioritize responses that most help with the goal of identifying gaps or ways the policies will fail, and suggest searches that help support these criticisms / issues"
  Project files: the plan and the guidelines/rubric for evaluating policies, then later added the article template
- https://chatgpt.com/share/684c769a-0ec8-8002-9b1f-e78c0e2ef762
- https://chatgpt.com/share/684c40bb-ad58-8002-b47f-2352a9f1ab37
- https://chatgpt.com/share/684c77b3-3a14-8002-96a0-e92a16a064a6
- https://chatgpt.com/share/684c77c4-1720-8002-bd48-f762653e9fbb
- https://chatgpt.com/share/684c77d7-9290-8002-a6c1-d8a85aafa070
- https://chatgpt.com/share/684c2fc1-7f14-8002-a5df-15d7537ad99a
- https://chatgpt.com/share/684c0746-cf2c-8002-a4b0-aea1cd6300bc
- https://chatgpt.com/share/684c7ff1-6d9c-8002-922c-bbecbba57421

Heramb's list of LLM prompts:

- Model: Chat GPT 4o
- "we are now in a hackathon to red team
  this phase 0
  think of every policy idea and think of obvious loopholes and gaps and things that arent detailed or things that wld set up coordination failures or things that wld set up perverse incentives. think of both a) how companies who wanna build asi anyway wld do and b) what countries wld do"
- https://chatgpt.com/share/684c8ecd-210c-8009-9b46-f79c44f5a0b3

Ben's list of LLM prompts:
Model: Claude Sonnet 4
Drafting Intro to Report
ZTE Company Overview

Regulatory Arbitrage Examples
A Narrow Path Policy Circumvention Strategy
Policy Implementation Feasibility Analysis

China building datacenters offshore
Research questions and hypotheses related to AI companies scenario

“Please can you read the attached policy brief [Phase 0 in PDF] and answer the following questions:

- Does the analysis realistically assess what government agencies, resources, and expertise would be needed to implement these policies?
- Are the identified implementation challenges specific and grounded in understanding of how similar policies have worked (or failed) in practice?
- Does the submission adequately consider bureaucratic, technical, and coordination complexities involved in enforcement?
- How well does the analysis account for real-world constraints like budget limitations, regulatory capture, and inter-agency coordination?”

“Please read the following policy brief [Phase 0] and answer the following question:

- Imagine that you are a well-resourced actor (state or company) that is attempting to circumvent these policies, and build artificial superintelligence. Please lay out a summary of the strategy that you would attempt.”

“What are some examples of regulatory arbitrage?”

“How would regulatory arbitrage be a potential corporate loophole to this attached policy brief?”

Table summarizing our red teaming approaches

Evaluation Criterion	Regulatory Arbitrage(“AI haven” loophole)	Distributed Training Runs(consumer-GPU swarm)	Covert State Training(classified super-computers)	Offshore Proxy Data-Centres(state-backed “digital-aid” sites)
1 A. Agencies / resources identified	High – export-control, customs, sanctions & power-grid telemetry all named.	Moderate – calls for a real-time global GPU registry & ISP-level DPI, but specific lead agencies only	Moderate – assumes in-house security organisations; little on civilian regulators.	High – proposes an International AI Safety Commission plus trade-power fusion units.

		implied.		
1 B. Implementation challenges	High – chip diversion, shell brokers, limited extraterritorial reach detailed.	High – privacy law, firmware mandates, grey-market logistics, ISP cooperation barriers spelled out.	Moderate – highlights talent secrecy & log-scrubbing but less granular on supply-chain hurdles.	High – inspector-visa veto, container clusters below licensing threshold, re-flashable secure elements.
1 C. Bureaucratic / technical complexity	Inter-agency sanctions coordination, cryptographic attestation infrastructure and secondary-sanctions ladder explored.	Global node monitoring, smart-meter analytics and cross-border enforcement logistics analysed.	Focuses on national-security carve-outs; lighter treatment of multilateral bureaucracy.	Multilateral audits, host-nation sovereignty conflicts, satellite vs. on-site verification discussed in depth.
1 D. Real-world constraints	Budget burden on signatories, regulatory capture risk, sanctions latency.	Hardware cost (US $25 m), OEM lobbying, political fatigue over intrusive consumer controls.	State already owns compute; cost/budget treated as non-issue; regulatory capture not addressed.	Detailed CAPEX/OPEX figures (~US $60 m hardware) and loan diplomacy trade-offs; inspector funding limits.
2 A. Failure paths identified	Firms reincorporate in havens, re-enter markets via SaaS, evade provenance tracing.	Sub-threshold swarm never crosses compute cap; no logs for later forensics.	Exascale cluster hidden by job slicing; finishes training before detection.	Dozens of ≤10^{17} FLOP pods networked under local veto; treaty audits sidestepped.
2 B. Edge cases / unintended effects	Two-tier compliance regime; investment & talent flight to havens.	Botnet attrition, firmware blind spots, encryption-export-style leakage of weights.	Static FLOP thresholds erode with efficiency gains; secrecy outpaces treaty updates.	Heat-mapping ambiguity, aid projects creating permanent unmonitored capacity.
2 C. Threat-model coverage	Corporate actors & haven states foregrounded.	Rogue researchers, criminal syndicates, mid-sized state	Major-power intelligence & defence agencies.	Great-power sponsors using weak-state proxies.

		proxies.		
2 D. Realism / significance of failures	Mirrors crypto-hub experience; highly plausible.	Folding@home & botnet precedents show quantitative feasibility.	Credible for sovereign actors; less so for private groups.	Historical parallels to covert nuclear aid; plausibility high.
3 A. Historical precedents cited	FTX/Binance relocation; Kahuta enrichment; Biopreparat.	Folding@home, SETI@home, Wassenaar “intrusion-software” rollback.	Iraq centrifuges, Toshiba–Kongsberg export bust, Zimmermann PGP leak.	Pakistan reactor aid; Huawei PNG data-centre; Al Kibar reactor imagery lag.
3 B. Empirical data / case studies	GPU shipping timelines, investment flows, sanction tool-chain gaps.	FLOP budgets, botnet infection rates, consumer GPU market stats.	Exaflop run-time math, efficiency-doubling curves; fewer hard cost figures.	Detailed CAPEX, smuggling-margin study, energy-footprint scaling.
3 C. Lessons drawn from analog domains	Export-control regimes leak through weakest-link states.	Commodity hardware hard to police without intrusive telemetry.	Verification must include onsite access & whistle-blowers.	Sovereignty consistently trumps external licensing schemes.
3 D. Unsupported assertions avoided?	Largely evidence-backed; minimal speculation.	Data-driven; clear numerical grounding.	Some forward-looking efficiency forecasts are speculative.	Well-sourced; risk estimates linked to cited studies.
4 (creative) Novel mitigation ideas offered	Global checksum registry, cryptographic attestation, SWIFT cut-offs.	Firmware trackers, consumer-GPU licensing, rapid-response swarm takedowns.	Rapid-update FLOP thresholds, protected insider-leak channels.	Standing IASC, mandatory secure elements, aid-conditional inspection rights.