Risk-Weighted Compute Permit Markets under Imperfect Monitoring

Joel N. Christoph[a,*]

[a]*Department of Economics, European University Institute, Villa La Fonte, Via delle Fontanelle, 18, 50014 San Domenico di Fiesole, Italy*

Abstract

This paper studies the design of a tradable permit market for frontier AI compute under imperfect monitoring. A regulator sets an aggregate cap on risk-weighted compute; developers trade permits and must retire them in proportion to metered compute usage multiplied by evaluation-contingent risk multipliers. Because metering and reporting are noisy and strategically manipulable, the regulator conducts stochastic audits and imposes convex penalties for detected shortfalls. We characterize incentive-compatible audit and penalty pairs under enforcement budget constraints, establishing conditions under which truthful reporting constitutes a Bayesian Nash equilibrium. Under these conditions, competitive permit trading implements a second-best efficient allocation. We derive the optimal audit intensity and penalty schedule as functions of the permit price and detection technology, and compare equilibrium welfare with command-and-control alternatives, identifying parameter regimes where the permit market yields strictly higher surplus. An extension introduces insurance priced on verifiable model evaluations, showing that bundling insurance with permits improves compliance and allocative efficiency when auditing capacity is limited. The analysis connects cap-and-trade design, costly state verification, and the economics of AI governance.

Keywords: mechanism design, tradable permits, imperfect monitoring, costly state verification, cap-and-trade, AI governance, JEL: D47, D82, H23, L51, Q58

1. Introduction

Frontier AI development is concentrated in a small number of compute-intensive training runs. This concentration creates a natural regulatory lever: govern access to large-scale

*Corresponding author.
Email address: joel.christoph@eui.eu (Joel N. Christoph)

compute rather than attempting to regulate downstream model behaviors case by case. A simple cap on aggregate compute, however, is too blunt. Training runs differ sharply in expected hazard, and a one-size-fits-all limit either over-restricts low-risk innovation or under-regulates high-risk activity.

This paper analyzes a market-based governance architecture built around *risk-weighted compute permits*. A regulator issues tradable permits denominated in units of risk-weighted compute, sets an aggregate cap, and requires developers to surrender permits proportional to (i) metered compute usage and (ii) a risk multiplier determined by verifiable evaluation outcomes. The permit market allocates scarce compute efficiently across developers through price signals, while risk weighting concentrates regulatory attention on training runs that pose the greatest expected hazard.

The central difficulty is enforcement. Unlike emissions measured at a smokestack, the compliance-relevant quantity for AI training is harder to observe: hardware metering can be noisy, evaluation outcomes can be manipulated, and the riskiness of a training run depends on details that are only partly verifiable. Standard permit market designs that assume perfect compliance monitoring do not apply directly.

We address this by coupling the permit market to a compliance mechanism consisting of stochastic audits and convex penalties. The regulator audits a randomly selected fraction of training runs and imposes penalties that increase convexly in the detected shortfall between required and surrendered permits. This architecture draws on the costly state verification literature (Townsend, 1979; Gale and Hellwig, 1985), extending it to a market setting with heterogeneous risk types and endogenous evaluation choices. The mechanism design approach follows a tradition—exemplified by Mitra (2001)—of characterizing incentive-compatible and efficient allocation rules under incomplete information, adapted here to a regulatory setting with tradable permits.

Contributions.. We make four contributions.

First, we formalize compute-permit governance as a mechanism design problem with noisy monitoring and costly verification (Section 3). The model captures the key features of AI compute governance: heterogeneous risk across training runs, imperfect observability of compute usage and safety effort, and limited enforcement resources.

Second, we derive sufficient conditions on audit probabilities and penalty curvature under which truthful permit retirement is a dominant strategy, and characterize audit and penalty pairs that sustain compliance as a Bayesian Nash equilibrium (Section 4). The conditions make explicit the tradeoff between monitoring intensity and penalty severity: when audit capacity is constrained, steeper penalties compensate.

Third, we show that under truthful compliance, the competitive permit market implements the cap-constrained efficient allocation, and we compare welfare against command-and-control regulation, identifying parameter regimes where permits dominate (Section 5).

Fourth, we extend the framework to incorporate insurance priced on verifiable evaluation outcomes. Bundling insurance with permits creates an additional compliance incentive that reduces the required audit intensity, expanding the set of enforcement budgets under which the permit market is viable (Section 7).

2. Related Work

Our analysis draws on three bodies of work.

Cap-and-trade and permit markets.. The theoretical foundations of tradable permit systems were established by Montgomery (1972), who showed that a competitive market in emission licenses achieves the cost-minimizing allocation among polluters. Subsequent work examined permit auction design (Cramton and Kerr, 2002), the role of transaction costs (Stavins, 1995), and the classic "prices versus quantities" tradeoff under uncertainty (Weitzman, 1974). Our contribution is to adapt the permit framework to a setting where the regulated quantity (risk-weighted compute) is imperfectly observable and the risk weight itself depends on verifiable but manipulable evaluation signals.

Costly state verification and auditing.. Townsend (1979) introduced costly state verification in optimal contract design. Gale and Hellwig (1985) applied it to debt contracts, and Mookherjee and Png (1989) characterized optimal auditing schemes when the auditor faces budget constraints. Border and Sobel (1987) analyzed auditing and penalties in a model with strategic agents. We extend this literature to a market environment where multiple agents

interact through permit trading, and where the object to be verified (risk-weighted compute) combines observable metering with an evaluation-contingent risk classification.

AI governance and compute regulation.. A growing literature examines compute as a governance lever for AI development. Sastry et al. (2024) provide a comprehensive analysis of how computing power can serve as a regulatory instrument, arguing that compute is detectable, excludable, and quantifiable relative to other AI inputs. Shavit (2023) proposes a hardware-based monitoring framework for verifying compliance with rules on large-scale training runs. Anderljung et al. (2023) outline a regulatory architecture for frontier AI models based on registration, reporting, and compliance mechanisms. These contributions focus on institutional design and monitoring technology; formal economic analysis of market-based compute governance mechanisms, however, remains limited. We provide a mechanism design treatment that connects the AI governance discussion to established results on permit markets and enforcement under imperfect monitoring.

3. Model

3.1. Developers, compute, and risk

There is a set of AI developers $N = \{1, \dots, n\}$. Each developer i undertakes a training run using compute $c_i \in \mathbb{R}_+$ and choosing unobservable safety effort $e_i \in [0, 1]$. The private benefit from the run is $v_i(c_i)$, where v_i is twice continuously differentiable, strictly increasing, and strictly concave, with $v_i'(0) = \infty$ and $\lim_{c\to\infty} v_i'(c) = 0$.

After the training run, a standardized evaluation protocol produces a verifiable signal $s_i \in \{L, H\}$ about the model's risk profile. The probability of a high-risk signal depends on the developer's safety effort and on an exogenous type $\theta_i \in \Theta$ capturing application-specific hazard:

$$\Pr(s_i = H \mid \theta_i, e_i) = g(\theta_i, e_i),$$

where g is increasing in θ_i and decreasing in e_i. The regulator maps the evaluation signal to a risk multiplier:

$$\lambda(s_i) = \begin{cases} 1 & \text{if } s_i = L, \\ \alpha & \text{if } s_i = H, \end{cases}$$

with $\alpha > 1$. The *risk-weighted compute* of developer i is

$$q_i \equiv \lambda(s_i) \cdot c_i.$$

3.2. Permit market

The regulator issues $Q > 0$ permits, each entitling the holder to one unit of risk-weighted compute. Permits are freely tradable in a competitive secondary market at price $p \geq 0$. Developer i must *retire* (surrender) $\hat{q}_i$ permits for their training run. The compliance requirement is that $\hat{q}_i \geq q_i = \lambda(s_i)c_i$.

3.3. Imperfect monitoring

The regulator does not observe c_i directly. Hardware-level metering produces a signal $m_i = c_i + \varepsilon_i$, where ε_i is mean-zero noise. The regulator can audit developer i at cost $k > 0$, which reveals c_i exactly. Audits are stochastic: each run is audited independently with probability $\rho \in (0, 1]$, subject to a budget constraint $\rho \leq \bar{\rho}$, where $\bar{\rho} = B/(nk)$ for total enforcement budget B.

If an audit detects a shortfall $\Delta_i = (\lambda(s_i)c_i - \hat{q}_i)_+$, the regulator imposes a penalty $\phi(\Delta_i)$.

Assumption 1 (Penalty regularity)**.** The penalty function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is continuously differentiable, increasing, strictly convex, with $\phi(0) = 0$, $\phi'(0) \geq 0$, and $\lim_{\Delta\to\infty} \phi(\Delta)/\Delta = \infty$.

3.4. Developer's problem

Given permit price p, audit probability ρ, and penalty function ϕ, developer i with realized risk multiplier $\lambda_i = \lambda(s_i)$ chooses compute c_i and permit retirement $\hat{q}_i$ to maximize:

$$U_i(c_i, \hat{q}_i) = v_i(c_i) - p\hat{q}_i - \rho\,\phi\Big((\lambda_i c_i - \hat{q}_i)_+\Big). \tag{1}$$

The first term is the benefit of compute. The second is the cost of acquiring and retiring permits. The third is the expected penalty from audit detection. We analyze the compliance decision (choice of $\hat{q}_i$ given c_i) separately from the compute choice (choice of c_i given compliance behavior).

4. Compliance under Imperfect Monitoring

This section establishes conditions under which truthful permit retirement, that is, surrendering $\hat{q}_i = \lambda_i c_i$ permits, is optimal for every developer.

Proposition 2 (Deterrence condition)**.** *Fix any compute level $c_i > 0$ and any realized risk multiplier λ_i. Truthful retirement $\hat{q}_i = \lambda_i c_i$ is a best response if*

$$\rho\,\phi(\Delta) \geq p\,\Delta \quad \textit{for all } \Delta \geq 0. \tag{2}$$

Proof. Consider a deviation in which developer i under-retires by $\Delta \geq 0$, setting $\hat{q}_i = \lambda_i c_i - \Delta$ instead of $\hat{q}_i = \lambda_i c_i$. This reduces permit expenditure by $p\Delta$. With probability ρ, an audit detects the shortfall Δ and imposes penalty $\phi(\Delta)$. The net change in expected payoff from the deviation is

$$\Delta U = p\Delta - \rho\,\phi(\Delta).$$

If (2) holds, then $\Delta U \leq 0$ for all $\Delta \geq 0$. Since over-retirement ($\Delta < 0$) is strictly costly (it wastes permits at price p per unit without any offsetting benefit), the developer's optimal retirement is exactly $\hat{q}_i = \lambda_i c_i$. □

The condition (2) is intuitive: the expected penalty from under-reporting must exceed the permit cost saved. Because ϕ is convex, the condition is tightest for small deviations. The following corollary makes this explicit.

Corollary 3 (Marginal deterrence)**.** *If ϕ is differentiable at 0 and $\rho\,\phi'(0) \geq p$, then condition (2) holds for all $\Delta \geq 0$.*

Proof. By convexity, $\phi(\Delta) \geq \phi(0) + \phi'(0)\Delta = \phi'(0)\Delta$ for all $\Delta \geq 0$. Thus $\rho\,\phi(\Delta) \geq \rho\,\phi'(0)\Delta \geq p\Delta$. □

Remark 4 (Audit and penalty tradeoff). Corollary 3 makes explicit the substitutability between audit intensity and penalty severity. A regulator with limited audit capacity (small ρ) can maintain deterrence by raising the marginal penalty $\phi'(0)$, while a regulator with cheap auditing (large ρ) can use milder penalties. The product $\rho\,\phi'(0) \geq p$ is the binding constraint.

Parametric example: quadratic penalties.. Suppose $\phi(\Delta) = a\Delta + b\Delta^2$ with $a, b > 0$. Then $\phi'(0) = a$, and Corollary 3 requires $\rho\, a \geq p$. The minimum audit probability for deterrence is

$$\rho^* = \frac{p}{a}. \tag{3}$$

For $\rho \geq \rho^*$, compliance is sustained regardless of the quadratic coefficient b (which provides additional deterrence against large deviations). This yields a simple design rule: set the linear penalty coefficient a to control the marginal deterrence threshold, and set the quadratic coefficient b to provide robustness against large deviations.

Budget-constrained audit design.. The regulator chooses ρ to minimize enforcement cost $nk\rho$ subject to the deterrence constraint $\rho a \geq p$, yielding optimal audit probability $\rho^* = p/a$ (provided $p/a \leq \bar{\rho}$). The total enforcement cost is nkp/a. This is decreasing in a: steeper marginal penalties reduce the required audit frequency and total enforcement expenditure. However, very large penalties may face legal or political constraints, bounding a from above in practice.

5. Permit Market Equilibrium and Efficiency

Assume that the deterrence condition of Proposition 2 holds, so that all developers comply truthfully. We now analyze the compute allocation induced by the permit market.

5.1. Competitive equilibrium

Under truthful compliance, developer i with expected risk multiplier $\bar{\lambda}_i \equiv \mathbb{E}[\lambda(s_i) \mid \theta_i, e_i]$ chooses compute to solve

$$\max_{c_i \geq 0} \; v_i(c_i) - p\, \bar{\lambda}_i\, c_i. \tag{4}$$

The first-order condition is

$$v_i'(c_i^*) = p\, \bar{\lambda}_i. \tag{5}$$

The permit price p^* clears the market:

$$\sum_{i=1}^{n} \bar{\lambda}_i\, c_i^*(p^*) = Q. \tag{6}$$

Proposition 5 (Efficient allocation under risk-weighted permits)**.** *Suppose truthful compliance holds and developers are price-takers. Then any market-clearing permit price p^* implements a solution to the cap-constrained planner's problem:*

$$\max_{(c_i \geq 0)_{i \in N}} \sum_{i=1}^{n} v_i(c_i) \quad \textit{subject to} \quad \sum_{i=1}^{n} \bar{\lambda}_i \, c_i \leq Q.$$

Proof. The Lagrangian of the planner's problem is

$$\mathcal{L} = \sum_{i=1}^{n} v_i(c_i) - \mu\Big(\sum_{i=1}^{n} \bar{\lambda}_i c_i - Q\Big).$$

The first-order conditions for interior solutions are $v_i'(c_i) = \mu\bar{\lambda}_i$ for all i. In the competitive permit market, each developer's first-order condition is (5): $v_i'(c_i^*) = p^*\bar{\lambda}_i$. Thus the market-clearing price p^* coincides with the planner's shadow value μ, and the competitive allocation $(c_i^*(p^*))_{i \in N}$ solves the planner's problem. □

The result formalizes the role of tradability: once compliance is enforced, the market uses permit prices to allocate compute efficiently under the cap. Risk weighting ensures that high-hazard training runs face a higher effective marginal cost of compute (scaled by $\bar{\lambda}_i$), concentrating regulatory pressure where it is most needed.

5.2. Comparison with command-and-control

Under a command-and-control regime, the regulator directly assigns compute quotas $\bar{c}_i$ to each developer. If the regulator has full information about developer values $v_i(\cdot)$ and risk types θ_i, it can replicate the planner's optimum. In practice, the regulator typically does not observe private values.

Proposition 6 (Welfare comparison)**.** *Let W^{permit} denote welfare under the risk-weighted permit market with truthful compliance, and let W^{CC} denote welfare under a command-and-control allocation $(\bar{c}_i)$ that respects the same aggregate cap Q.*

(i) If the regulator knows all v_i and $\bar{\lambda}_i$, then $W^{\text{CC}} = W^{\text{permit}}$.

(ii) If the regulator does not observe v_i and assigns uniform quotas $\bar{c}_i = Q/(n\bar{\lambda}_{\text{avg}})$, then $W^{\text{permit}} \geq W^{\text{CC}}$, with strict inequality whenever developers are heterogeneous in marginal values at the uniform allocation.

Proof. Part (i) follows because a fully informed regulator can replicate the solution to the planner's problem.

For part (ii), note that the permit market allocation satisfies $v_i'(c_i^*) = p^*\bar{\lambda}_i$ for all i, which is the planner's optimality condition. The uniform command-and-control allocation generally violates this condition: if $v_i'(\bar{c}_i) \neq v_j'(\bar{c}_j) \cdot (\bar{\lambda}_j/\bar{\lambda}_i)$ for some pair i, j, then reallocating compute from the low-marginal-value developer to the high-marginal-value developer (at constant risk-weighted total) increases aggregate welfare. The permit market performs exactly this reallocation through voluntary trade. □

Remark 7 (When command-and-control can dominate). The comparison above conditions on truthful compliance in the permit market, which requires enforcement expenditure nkp/a. If the enforcement budget is insufficient to sustain compliance (i.e., $B < nkp/a$), the permit market may unravel as evasion becomes attractive. In such regimes, a command-and-control system with simpler enforcement requirements (e.g., hardware shutoffs) could outperform the permit market despite its allocative inefficiency. The relevant comparison is therefore between the allocative gains from market-based reallocation and the additional enforcement costs of sustaining compliance.

6. Optimal Audit Policy under Budget Constraints

We now characterize the regulator's optimal enforcement policy when the enforcement budget is binding.

Proposition 8 (Cost-minimizing audit intensity)**.** *Suppose penalties take the form* $\phi(\Delta) = a\Delta + b\Delta^2$ *with* $a, b > 0$*. Given permit price* p*:*

(i) Deterrence for all shortfall levels $\Delta \geq 0$ *holds if and only if* $\rho\, a \geq p$*.*

(ii) The cost-minimizing audit probability that preserves deterrence is $\rho^* = p/a$*, provided* $p/a \leq \bar{\rho}$*.*

(iii) Total enforcement cost under ρ^* *is* $C^* = nkp/a$*, which is decreasing in the linear penalty coefficient* a*.*

Proof. For part (i), the deterrence condition (2) requires $\rho(a\Delta + b\Delta^2) \geq p\Delta$ for all $\Delta \geq 0$. Dividing by $\Delta > 0$ yields $\rho(a + b\Delta) \geq p$ for all $\Delta > 0$. Since this must hold as $\Delta \downarrow 0$, the

binding constraint is $\rho\, a \geq p$. Conversely, if $\rho\, a \geq p$, then $\rho(a + b\Delta) \geq \rho\, a \geq p$ for all $\Delta \geq 0$, confirming deterrence.

Part (ii) follows: among audit probabilities satisfying $\rho\, a \geq p$, the smallest is $\rho^* = p/a$.

Part (iii): total enforcement cost is $nk\rho^* = nkp/a$, which is decreasing in a. □

This result provides a simple calibration rule. The regulator observes the permit price p (which is endogenous but can be estimated from market data), chooses the linear penalty coefficient a subject to legal or political constraints, and sets the audit probability accordingly. When permits become expensive (high p), evasion incentives rise, requiring either more frequent audits or steeper penalties.

Comparative statics.. Several relationships follow from the analysis. A tighter cap Q raises the permit price p, which increases required audit intensity or penalty severity for compliance. Better metering technology (lower ε variance) improves audit signal quality, which can be modeled as reducing the effective audit cost k; this lowers enforcement expenditure without changing the deterrence condition. A wider gap between risk multipliers (α larger) increases the stakes of misclassification, raising the return to auditing evaluation outcomes in addition to compute usage.

7. Extension: Insurance Bundled with Permits

When the enforcement budget B is insufficient to sustain compliance at the unconstrained optimal audit level, the regulator can supplement audit-based deterrence with an insurance requirement.

Suppose an insurer offers coverage for catastrophic incidents at premium $\pi(\lambda_i)$ per unit of compute, where π is increasing: $\pi(\alpha) > \pi(1)$. The developer's objective becomes

$$U_i^{\text{ins}} = v_i(c_i) - p\hat{q}_i - \pi(\lambda_i)c_i - \rho\,\phi\Big((\lambda_i c_i - \hat{q}_i)_+\Big). \tag{7}$$

The insurance premium does not directly change the permit retirement decision (which depends on p, ρ, and ϕ). However, it affects the *evaluation manipulation* incentive. A developer who can influence the evaluation signal s_i (e.g., by misrepresenting training specifications or selectively running benchmarks) faces a tradeoff: lowering λ_i reduces permit costs but

does not reduce the insurance premium if the insurer conducts independent verification. Conversely, if the insurer conditions premiums on the same verifiable signal, the cost of gaming the evaluation rises.

Proposition 9 (Insurance-enhanced deterrence)**.** *Suppose the insurer observes s_i independently and charges $\pi(\lambda_i)$ per unit compute, with $\pi(\alpha) - \pi(1) = \delta > 0$. If a developer can manipulate the evaluation to produce $s_i = L$ instead of $s_i = H$ at private cost $\gamma > 0$, then adding insurance weakly reduces the set of developers who find manipulation profitable. Specifically, manipulation is profitable only if $\gamma < p(\alpha - 1)c_i$ without insurance, but requires $\gamma < [p(\alpha - 1) - \delta]c_i$ with insurance when $\delta < p(\alpha - 1)$.*

Proof. Without insurance, successfully manipulating s_i from H to L reduces required permit retirement by $(\alpha - 1)c_i$, saving $p(\alpha - 1)c_i$ in permit costs. Manipulation is profitable if $\gamma < p(\alpha - 1)c_i$.

With insurance, the manipulation also affects the premium: the developer pays $\pi(1)c_i$ instead of $\pi(\alpha)c_i$, saving δc_i. But since the insurer conducts independent verification, the manipulated signal does not change the insurer's assessment. The insurer still charges $\pi(\alpha)c_i$. Thus the net benefit from manipulation is only the permit savings: $p(\alpha - 1)c_i - \delta c_i$ (because the insurer's independent assessment raises the developer's cost by δc_i relative to the case where the insurer accepted the manipulated signal). Manipulation is profitable only if $\gamma < [p(\alpha - 1) - \delta]c_i$.

When $\delta > 0$, the threshold for profitable manipulation is strictly lower with insurance than without. □

The insurance module thus complements the audit-based compliance mechanism by raising the cost of evaluation manipulation. This is valuable precisely when audit resources are scarce: the insurer's independent verification substitutes for regulatory auditing of evaluation signals, allowing the regulator to focus audit capacity on verifying compute usage.

8. Discussion

Implementation requirements.. The proposed mechanism requires three institutional components: (i) a permit registry and trading platform, which can build on existing financial market

infrastructure; (ii) a standardized evaluation protocol for mapping training outcomes to risk multipliers, which requires technical capacity but not real-time surveillance; and (iii) an audit and penalty system, which requires legal authority and enforcement resources but operates ex post rather than continuously. The hardware-level monitoring framework proposed by Shavit (2023) offers a concrete technical foundation for the metering and audit functions assumed in our model.

International enforcement.. In multi-jurisdiction settings where no single regulator has global authority, the mechanism can operate through market access and supply-chain leverage. Cloud compute providers and chip manufacturers can serve as chokepoints: access to state-of-the-art hardware or cloud services can be conditioned on participation in the permit system. This approach parallels international financial regulation, where market access requirements substitute for direct enforcement authority. As Sastry et al. (2024) observe, the concentrated structure of the AI chip supply chain makes compute a particularly effective point of regulatory intervention.

Dynamic extensions.. The static model analyzed here can be extended to accommodate permit banking across periods (allowing developers to smooth compute usage over time), dynamic adjustment of the cap based on realized incident rates and technological change, and differentiated permit categories for distinct model domains (e.g., language models versus robotics versus biological design tools). These extensions introduce intertemporal incentive considerations that merit separate analysis.

Limitations.. Several assumptions warrant discussion. The model treats audit outcomes as deterministic conditional on the audit occurring; in practice, audits may produce noisy signals, requiring modifications to the penalty structure (see Appendix A). We assume a single risk classification (L or H); a finer classification would improve targeting but complicate the mechanism. The competitive equilibrium analysis assumes price-taking behavior, which may not hold if a small number of large developers have market power in the permit market. Finally, the welfare analysis conditions on a given cap Q; optimal cap-setting is a separate (and important) design problem that involves the regulator's assessment of aggregate risk tolerance.

9. Conclusion

Risk-weighted compute permits offer a tractable approach to AI governance that combines a quantitative cap with market-based flexibility. By denominating permits in risk-weighted units, the mechanism targets regulatory attention on high-hazard training runs without categorically restricting any particular type of development. The key challenge is enforcement under imperfect monitoring; coupling the permit market to a calibrated audit and penalty scheme yields a robust second-best allocation. Integration with evaluation-contingent insurance further strengthens compliance incentives and reduces the enforcement budget required to sustain the mechanism. The framework provides a foundation for compute governance that balances innovation and safety, and that can accommodate international participation through market access conditions rather than direct jurisdictional authority.

Acknowledgements

I thank seminar participants at the Centre for the Governance of AI and Harvard Kennedy School for helpful discussions. This paper is submitted to the Special Issue of *Mathematical Social Sciences* in memory of Manipushpak Mitra, whose contributions to mechanism design theory—particularly on incentive-compatible allocation under incomplete information (Mitra, 2001)—inform the approach taken here.

Declaration of Competing Interests

The author has no competing interests to declare.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The author acknowledges doctoral funding from the European University Institute (France scholarship) and the Japan-IMF Scholarship Program for Advanced Studies.

Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work the author used generative AI tools in order to assist with editing and formatting of non-technical exposition and LaTeX code. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Appendix A. Robustness: Noisy Audits

The baseline model assumes that an audit perfectly reveals c_i. If instead the audit produces a noisy signal $\tilde{c}_i = c_i + \eta_i$ with η_i mean-zero and known distribution, the deterrence condition must account for the probability that the audit fails to detect a shortfall. Let $d(\Delta) = \Pr(\tilde{c}_i \geq c_i - \Delta)$ denote the probability that the noisy audit correctly identifies an under-retirement of magnitude Δ. The modified deterrence condition becomes

$$\rho\, d(\Delta)\, \phi(\Delta) \geq p\Delta \quad \text{for all } \Delta \geq 0.$$

If $d(\Delta) \to 1$ as the audit technology improves (i.e., $\mathrm{Var}(\eta_i) \to 0$), this converges to the baseline condition (2). For fixed d, noisier audits require either higher ρ or steeper ϕ to maintain deterrence.

Appendix B. Proof Details for Proposition 6

Part (ii) of Proposition 6 follows from the general principle that a competitive equilibrium maximizes the sum of concave objectives subject to a linear constraint. The uniform command-and-control allocation generically fails the equi-marginal condition $v_i'(\bar{c}_i)/\bar{\lambda}_i = v_j'(\bar{c}_j)/\bar{\lambda}_j$ for all i, j, because both v_i' and $\bar{\lambda}_i$ vary across developers. Therefore, there exist feasible reallocations that increase total welfare, and the competitive equilibrium achieves the maximum.

References

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., et al. (2023). Frontier AI regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

Border, K. C. and Sobel, J. (1987). Samurai accountant: A theory of auditing and plunder. *Review of Economic Studies*, 54(4):525–540.

Cramton, P. and Kerr, S. (2002). Tradeable carbon permit auctions: How and why to auction not grandfather. *Energy Policy*, 30(4):333–345.

Gale, D. and Hellwig, M. (1985). Incentive-compatible debt contracts: The one-period problem. *Review of Economic Studies*, 52(4):647–663.

Mitra, M. (2001). Mechanism design in queueing problems. *Economic Theory*, 17(2):277–305.

Montgomery, W. D. (1972). Markets in licenses and efficient pollution control programs. *Journal of Economic Theory*, 5(3):395–418.

Mookherjee, D. and Png, I. P. L. (1989). Optimal auditing, insurance, and redistribution. *Quarterly Journal of Economics*, 104(2):399–415.

Sastry, G., Heim, L., Belfield, H., Anderljung, M., Brundage, M., Hazell, J., et al. (2024). Computing power and the governance of artificial intelligence. *arXiv preprint arXiv:2402.08797*.

Shavit, Y. (2023). What does it take to catch a Chinchilla? Verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*.

Stavins, R. N. (1995). Transaction costs and tradeable permits. *Journal of Environmental Economics and Management*, 29(2):133–148.

Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2):265–293.

Weitzman, M. L. (1974). Prices vs. quantities. *Review of Economic Studies*, 41(4):477–491.