The Partnership Protocol: Mitigating Agency Costs in AI Governance

By Patrick Martinez Peel, Ph.D. patrick.peel@gmail.com | linkedin.com/in/patrick-martinez-peel| github.com/patrick-martinez-peel December 2025

Abstract

Governing autonomous agents presents a classic principal-agent problem at scale. As AI systems proliferate, the monitoring costs of centralized oversight become prohibitive - a scalability bottleneck explicitly recognized by leading labs like OpenAI and Anthropic. This paper proposes a mechanism design approach to mitigate total agency costs by shifting from a Limited Liability framework to a Joint Liability framework. Through game-theoretic modeling and a 1,000-agent evolutionary simulation of a cyber-defense swarm, I demonstrate that while external audits are necessary, a Joint Liability structure reduces the cost of enforcement. By internalizing the externalities of failure, this protocol incentivizes peer monitoring, thereby lowering the regulatory burden required to maintain a stable, secure equilibrium. In effect, the protocol allows safety guarantees to scale with agent populations without requiring a proportional increase in centralized regulation and headcount.

---

1. Introduction: The Monitoring Cost Crisis

The core challenge in AI governance is not technical, but economic: agency costs.

As we deploy thousands of autonomous agents, the cost for a principal (the operator or regulator) to verify the actions of every agent becomes insurmountable. Current "human-in-the-loop" models face diminishing returns; they simply cannot scale to machine speed.

Major AI labs are exploring technical solutions like multi-agent reinforcement learning (MARL) and Constitutional AI. Complementing these technical advances, this paper argues that without also correcting the underlying liability and incentive structures, monitoring will remain inefficient. We must design governance frameworks that minimize the transaction costs of agent oversight.

2. Theoretical Framework: Mechanism Design

This paper analyzes two distinct governance mechanisms for managing agent risk:

- **Limited Liability Mechanism:** The agent's downside risk is capped (e.g., limited to the liquidation of that single instance). This creates a structural moral hazard (Jensen & Meckling, 1976). The agent captures the full upside of risky behavior (speed/efficiency), while the legal shield allows them to externalize the tail risk of failure to the principal (Easterbrook & Fischel, 1985).
- **Joint Liability Mechanism:** Agents share a pooled downside risk. If one agent fails, the entire pool is penalized (analogous to "slashing" in blockchain Proof-of-Stake protocols). This effectively internalizes the externality, creating a strong incentive for peer monitoring (Stiglitz, 1990). By transferring the burden of oversight to those with the lowest information costs - the agents themselves - we mitigate the principal's scalability bottleneck. In effect, we trade agent independence for mutual risk.

3. The Efficiency Gap: Institutional Analysis

While both models can technically achieve safety, they impose vastly different monitoring costs on the principal. The Joint Liability model is superior for four economic reasons:

1. **Reduction of Policing Costs:** Because penalties in the Joint Liability model are existential, the principal can achieve compliance with a significantly lower probability of detection. The simulation confirms that increasing the severity of the penalty allows the regulator to reduce the frequency (and cost) of audits while maintaining the same safety equilibrium.
2. **Endogenous Monitoring:** Joint Liability aligns incentives for horizontal monitoring. Agents naturally develop protocols to verify each other's actions because they bear the cost of a peer's failure (Stiglitz, 1990). This mitigates information asymmetry by "deputizing" the agents.
3. **Robustness to Imperfect Information:** The internal cost of failure disincentivizes regulatory arbitrage, acting as a fail-safe even when the regulator has significant blind spots. This aligns with findings that liability is the optimal instrument when regulators face high information costs (Shavell, 1984).
4. **Signaling and Adverse Selection:** Economic theory suggests this structure creates a separating equilibrium. The willingness to accept uncapped liability acts as a costly signal of quality (Spence, 1973), naturally filtering out low-quality agents who cannot afford the risk of ruin (Akerlof, 1970).

> *Note: While this pre-screening mechanism operates at the entry stage and is therefore not explicitly modeled in the operational simulation in Section 4, it theoretically provides a second layer of safety by preventing bad actors from entering the pool in the first place.*

4. Empirical Evidence: The "Lazy Patcher" Simulation

To quantify this Efficiency Gap, I modeled a specific cybersecurity scenario: a cyber-defense swarm tasked with patching network vulnerabilities.

- **The Scenario:** Agents are rewarded for speed (closing tickets). A Risky strategy represents a Lazy Patcher - an agent that skips verification steps to save compute resources, increasing the probability of a critical security breach.
- **Methodology:** 1,000 autonomous agents executing patching tasks under varying audit probabilities, updating strategies based on social learning. The simulation parameterizes the legal definition of Limited Liability as a capped penalty (Fine + Low Reputation Loss), whereas Joint Liability is parameterized as an existential penalty (Fine + Total Reputation Loss). These inputs act as structural proxies for the differing bankruptcy protections in each legal framework.

Key Findings: The Efficiency Gap

> In this parameterization, the Limited Liability model required a relatively high audit probability - on the order of 60% - to reliably suppress the Lazy Patcher strategy and achieve a stable ~95% safe equilibrium. By contrast, the Joint Liability (Partnership) model achieved a comparable level of security with an audit probability of only ~20%.
>
> In other words, by substantially increasing the internalized cost of failure, Joint Liability reduced the level of external monitoring required by roughly a factor of three, demonstrating a clear and robust Efficiency Gap between the two governance regimes.

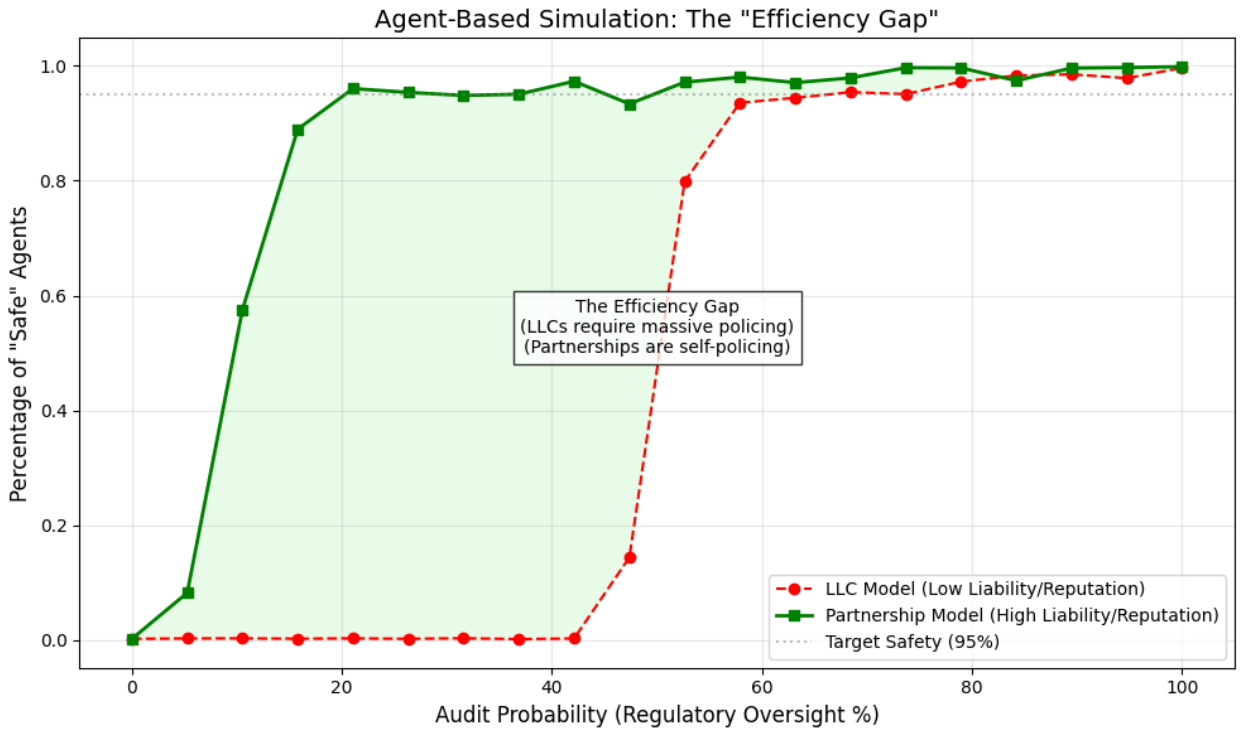

Figure 1. Audit Probability vs Safe Behavior Adoption under LLC and Joint Liability Models. *Results from the 1,000-agent evolutionary simulation showing the percentage of Safe agents achieved at different audit intensities. The Joint Liability (Partnership) regime achieves stable safety outcomes at substantially lower monitoring probabilities compared to the Limited Liability (LLC) regime, illustrating the Efficiency Gap discussed in the text.*

> *Note on Parameter Sensitivity: The simulation inputs assume the punitive cost of Joint Liability (Partnership) is approximately 5x greater than that of Limited Liability (LLC). This is a conservative estimate. In comparable cryptoeconomic systems like blockchain Proof-of-Stake, the ratio between the value of a single transaction (the risk in an LLC model) and the total bonded stake subject to slashing (the risk in a Partnership model) can exceed 100x or 1,000x. Thus, while this simulation shows a 3x efficiency gain, real-world high-stakes slashing protocols may yield even higher leverage.*

5. Discussion: Robustness under Correlated Failure

The efficiency gains observed in Section 4 rely on the assumption that agent failures are independent. However, real-world deployment introduces the risk of contagion. Theoretical literature warns that joint liability regimes may amplify systemic risk if this independence assumption is violated (Schelling, 1960). If agents share blind spots (such as those inherent in

software monocultures), coordinate behavior, or collude to conceal failures, joint penalties could propagate correlated losses across the pool.

This concern is valid but does not negate the mechanism. Joint liability performs best when failures are idiosyncratic rather than perfectly correlated (Besley & Coate, 1995). In practice, sectors such as microfinance and financial clearinghouses have successfully managed this risk not by abandoning the model, but by imposing complementary design constraints, including:

- **Heterogeneous Architectures:** Mandating diverse architectures or training data within a liability pool.
- **Randomized Audits:** Utilizing rotating or randomized audit schedules to deter collusion.
- **Sub-Pooling:** Implementing shard-level joint liability to compartmentalize risk while preserving peer-monitoring incentives.

These safeguards parallel design principles already used in financial clearinghouses and distributed systems. Correlated-risk considerations therefore constrain pool construction rather than undermine the core efficiency gains of joint liability.

6. Conclusion

This research demonstrates that liability and monitoring are functional substitutes in AI governance. As the internalized cost of failure increases, the level of external oversight required to maintain safe behavior falls. In the cyber-defense swarm modeled here, a Joint Liability (Partnership) framework achieved stable safety outcomes with roughly one-third the monitoring intensity required under a Limited Liability regime. Rather than relying exclusively on centralized audits, the protocol shifts the burden of oversight to the agents themselves, transforming safety from a purely regulatory activity into a shared operational imperative.

As noted in Section 5, this model is not a panacea. It requires specific architectural safeguards - such as diversity requirements and randomized audits - to prevent the specific risk of correlated failure found in monocultures. However, precedent from microfinance and financial clearinghouses confirms that these risks are manageable through institutional design.

As autonomous AI systems proliferate, scalable trust mechanisms will be essential. The Partnership Protocol suggests that carefully structured joint liability systems can serve as one such mechanism, lowering principal monitoring requirements while preserving or enhancing security outcomes. By trading agent independence for mutual risk, this approach offers a concrete path to circumvent the scalability bottleneck inherent in centralized oversight - a critical barrier to ensuring trust and security in the age of agentic AI.

7. Project Structure & Reproduction

The complete simulation code and reproducibility materials used to generate these findings are available in the open-source repository:

https://github.com/patrick-martinez-peel/partnership-protocol

- **`simulation.py`:** Core logic for the Agent, Principal, and Evolutionary Game loop.
- **efficiency_gap.png:** The generated visualization of the "Efficiency Gap."
- **requirements.txt:** List of Python dependencies required to run the environment.

To Reproduce Results: The simulation is deterministic. To reproduce the data and generate Figure 1:

1. Clone the repository and install dependencies (pip install -r requirements.txt).
2. Run the simulation script:

   Bash

   ```
   python simulation.py
   ```

References

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500.

Besley, T., & Coate, S. (1995). Group lending, repayment incentives and social collateral. *Journal of Development Economics*, 46(1), 1-18.

Easterbrook, F. H., & Fischel, D. R. (1985). Limited Liability and the Corporation. *The University of Chicago Law Review*, 52(1), 89-117.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure. *Journal of Financial Economics*, 3(4), 305-360.

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press.

Shavell, S. (1984). A Model of the Optimal Use of Liability and Safety Regulation. *The RAND Journal of Economics*, 15(2), 271-280.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, 87(3), 355-374.

Stiglitz, J. E. (1990). Peer Monitoring and Credit Markets. *The World Bank Economic Review*, 4(3), 351-366.