
Modelling the impact of verification in cross-border AI training projects¹

Fabio Marinello
Independent

With
Apart Research

Abstract

This paper develops a stylized game-theoretic model of cross-border AI training projects in which multiple states jointly train frontier models while retaining national control over compute resources. We focus on decentralized coordination regimes, where actors publicly pledge compute contributions but privately choose actual delivery, creating incentives to free-ride on a shared public good. To address this, the model introduces explicit verification mechanisms, represented as a continuous monitoring intensity that improves the precision of noisy signals about each actor's true compute contribution. Actors choose compute provision strategically, balancing the benefits of higher system capability against compute costs and the risk of detection and punishment. Simulation results reveal several governance-relevant findings: (1) verification effectively deters free-riding, with cheating decreasing smoothly as monitoring intensity increases; (2) stronger verification accelerates collective capability development by approximately 27%; (3) a non-monotonic "verification valley" exists where intermediate monitoring levels produce worse welfare outcomes than either minimal or maximal verification; (4) verification precision (measurement noise) is a more influential policy lever than penalty magnitude for controlling cheating behavior; and (5) verification regimes create distributional consequences, with smaller actors benefiting while larger actors bear net costs. These findings suggest that policymakers designing international AI governance institutions face a commitment problem: half-measures in verification are counterproductive, and effective regimes require either accepting some free-riding or investing substantially in monitoring infrastructure.

¹ Research conducted at the [Technical AI Governance Challenge](#), 2026

Keywords: cross-border AI governance; verification and monitoring; game theory.

1. Introduction

Cross-border AI projects. In the following, we define a global AI project as a durable, cross-border initiative in which multiple states—and relevant firms and laboratories under their jurisdiction—jointly build, govern, and operate a component of the frontier AI ecosystem under binding commitments embedded in a precise and enforceable governance framework. In principle, such a project could target a wide range of objectives: it could be oriented toward fundamental research (for example, developing learning paradigms beyond the transformer family), toward AI safety research (for example, interpretability, auditing, and guardrail development), or toward shared evaluation and incident response infrastructure; see [Ho et Al, 2023](#) for a breakdown of possible international institutions for AI development. In this paper, we focus on projects aimed at training and deploying frontier AI models, addressing the technical constraints and governance challenges associated with cross-border endeavours.

Institution design. There are at least two distinct motivations for cross-border coordination in frontier AI development. First, for middle powers, such an endeavor entails large fixed costs in compute, operations, and safety infrastructure that do not scale efficiently to purely national programs. Acting independently leads to duplicated capital investments, heightened exposure to supply chain and political shocks, difficulty internalizing cross-border safety and reputational externalities, and limited capacity to credibly signal safety commitments or shape global standards. Second, coordination can function as a race-stabilization mechanism even among leading AI developers. When frontier scaling is slowed or paused due to safety concerns, unilateral restraint is inherently unstable under competitive uncertainty. A shared international project can transform restraint into a collectively governed continuation of development, aligning pacing, safety thresholds, and release conditions while reducing incentives to defect.

Several organizational designs could support such coordination, each with distinct tradeoffs and challenges. One approach is to pool resources into a centralized AI facility in which talent, compute, and data are colocated. Another is to preserve national control over resources while coordinating training through decentralized paradigms such as [federated](#) or periodically synchronized learning. We focus on the latter design. Although technically more complex, it is more revealing from a game-theoretic perspective, as it directly confronts issues of sovereignty, incentive compatibility, and strategic interaction. Accordingly, we assume the technical feasibility and economic viability of decentralized coordination and introduce a stylized mathematical model that isolates the core mechanisms through which such an arrangement could operate.

Regardless of the institutional design, participants in cross-border AI training projects have rational grounds to distrust one another, even when they are broadly aligned on the objective of developing safer and more capable frontier systems. Because the resulting model constitutes a shared public good, each actor faces incentives to free-ride by underproviding compute relative to its pledged contribution while continuing to benefit from others’ efforts. Verification mechanisms, both in-person or hardware-enabled, can be used to mitigate such opportunistic behavior by improving observability and enforcement of commitments. The central aim of our model is to analyze how the presence and strength of verification affect incentives to cooperate, belief formation about others’ behavior, and the overall feasibility and stability of sustained joint training arrangements.

Our main contributions are:

1. *Building a game-theoretic model to describe a form of cross-border decentralized training,*
2. *Studying the impact of verification methods on cooperation in such a setting.*

2. Related Work

A small but growing body of research applies game theory to core problems in AI governance. *A Game-Theoretic Framework for AI Governance* models interactions between regulators and AI developers as a Stackelberg game, highlighting how commitment power and enforcement shape equilibrium compliance. Complementing this perspective, a recent RAND report, *Strategies and Detection Gaps in a Game-Theoretic Model of Compute Governance*, studies how strategic behavior and limited observability affect the effectiveness of compute-based monitoring regimes. Earlier work by Armstrong, Bostrom, and Shulman develops a stylized game-theoretic model of an AI arms race, showing how competitive pressure to be first can incentivize risk taking and underinvestment in safety.

More broadly, compute governance is widely viewed as a central lever for technical AI governance, motivated by the empirical link between training compute and frontier capabilities and by the relative observability of large-scale compute use. This perspective underlies a range of proposals for thresholds, reporting requirements, and monitoring regimes tied to training runs. In parallel, several authors have proposed international institutions and cooperative mechanisms to manage cross-border externalities and stabilize race dynamics. However, much of this literature remains institutional or qualitative, leaving open how verification quality and repeated interaction shape the feasibility of sustained international cooperation.

3. Methods

Basic architecture. We model N actors, indexed by $i = 1, \dots, N$. Time proceeds in discrete rounds $t = 1, \dots, T$. All computation happens locally within each actor's infrastructure: cross-border interaction is limited to public announcements of compute commitments, shared observations of capability progress, and shared signals from verification mechanisms.

Each actor commits ex ante to a fixed per-period compute contribution $\hat{c}_i \geq 0$ (measured in FLOP). In each period, the actor then chooses a private action $a_{i,t} \geq 0$, corresponding to the compute actually delivered. The divergence between these quantities captures the scope for strategic under-delivery. The total effective compute delivered in round t is therefore $A_t = \sum_{i=1}^N a_{i,t}$, and similarly the effective pledged compute is $\hat{A}_t = \sum_{i=1}^N \hat{c}_i$.

Motivated by empirical scaling laws, we model the relationship between cumulative invested compute $I_t = \sum_{s=1}^t A_s$ and system capability using a simple power-law specification, $q_t = \lambda I_t^\gamma$, where $\lambda > 0$ is a scale parameter and $\gamma \in [0, 1]$ governs the degree of diminishing returns to compute. However, actors do not have direct access to the true capability level. Benchmarks have measurement error, test sets vary, and performance fluctuates; at best, they can perform a measurement subject to noise and produce each round a public observation

$$o_t = q_t + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2).$$

Noise in q_t obscures the link between observed progress and actual compute, making defection harder to detect and increasing reliance on explicit verification rather than outcome-based inference.

Verification mechanism. Since delivered compute $a_{i,t}$ is private, we need mechanisms to verify contributions. These might include hardware attestation, energy metering, or on-site third-party audits. We summarize verification effort and technology with a scalar $v_t \in [0, 1]$. Verification produces a public signal for each actor:

$$r_{i,t} = a_{i,t} + \xi_{i,t}, \quad \xi_{i,t} \sim \mathcal{N}(0, \tau_i^2(v_t)),$$

where the noise variance $\tau_i^2(v_t)$ decreases with verification strength; a simple form is

$$\tau_i^2(v) = \frac{\tau_{i,0}^2}{1 + \beta_i v}, \quad \beta_i > 0.$$

Higher v_t means more resources devoted to monitoring (better hardware-enabled attestation, trusted metering, audits, inspections), yielding more accurate signals of each actor's actual compute delivery. Its value can be fixed and set exogenously, or more generally it can be endogenous and time-varying.

Actors publicly pledged $\hat{c}_{i,t}$; we want to know whether they delivered roughly that amount. We thus compute a score

$$Z_{i,t} := \frac{\hat{c}_i - r_{i,t}}{\sqrt{\tau_i^2(v_t)}}.$$

which intuitively measures how many standard deviations below the pledge the observed delivery is. We then compare $Z_{i,t}$ to a fixed threshold z_α , the $(1 - \alpha)$ -quantile of a standard normal distribution; for example, if $\alpha = 0.05$ is the false positive rate we are willing to tolerate, $z_\alpha \approx 1.64$. This provides a signal of whether an actor is defecting or keeping to its word.

Payoff structure. Detection can have several possible consequences, from expulsion to a penalty that might be fixed or proportional to the deviation from the pledged amount of compute. Let actor i 's utility in round t be

$$u_i(t) = b_i(q_t) - k_i(a_{i,t}) - P_i \mathbf{1}\{Z_{i,t} > z_\alpha\}.$$

where $b_i(q_t)$ is the benefit from capability, $k_i(a_{i,t})$ the cost of providing compute, and $\mathbf{1}(x > 0)$ is the indicator function over the interval defined by $x > 0$. Actions are chosen to maximize expected utility based on public history $H_t = \{((\hat{c}_j)_{j \leq N}, o_s, (r_{j,s})_{j \leq N})\}_{s \leq t}$, giving

$$a_{i,t} \in \arg \max_{a \geq 0} \mathbb{E}[u_i(t) \mid H_{t-1}, a_{i,t} = a]$$

where the expected value can be expanded into

$$\mathbb{E}[u_i(t) \mid H_{t-1}, a_{i,t}] = \mathbb{E}[b_i(q_t) \mid H_{t-1}, a_{i,t}] - k_i(a_{i,t}) - P_i \cdot \Pr(Z_{i,t} > z_\alpha \mid a_{i,t}, v_t).$$

Myopic per-round optimization can be extended to infinite-horizon discounted optimization using standard game-theoretic tools such as the Bellman equation.

4. Results

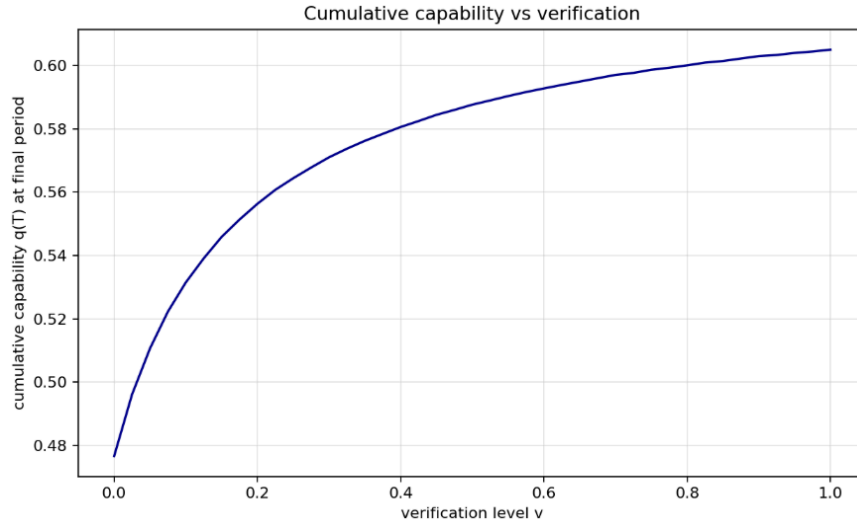


Figure 1 shows how verification affects the total AI capability achieved. At $v=0$, capability is lowest (~ 0.475) due to widespread cheating. At $v=1$, capability reaches ~ 0.605 — about 27% higher. This demonstrates that effective verification doesn't just ensure fairness; it accelerates collective capability development by ensuring actors contribute their pledged resources.

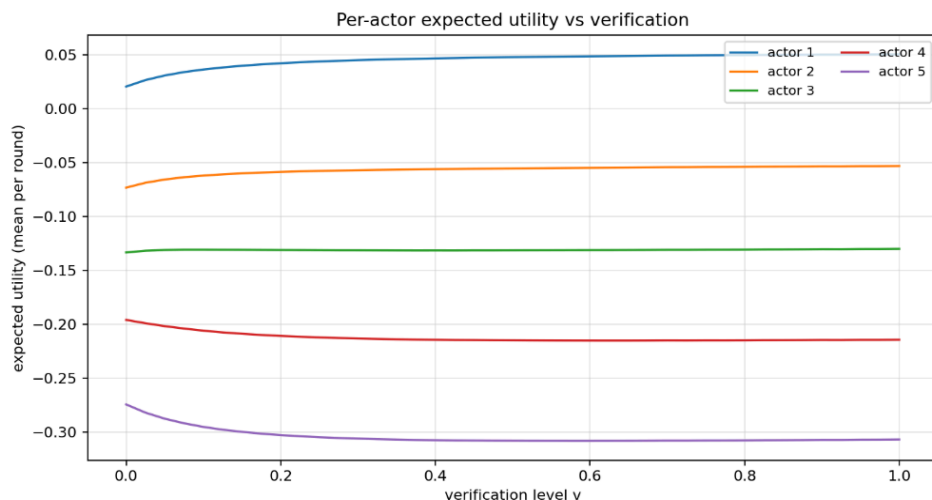


Figure 2 shows winners and losers under different verification regimes. Actor 1 (smallest pledge, lowest cost) has positive utility; actors 3-5 have negative utility (costs exceed benefits). Utilities slightly improve with higher v for most actors. This reveals distributional consequences — smaller actors benefit while larger actors bear net costs, raising equity concerns in governance design.

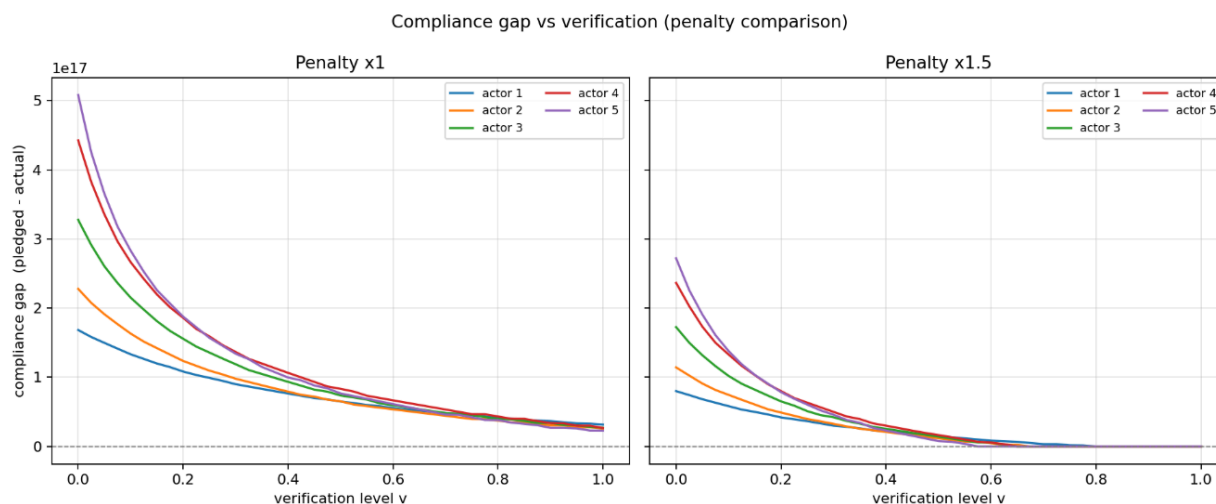


Figure 3 shows how under-contribution decreases as verification intensity increases. At $v=0$, actors cheat substantially (gap up to 5×10^{17} FLOP), with larger-pledge actors cheating more in absolute terms. By $v=1$, all actors approach full compliance. This demonstrates that verification is an effective deterrent: stronger monitoring reduces free-riding behavior in a smooth, predictable manner.

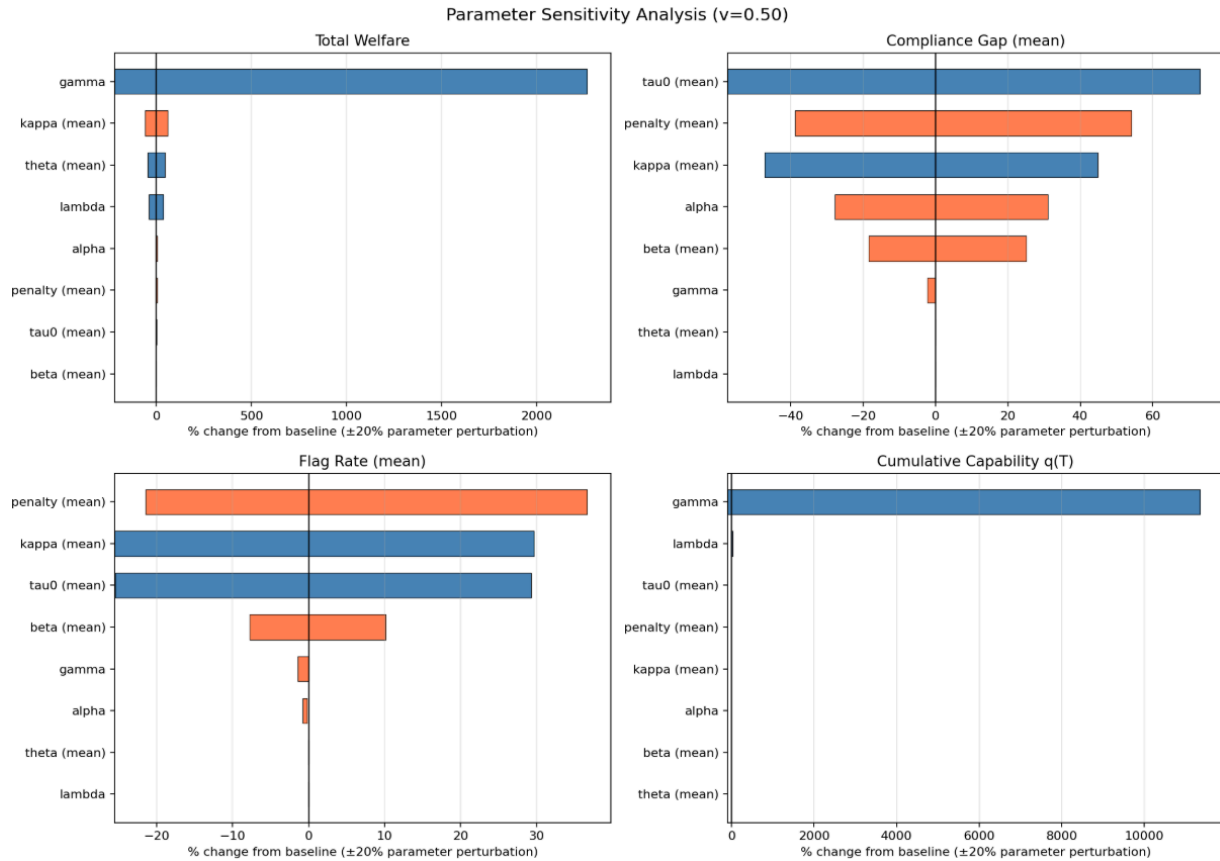


Figure 5 is a sensitivity analysis of the model's parameters. Gamma (scaling exponent) dominates welfare and capability outcomes, but this is a technology fundamental, not a policy lever. For policy-relevant parameters, tau0 (measurement noise) is the most influential for compliance behavior, followed by penalty and kappa (cost). This shows that improving verification precision is more effective than adjusting penalties for controlling cheating. The test threshold alpha affects compliance but has negligible welfare impact. Blue bars indicate parameters where increases improve the outcome; coral bars indicate where they worsen it.

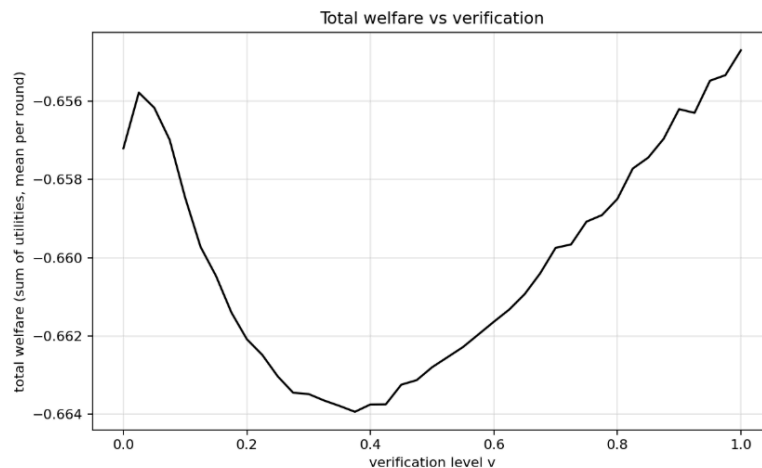


Figure 4 reveals a non-monotonic "verification valley:" welfare first decreases ($v=0$ to $v\approx 0.4$), then increases ($v\approx 0.4$ to $v=1$). The worst welfare occurs at intermediate verification. This suggests a policy dilemma: either commit to low verification (accept cheating, avoid penalty costs) or commit to high verification (enforce compliance); half-measures perform worst.

5. Discussion and Limitations

The verification-compliance relationship. Our primary finding is that verification intensity and compliance exhibit a smooth, monotonic relationship: as monitoring precision improves, free-riding decreases predictably. This suggests that verification mechanisms can be calibrated to achieve target compliance levels, providing policymakers with a tractable governance lever. Importantly, verification and penalties function as partial substitutes—a 50% penalty increase produces roughly equivalent deterrence to a moderate verification increase—offering flexibility in institutional design.

The verification valley. Perhaps the most policy-relevant finding is the non-monotonic relationship between verification and welfare. Intermediate verification levels produce worse outcomes than either minimal or maximal monitoring. This "verification valley" arises because moderate monitoring is insufficient to deter cheating but still imposes penalty costs when violations are detected; institutions should therefore either commit to robust verification infrastructure or explicitly accept a degree of non-compliance.

Distributional consequences. The model reveals that verification regimes create winners and losers. Smaller actors (with lower pledges and costs) achieve positive expected utility, while larger actors bear net costs even under full compliance. This distributional asymmetry raises equity concerns and suggests that sustainable international cooperation may require side payments or differentiated obligations to maintain participation by larger contributors.

Parameter sensitivity. Sensitivity analysis identifies verification precision as the most influential policy-controllable parameter for compliance outcomes, followed by penalty magnitude and actor cost structures. Notably, the scaling law parameters dominate welfare and capability outcomes but are technology fundamentals beyond policy control. This suggests that governance efforts should prioritize investments in monitoring technology over fine-tuning penalty schedules.

Limitations. The model is deliberately stylized and abstracts away many technical, organizational, and political frictions. Verification is represented by a single scalar, collapsing what are in practice layered, institution-specific monitoring and enforcement mechanisms. Compute pledges and punishment schemes are treated as exogenous, rather than as objects of negotiation. The analysis employs myopic per-round optimization rather than fully dynamic strategies with reputation

effects. Actor heterogeneity is limited to parameter differences rather than modeling distinct strategic types (e.g., committed cooperators versus opportunistic defectors). Finally, the model assumes actors can observe the verification level and respond optimally, which may overstate behavioral responsiveness in practice.

Future work. Natural extensions include endogenizing verification intensity by introducing explicit monitoring costs, modeling richer enforcement regimes with graduated sanctions and exclusion, incorporating repeated-game dynamics such as reputation formation and coalition behavior, and grounding parameters in empirical evidence from existing international monitoring regimes (e.g., nuclear safeguards, arms control verification). Integrating technical constraints from federated and distributed learning—such as synchronization costs, communication overhead, and performance degradation over large distances—would help connect the governance model more closely to real-world system design.

6. Conclusion

This paper develops a game-theoretic model of verification in cross-border AI training projects, formalizing how monitoring intensity affects compliance incentives when actors can privately deviate from public compute pledges. Our simulations demonstrate that verification is an effective deterrent against free-riding. However, the relationship between verification and welfare is non-monotonic, with intermediate monitoring levels producing worse outcomes than either minimal or maximal verification—a "verification valley" that poses a commitment problem for institutional designers. Verification precision is a more influential policy lever than penalty magnitude for controlling compliance, suggesting that investments in monitoring technology may be more cost-effective than escalating sanctions. The model also highlights distributional consequences: verification regimes benefit smaller actors while imposing net costs on larger contributors, potentially requiring differentiated obligations or compensation mechanisms to sustain cooperation. For policymakers considering international AI governance institutions, the key implication is that half-measures in verification may be counterproductive.

Code and Data

Code repository: <https://github.com/pythonski/federated-learning-hackathon.git>

References

Besiroglu, T., Erdil, E., Barnett, M., & You, J. (2024). Chinchilla Scaling: A replication attempt. arXiv preprint arXiv:2404.10102.

Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., & Snidal, D. (2023). International Institutions for Advanced AI. arXiv preprint arXiv:2307.04699.

Moon, A., Vedula, P., Geneson, J., Bar-on, S. (2025). Strategies and Detection Gaps in a Game-Theoretic Model of Compute Governance. RAND Corporation Research Report RRA3686-1.

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeier, J. (2019). A Survey on Distributed Machine Learning. ACM Computing Surveys, 52(4).

O’Gara, A., Kulp, G., Hodgkins, W., Petrie, J., Immler, V., Aysu, A., Basu, K., Bhasin, S., Picek, S., & Srivastava, A. (2025). Hardware-Enabled Mechanisms for Verifying Responsible AI Development. Presented at the ICML 2025 Workshop on Technical AI Governance (TAIG).

LLM Usage Statement

ChatGPT was used to brainstorm approaches and help draft sections. Claude was used to write most of the code. The author alone is responsible for any errors in the text, which are regrettably likely given the time constraints.

Appendix

Dual-Use Risks. This framework could theoretically inform actors seeking to evade detection by identifying optimal cheating strategies that minimize flag probability while maximizing under-contribution. The sensitivity analysis reveals which parameters most affect detection, potentially guiding adversarial efforts to exploit weak verification regimes.

Ethical Considerations. The model treats compute contributions as the sole dimension of cooperation, abstracting over safety practices, data governance, and capability disclosure that are equally critical to responsible AI development. Verification regimes that focus narrowly on compute may inadvertently incentivize compliance gaming, delivering pledged compute while cutting corners on safety. The distributional findings (smaller actors benefit, larger actors bear costs) raise equity concerns: governance frameworks should consider differentiated obligations to avoid penalizing actors whose participation is essential for global coverage.