The Price of Over-Delegation: Stackelberg Liability Design for Agentic AI Handoffs

Tomoya Hoshino HOSHINO_TOMOYA@DN.SMBC.CO.JP
Sumitomo Mitsui Banking Corporation, Japan

Abstract

As LLM-based agents are increasingly deployed in sequential delegation chains, each handoff can obscure accountability for the final output, leading to context loss, audit overhead, and diffusion of responsibility. We formulate this governance problem as a Stackelberg game: a regulator sets a liability share, and developers choose a workflow partition via a boundary-insertion game on a sequential workflow DAG. The induced game is an exact potential game for every liability share $\gamma \in (0, 1]$, and under a continuous relaxation admits a unique interior equilibrium. We prove an over-fragmentation theorem: when developers only partially internalize handoff externalities ($\gamma < 1$), the equilibrium delegation depth strictly exceeds the social optimum, and the resulting welfare loss admits a scale-free closed-form expression independent of workflow size, agent productivity, and handoff-cost scale. We characterize the optimal liability share γ^* via a first-order condition that balances the marginal welfare gain against the marginal enforcement cost, and derive comparative statics. Under optimal regulation, residual welfare loss scales quadratically with enforcement cost, suggesting that reductions in enforcement costs yield more-than-proportional welfare gains.

1. Introduction

As LLM-based agents are increasingly deployed in sequential delegation chains, handoffs can obscure accountability for final outputs, causing context loss and audit overhead [4, 12, 14]. Unlike human organizations, where responsibility allocation is typically designed *ex ante*, agentic AI systems may extend delegation chains autonomously at runtime, raising the question of who should bear handoff externalities. Prior work studies Stackelberg-style AI governance [16], accountability limits in human–agent collectives [13], and specialization–coordination trade-offs in organizations and AI task chains [1, 5]. However, these works do not characterize equilibrium distortions under partial liability internalization or the associated governance design problem (Section 2).

We provide a tractable benchmark by formulating the developers' partition choice as a boundary-insertion game on a sequential DAG with homogeneous developers. The induced game is an exact potential game [9] for every liability share $\gamma \in (0, 1]$, and the regulator sets γ as a Stackelberg leader. Our contributions are:

1. We formulate handoffs in multi-agent AI workflows as a Stackelberg governance problem and show that the developers' partition choice induces an exact potential game.

2. We prove an *over-fragmentation theorem*: when $\gamma < 1$, equilibrium delegation depth strictly exceeds the social optimum, with a scale-free closed-form welfare loss independent of workflow size, agent productivity, and handoff-cost scale.

© T. Hoshino.

3. We characterize the optimal γ^* via a first-order condition and comparative statics, and show that residual welfare loss scales quadratically with enforcement cost.

2. Related Work

Game-theoretic AI governance. Stackelberg formulations of AI governance [16], simulation-based evaluations of multi-agent governance [3, 15], and impossibility results for accountability in human–agent collectives [13] advance the field but leave open the characterization of equilibrium distortions under partial liability internalization and optimal regulatory design for structured workflows. We address this gap for handoff externalities on workflow DAGs.

Organizational economics. The specialization–coordination trade-off is classical [1, 7], and Demirer et al. [5] extend this analysis to AI task chaining. These works study single-planner optimization; we analyze the strategic setting in which multiple developers choose partitions based on decentralized incentives, producing systematic over-fragmentation in Nash equilibrium.

Congestion games and price of anarchy. Our boundary-insertion game is a Rosenthal-type congestion game [10] that admits an exact potential [9]. Unlike smoothness-based price-of-anarchy bounds [11], which yield worst-case results over broad game classes, we exploit the structure of handoff externalities to derive an exact, scale-free welfare-loss expression that depends only on γ and α.

3. Model

We assume a sequential workflow (line DAG), homogeneous developers, and a continuous relaxation.[1]

3.1. Workflow and Agents

A workflow of m sequential subtasks forms a line DAG $H = (V, E)$ with vertices $V = \{1, \ldots, m\}$ and edges $E = \{(\ell, \ell + 1)\}_{\ell=1}^{m-1}$. It is partitioned among k agents into contiguous intervals $\{S_1, \ldots, S_k\}$, where k is endogenous and chosen by developers.[2] Under homogeneity, each agent's scope is $|S_i| = m/k$.[3] Agent i produces output $a \cdot |S_i|^{1-\alpha}$, where $a > 0$ is a productivity scale and $\alpha \in (0, 1)$ governs the specialization gain, yielding aggregate output $am^{1-\alpha}k^{\alpha}$. We model the effective coordination overhead as ωk ($\omega > 0$); a literal chain of k agents has $k - 1$ handoffs, but using $\omega(k - 1)$ leaves the first-order conditions unchanged.[4] Social welfare is

$$W(k) = am^{1-\alpha}k^{\alpha} - \omega k. \tag{1}$$

1. Homogeneity abstracts away differences in agent capability but preserves the fragmentation bias. It is not what yields the potential structure, which follows from aggregativeness; it sharpens the closed-form loss $\mathcal{L}$ (Appendix A.7). Relaxations are discussed in Section 6.
2. Developers are the economic actors who design the workflow partition, such as orchestrators and deployers.
3. We relax integrality and analyze $k \in \mathbb{R}_{>0}$; the qualitative results are preserved in the discrete case.
4. The ωk convention shifts $W(k)$ by a constant $-\omega$ relative to $\omega(k-1)$. This affects neither the optimality/equilibrium conditions nor the relative welfare loss $\mathcal{L}$ (Corollary 6), since the constant cancels in the normalization.

3.2. Handoff Externality and Boundary-Insertion Game

Each developer internalizes only a fraction $\gamma \in (0, 1]$ (the *liability share*) of the social cost of a handoff; the remainder $(1 - \gamma)$ is externalized. The parameter γ is set by the regulator through logging mandates, audit requirements, and similar instruments; $\gamma = 1$ implements the first-best partition absent enforcement costs.

We formulate the partition choice as a boundary-insertion game: a player at each edge $(\ell, \ell + 1) \in E$ chooses $s_\ell \in \{0, 1\}$ (insert a handoff or not). These boundary players are analytical devices, not additional economic actors: they recast the developers' aggregate partition choice as a set of local binary choices, one per boundary. The resulting number of agents is $k(\mathbf{s}) = 1 + \sum_\ell s_\ell$, and the payoff of player ℓ is

$$u_\ell(\mathbf{s}) = s_\ell \cdot \left[\alpha a m^{1-\alpha} k(\mathbf{s})^{\alpha-1} - \gamma\omega\right], \tag{2}$$

i.e., the marginal specialization benefit minus the internalized handoff cost (zero if $s_\ell = 0$).

Proposition 1 (Potential Structure) *For every $\gamma \in (0, 1]$, the following hold:*

1. *The discrete boundary-insertion game is a Rosenthal-type exact potential game [9, 10]. Because the strategy space is finite, a pure-strategy Nash equilibrium exists.*
2. *Under the continuous relaxation, the potential is $\Phi_\gamma(k) = am^{1-\alpha}(k^\alpha - 1) - \gamma\omega(k - 1)$. Since Φ_γ is strictly concave, any interior equilibrium is unique and coincides with the maximizer of Φ_γ.*

The proof is given in Appendix A.2.

The first-order condition $\Phi'_\gamma(k) = 0$ yields the equilibrium delegation depth

$$\hat{k}(\gamma) = \left(\frac{\alpha a m^{1-\alpha}}{\gamma\omega}\right)^{\frac{1}{1-\alpha}} = k^* \cdot \gamma^{-\frac{1}{1-\alpha}}, \tag{3}$$

where k^* is the social optimum (Theorem 4). When $\gamma < 1$, we have $\hat{k}(\gamma) > k^*$: developers underweight handoff costs and fragment the workflow excessively.

Remark 2 (Best-response convergence) *Because the boundary-insertion game is an exact potential game on a finite strategy space (Proposition 1), it has the finite improvement property [9]: every path of unilateral better responses is finite and ends at a pure-strategy Nash equilibrium, so better- and best-response dynamics converge in finitely many steps. Under the continuous relaxation, gradient ascent on the strictly concave potential Φ_γ, with a sufficiently small step size, converges to the unique interior equilibrium $\hat{k}(\gamma)$ in (3). Developers can thus reach the equilibrium through decentralized adjustment.*

Remark 3 (Beyond the line) *The construction in Proposition 1 depends on the workflow only through aggregativeness: each boundary player's marginal value depends on the profile s only through the agent count $k(s)$. On any workflow whose undirected graph is a tree, cutting an edge raises the number of components by one, independently of the other edges, so $k(s) = 1 + \sum_e s_e$. Proposition 1, Theorem 5, and Corollary 6 then extend verbatim from the line to every tree. The identity fails once the undirected graph contains a cycle: in the reconverging workflow $1 \to \{2, 3\} \to 4$, removing an edge can leave the graph connected, so aggregativeness no longer holds and a boundary's marginal value depends on which other boundaries are cut. Proposition 9 in Appendix A.6 states this and gives the counterexample.*

3.3. Regulator as Stackelberg Leader

The regulator commits to $\gamma \in [\gamma_0, 1]$ ex ante; developers then choose the equilibrium partition $\hat{k}(\gamma)$; the workflow executes and welfare is realized. The baseline $\gamma_0 \in (0, 1)$ reflects nonregulatory internalization (reputation, market discipline). Raising γ above γ_0 requires costly enforcement, so the regulator solves

$$\max_{\gamma \in [\gamma_0, 1]} R(\gamma) = W\big(\hat{k}(\gamma)\big) - C(\gamma - \gamma_0), \tag{4}$$

where $C\colon [0, 1 - \gamma_0] \to \mathbb{R}_{\geq 0}$ is the enforcement cost ($C(0) = 0$, $C' > 0$, $C'' \geq 0$). We focus on the linear case $C(\Delta) = c_0 \Delta$. Raising γ from γ_0 is equivalent to a per-handoff Pigouvian levy $t = (\gamma - \gamma_0)\omega$; when enforcement is costly, the full correction need not be optimal (Appendix B).

4. Main Results

4.1. Social Optimum

As a benchmark, we characterize the delegation depth that maximizes social welfare $W(k)$ (Eq. (1)).

Theorem 4 (Optimal Delegation Depth) *Under the continuous relaxation, the socially optimal delegation depth is* $k^* = \big(\alpha a m^{1-\alpha}/\omega\big)^{1/(1-\alpha)}$.

The optimum k^* balances the specialization gain α against the handoff cost ω. The proof is given in Appendix A.1.

4.2. Over-Fragmentation

Our first main result shows that partial liability internalization systematically produces over-fragmentation.

Theorem 5 (Over-Fragmentation) *For any liability share $\gamma \in (0, 1)$, the equilibrium delegation depth (Proposition 1(ii)) strictly exceeds the social optimum:*

$$\hat{k}(\gamma) = k^* \cdot \gamma^{-\frac{1}{1-\alpha}} > k^*. \tag{5}$$

Over the extended domain $\gamma \in (0, 1]$, equality holds if and only if $\gamma = 1$ (full internalization).

Because developers internalize only a fraction γ of handoff costs, they fragment the workflow excessively. This formalizes one mechanism by which partial liability internalization leads to responsibility diffusion in agentic AI. The proof is given in Appendix A.2.

Corollary 6 (Equilibrium Welfare Loss) *Let $r \equiv \gamma^{-1/(1-\alpha)} > 1$. The relative welfare loss, normalized by first-best welfare, is*

$$\mathcal{L}(\gamma) \equiv \frac{W(k^*) - W(\hat{k}(\gamma))}{W(k^*)} = \frac{(1-\alpha) + \alpha\, r - r^{\alpha}}{1-\alpha}. \tag{6}$$

$\mathcal{L}$ depends only on γ and α and is independent of m, a, and ω (scale-free).[5]

5. When γ is sufficiently small, $W(\hat{k}(\gamma))$ can become negative, in which case $\mathcal{L}$ exceeds one.

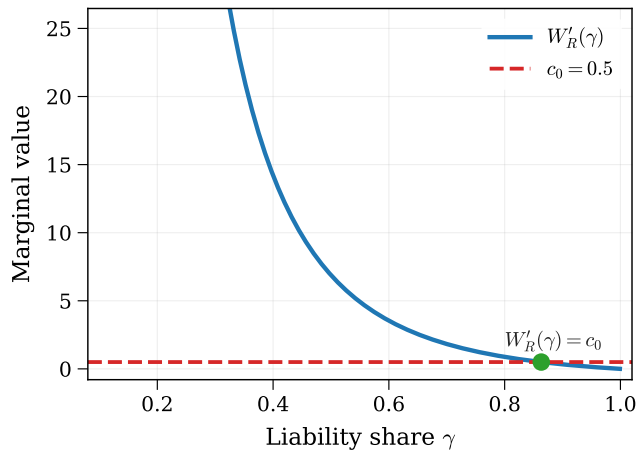

Figure 1: First-order condition for optimal liability regulation. The intersection of $W_R'(\gamma)$ (solid) and c_0 (dashed) determines $\gamma^* = 0.863$ ($\mathcal{L} \approx 0.007$). Parameters: $m = 10$, $\alpha = 0.3$, $a = \omega = 1.0$, $\gamma_0 = 0.1$, $c_0 = 0.5$.

The scale-free property means that over-fragmentation severity is determined entirely by the liability share and the specialization gain, making $\mathcal{L}$ a design diagnostic comparable across workflows of different scales. For instance, with $\alpha = 0.3$ and $\gamma = 0.4$, $\mathcal{L} \approx 0.47$: roughly half of first-best welfare is lost.[6]

4.3. Optimal Liability Regulation

The equilibrium welfare $W_R(\gamma) \equiv W(\hat{k}(\gamma))$ is strictly increasing and strictly concave on $(0, 1)$ (Appendix A.3), so $R(\gamma) = W_R(\gamma) - C(\gamma - \gamma_0)$ has a unique maximizer.

Theorem 7 (Optimal Liability Regulation) *Let $C\colon [0, 1 - \gamma_0] \to \mathbb{R}_{\geq 0}$ be an enforcement cost function with $C(0) = 0$, $C' > 0$, and $C'' \geq 0$. The following hold for the regulator's problem (4):*

(a) *The optimal liability share γ^* exists and is unique, possibly at the boundary of $[\gamma_0, 1]$.*

(b) *If $W_R'(\gamma_0) > C'(0)$, then $\gamma^* \in (\gamma_0, 1)$ and is characterized by*

$$\frac{\omega k^*}{1 - \alpha} (1 - \gamma^*) (\gamma^*)^{-\eta} = C'(\gamma^* - \gamma_0), \tag{7}$$

where $\eta \equiv \frac{2-\alpha}{1-\alpha} > 2$.

(c) *For an interior solution under the linear cost $C(\Delta) = c_0 \Delta$, the comparative statics are:*

$$\frac{\partial \gamma^*}{\partial c_0} < 0, \qquad \frac{\partial \gamma^*}{\partial m} > 0, \qquad \frac{\partial \gamma^*}{\partial a} > 0, \qquad \frac{\partial \gamma^*}{\partial \omega} < 0. \tag{8}$$

Part (b) characterizes optimal regulation as the point where the marginal welfare gain $W_R'(\gamma)$ equals the marginal enforcement cost. Part (c) implies that longer workflows and higher-productivity

6. Since $\mathcal{L} = 1 - 1/\text{PoA}$, Corollary 6 also provides an exact closed-form PoA, unlike worst-case smoothness-based bounds [11]. The PoA ratio requires $W(\hat{k}) > 0$, i.e., $\gamma > \alpha$; $\mathcal{L}$ remains well-defined on $(0, 1]$.

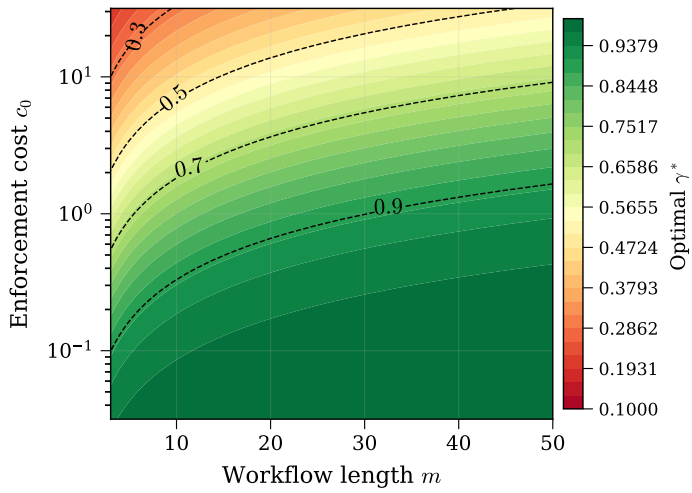

Figure 2: Regulator's design map: γ^* as a function of m and c_0. Contours mark $\gamma^* = 0.3, 0.5, 0.7, 0.9$. Parameters: $\alpha = 0.3$, $a = \omega = 1.0$, $\gamma_0 = 0.1$.

tasks are associated with higher optimal liability shares. Here ω is the per-handoff coordination overhead, not the severity of a failure; if a is read as scaling the social value of output, then higher-stakes workflows correspond to larger a, for which $\partial\gamma^*/\partial a > 0$ implies a higher optimal liability share. The remaining sign $\partial\gamma^*/\partial\omega < 0$ is therefore less paradoxical than it appears and is discussed in Appendix C.4. The proof is given in Appendix A.4.

Corollary 8 (Low-Cost Regime) *When $c_0 \ll \omega k^*/(1-\alpha)$, the optimal liability share satisfies $\gamma^* \approx 1 - c_0(1-\alpha)/(\omega k^*)$, and the residual welfare loss is*

$$\Delta W \equiv W(k^*) - W_R(\gamma^*) \approx \frac{c_0^2\,(1-\alpha)}{2\,\omega\,k^*}. \tag{9}$$

The quadratic scaling $\Delta W \propto c_0^2$ implies that reductions in enforcement costs yield more-than-proportional welfare gains (Appendix A.5).

5. Numerical Illustration

We illustrate the regulatory-design results using baseline parameters $m = 10$, $\alpha = 0.3$, $a = \omega = 1.0$, and $\gamma_0 = 0.1$.

Figure 1 plots both sides of the first-order condition (7) under linear enforcement cost ($c_0 = 0.5$). The intersection determines $\gamma^* = 0.863$, with relative welfare loss $\mathcal{L} \approx 0.007$, illustrating that optimal regulation achieves a near-first-best outcome while neither full internalization ($\gamma = 1$) nor the baseline ($\gamma = \gamma_0$) is optimal.

Figure 2 maps γ^* over workflow length m and enforcement cost c_0. For long workflows and low c_0, γ^* approaches 1; for short workflows and high c_0, it approaches γ_0, consistent with the comparative statics of Theorem 7(c). The quadratic scaling $\Delta W \propto c_0^2$ (Corollary 8) is verified numerically in Appendix C.

6. Discussion and Conclusion

We introduced a Stackelberg boundary-insertion game to study handoff externalities under partial liability internalization in sequential AI workflows. The exact potential structure yields an over-fragmentation theorem, a scale-free welfare-loss characterization, and a closed-form optimal liability share, providing a tractable benchmark for accountability design in delegation chains.

Policy implications. The quadratic scaling $\Delta W \propto c_0^2$ (Corollary 8) suggests that reductions in enforcement costs, through audit logging, compliance tooling, and related infrastructure, can yield more-than-proportional reductions in residual welfare loss, consistent with regulatory moves toward log retention for high-risk AI systems [6]. The comparative statics and Figure 2 suggest that longer workflows and higher-productivity tasks are natural candidates for stronger liability internalization. We assume a linear handoff cost for tractability; the over-fragmentation direction is robust to convex coordination costs, which preserve the bias without changing its direction (Appendix A.8).

Scope and extensions. This paper provides a stylized benchmark under a line DAG, homogeneous developers, and a static setting. Our quantitative illustration uses a single parameter point, but the central qualitative results do not depend on it: the welfare loss $\mathcal{L}(\gamma)$ is scale-free (Corollary 6), so over-fragmentation and its dependence on γ and α hold across workflow sizes, productivities, and handoff-cost scales. Although stated for the line, our results are not specific to it: they carry over unchanged to any tree workflow and fail on reconverging DAGs (Remark 3). Heterogeneous developers could lead to Bayesian Stackelberg formulations; iterative or dynamically routed workflows lie outside the DAG abstraction and are left for future work, though the static equilibrium is reachable by decentralized adjustment (Remark 2). Model diversity within the delegation chain [2, 8] and simulation-based evaluation [15] are important complementary directions.

Acknowledgment

We thank the anonymous reviewers for their constructive comments and suggestions, which helped improve the clarity and overall quality of this paper.

Camera-Ready Edits Summary

Changes relative to the reviewed submission. The central claims, theorem statements, figures, and numerical results are unchanged.

1. Revisions addressing reviewer comments

The following revisions were adopted in full.

Theme	Revision
Convergence to equilibrium	Added Remark 2, which uses the finite improvement property and gradient ascent on the strictly concave potential to show that the equilibrium is reachable by decentralized adjustment.
Generality beyond the line DAG	Added Remark 3 and Proposition 9: the results extend verbatim to any tree workflow and fail on reconverging DAGs, with an explicit counterexample.
Role of the homogeneity assumption	Added Appendix A.7 separating developers (strategic players) from agents (execution units) and clarifying that aggregativeness, not symmetry, yields the potential structure.
Robustness to convex coordination costs	Added Appendix A.8 showing that the over-fragmentation direction persists under convex costs, with a corresponding sentence added to the discussion.
Empirical grounding and calibration	Added Appendix A.9 outlining how the model could be calibrated and what such data can and cannot test; introduced reference [17] as the supporting benchmark.
Interpretation of $\partial\gamma^*/\partial\omega$	Added a paragraph distinguishing coordination overhead from failure severity and relating higher-stakes workflows to larger a.

2. Corrections and minor edits

- Restored the in-text pointer to the proof of the optimal-liability theorem (Theorem 7) in Appendix A.4; the proof itself was already present.
- Introduced reference [17] (App. A.9) to support the calibration discussion; no other references were added or removed.
- Made minor clarifications and presentation edits, including an explicit graph definition of the workflow in Section 3.1, a tightened scope-and-extensions discussion in Section 6, and routine style normalization.
- Renumbered the numbered environments following the insertion of Remarks 2 and 3 (for example, the optimal-delegation result is now Theorem 4 and the optimal-liability result is now Theorem 7).

3. What remains unchanged

No new experiments were added; Appendix A.9 discusses calibration conceptually rather than reporting experiments. The figures and the single illustrative parameter point are unchanged, and all numerical values were re-checked for consistency.

References

[1] Gary S. Becker and Kevin M. Murphy. The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, 107(4):1137–1160, 1992. doi: 10.2307/2118383.

[2] Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/17a234c91f746d9625a75cf8a8731ee2-Abstract-Conference.html.

[3] Marcantonio Bracale Syrnikov, Federico Pierucci, Marcello Galisai, Matteo Prandi, Piercosma Bisconti, Francesco Giarrusso, Olga Sorokoletova, Vincenzo Suriani, and Daniele Nardi. Institutional AI: Governing LLM collusion in multi-agent Cournot markets via public governance graphs, 2026. URL https://arxiv.org/abs/2601.11369. arXiv preprint arXiv:2601.11369.

[4] Competition and Markets Authority. Agentic AI and consumers. Research and analysis, Competition and Markets Authority, March 2026. URL https://www.gov.uk/government/publications/agentic-ai-and-consumers/agentic-ai-and-consumers.

[5] Mert Demirer, John J. Horton, Nicole Immorlica, Brendan Lucier, and Peyman Shahidi. Chaining tasks, redefining work: A theory of AI automation. NBER Working Paper 34859, National Bureau of Economic Research, February 2026. URL https://www.nber.org/papers/w34859.

[6] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 june 2024 laying down harmonised rules on artificial intelligence (AI act). Official Journal of the European Union, 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[7] Luis Garicano. Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904, 2000. doi: 10.1086/317671.

[8] Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021. doi: 10.1073/pnas.2018340118.

[9] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14 (1):124–143, 1996. doi: 10.1006/game.1996.0044.

[10] Robert W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973. doi: 10.1007/BF01737559.

[11] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM*, 62(5): 32:1–32:42, 2015. doi: 10.1145/2806883.

[12] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. Practices for governing agentic AI systems. White paper, OpenAI, December 2023. URL https://openai.com/index/practices-for-governing-agentic-ai-systems/.

[13] Haileleol Tibebu and Hewan Shemtaga. The accountability horizon: An impossibility theorem for governing human–agent collectives, 2026. URL https://arxiv.org/abs/2604.07778. arXiv preprint arXiv:2604.07778.

[14] Helen Toner, John Bansemer, Kyle Crichton, Matt Burtell, Thomas Woodside, Anat Lior, Andrew J. Lohn, Ashwin Acharya, Beba Cibralic, Chris Painter, Cullen O'Keefe, Iason Gabriel, Kathleen Fisher, Ketan Ramakrishnan, Krystal Jackson, Noam Kolt, Rebecca Crootof, and Samrat Chatterjee. Through the chat window and into the real world: Preparing for AI agents. Workshop report, Center for Security and Emerging Technology, October 2024. URL https://cset.georgetown.edu/wp-content/uploads/CSET-Through-the-Chat-Window-and-Into-the-Real-World.pdf.

[15] Vedanta S P and Ponnurangam Kumaraguru. I can't believe it's corrupt: Evaluating corruption in multi-agent governance systems, 2026. URL https://arxiv.org/abs/2603.18894. arXiv preprint arXiv:2603.18894.

[16] Na Zhang, Kun Yue, and Chao Fang. A game-theoretic framework for AI governance, 2023. URL https://arxiv.org/abs/2305.14865. arXiv preprint arXiv:2305.14865.

[17] Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.421. URL https://aclanthology.org/2025.acl-long.421/.

Appendix A. Proofs

A.1. Proof of Theorem 4 (Optimal Delegation Depth)

The first-order condition of $W(k) = am^{1-\alpha}k^{\alpha} - \omega k$ is

$$W'(k) = \alpha a m^{1-\alpha} k^{\alpha-1} - \omega = 0.$$

Solving for k gives

$$k^* = \left(\frac{\alpha a m^{1-\alpha}}{\omega}\right)^{\frac{1}{1-\alpha}}.$$

Since $W''(k) = \alpha(\alpha-1)am^{1-\alpha}k^{\alpha-2} < 0$ for $\alpha \in (0,1)$, k^* is the global maximizer of W.

Using the first-order condition $\alpha a m^{1-\alpha}(k^*)^{\alpha-1} = \omega$, the optimal welfare is

$$W(k^*) = am^{1-\alpha}(k^*)^{\alpha} - \omega k^* = \frac{\omega k^*}{\alpha} - \omega k^* = \frac{\omega k^*(1-\alpha)}{\alpha}. \qquad \square$$

A.2. Proof of Proposition 1, Theorem 5, and Corollary 6

Proof of Proposition 1 (Potential Structure). In the boundary-insertion game (Eq. (2)), consider a deviation by player ℓ from $s_\ell = 0$ to $s_\ell = 1$. Let $\kappa_{-\ell} \equiv \sum_{j \neq \ell} s_j$. The number of agents before and after the deviation is $k_0 = 1 + \kappa_{-\ell}$ and $k_1 = k_0 + 1$, respectively. The change in player ℓ's payoff is

$$\Delta u_\ell = R(k_1) - \gamma\omega,$$

where $R(k) = \alpha a m^{1-\alpha} k^{\alpha-1}$. Define the discrete Rosenthal potential as

$$\Phi_\gamma^{\text{disc}}(\mathbf{s}) = \sum_{i=1}^{k(\mathbf{s})-1} \big[R(1+i) - \gamma\omega\big].$$

Its change under the deviation is

$$\Delta\Phi_\gamma^{\text{disc}} = \big[R(k_1) - \gamma\omega\big] = \Delta u_\ell.$$

Since $\Delta u_\ell = \Delta\Phi_\gamma^{\text{disc}}$ holds for every player ℓ and every strategy profile $\mathbf{s}_{-\ell}$, the game satisfies the definition of an exact potential game [9, 10]. Because γ affects only the coefficient of the cost term in the potential, this structure holds for all $\gamma \in (0,1]$.

Under the continuous relaxation $k \in \mathbb{R}_{>0}$, we define the continuous analogue of the discrete potential by integrating the marginal private gain:

$$\Phi_\gamma(k) \equiv \int_1^k \big[\alpha a m^{1-\alpha} u^{\alpha-1} - \gamma\omega\big]\, du = am^{1-\alpha}(k^\alpha - 1) - \gamma\omega(k-1).$$

Since $\Phi_\gamma''(k) = \alpha(\alpha-1)am^{1-\alpha}k^{\alpha-2} < 0$, Φ_γ is strictly concave. Thus, if the unconstrained maximizer lies in the interior of the feasible interval, the continuous relaxation admits a unique interior equilibrium, which coincides with the maximizer of Φ_γ. Otherwise, the solution is the corresponding boundary solution. $\square$

Proof of Theorem 5 (Over-Fragmentation). By Proposition 1, the equilibrium is characterized by the first-order condition of the potential $\Phi_\gamma(k)$:

$$\Phi'_\gamma(k) = \alpha a m^{1-\alpha} k^{\alpha-1} - \gamma\omega = 0.$$

Solving for $\hat{k}$ yields

$$\hat{k}(\gamma) = \left(\frac{\alpha a m^{1-\alpha}}{\gamma\omega}\right)^{\frac{1}{1-\alpha}} = k^* \cdot \gamma^{-\frac{1}{1-\alpha}}.$$

For $\gamma \in (0,1)$, $\gamma^{-1/(1-\alpha)} > 1$, so $\hat{k}(\gamma) > k^*$. Over the extended domain $\gamma \in (0,1]$, equality holds if and only if $\gamma = 1$. □

Derivation of Corollary 6 (Equilibrium Welfare Loss). Let $r \equiv \gamma^{-1/(1-\alpha)} = \hat{k}(\gamma)/k^*$. The equilibrium welfare is

$$\begin{aligned} W(\hat{k}(\gamma)) &= a m^{1-\alpha}(k^* r)^\alpha - \omega(k^* r) \\ &= \omega k^* \left(\frac{r^\alpha}{\alpha} - r\right), \end{aligned}$$

where the last line uses $a m^{1-\alpha}(k^*)^\alpha = \omega k^*/\alpha$ (from the first-order condition). Since $W(k^*) = \omega k^*(1-\alpha)/\alpha$, the relative welfare loss is

$$\begin{aligned} \mathcal{L}(\gamma) &= \frac{W(k^*) - W(\hat{k}(\gamma))}{W(k^*)} \\ &= \frac{(1-\alpha)/\alpha - (r^\alpha/\alpha - r)}{(1-\alpha)/\alpha} \\ &= \frac{(1-\alpha) + \alpha\, r - r^\alpha}{1-\alpha}. \end{aligned}$$

Since r depends only on γ and α, $\mathcal{L}$ is independent of m, a, and ω. At $\gamma = 1$, $r = 1$ and $\mathcal{L}(1) = [(1-\alpha) + \alpha - 1]/(1-\alpha) = 0$. Monotonicity of $\mathcal{L}$ in γ follows from $d\mathcal{L}/d\gamma < 0$, since $dr/d\gamma < 0$ and $d\mathcal{L}/dr = (\alpha - \alpha r^{\alpha-1})/(1-\alpha) > 0$ for $r > 1$. □

A.3. Properties of Equilibrium Welfare

We show that the equilibrium welfare $W_R(\gamma) \equiv W(\hat{k}(\gamma))$ is strictly increasing and strictly concave on $(0,1)$.

First derivative. By the chain rule,

$$W'_R(\gamma) = W'(\hat{k}(\gamma)) \cdot \hat{k}'(\gamma).$$

We compute each factor. From $\hat{k}(\gamma) = k^* \gamma^{-1/(1-\alpha)}$,

$$\hat{k}'(\gamma) = -\frac{k^*}{1-\alpha} \cdot \gamma^{-\frac{2-\alpha}{1-\alpha}}.$$

Substituting $k = \hat{k}(\gamma)$ into $W'(k) = \alpha a m^{1-\alpha} k^{\alpha-1} - \omega$ and noting that $\hat{k}(\gamma)^{\alpha-1} = (k^*)^{\alpha-1} \cdot \gamma^{-(\alpha-1)/(1-\alpha)} = (k^*)^{\alpha-1} \cdot \gamma$ (since $-(\alpha-1)/(1-\alpha) = 1$), we use the first-order condition $\alpha a m^{1-\alpha}(k^*)^{\alpha-1} = \omega$ to obtain

$$W'(\hat{k}(\gamma)) = \omega\gamma - \omega = \omega(\gamma - 1).$$

Therefore,

$$W_R'(\gamma) = \omega(\gamma - 1) \cdot \left(-\frac{k^*}{1-\alpha}\right)\gamma^{-\eta} = \frac{\omega k^*}{1-\alpha}(1-\gamma)\gamma^{-\eta},$$

where $\eta \equiv (2-\alpha)/(1-\alpha)$. For $\gamma \in (0,1)$, $(1-\gamma) > 0$ and $\gamma^{-\eta} > 0$, so $W_R'(\gamma) > 0$, establishing strict monotonicity.

Strict concavity. Differentiating $W_R'(\gamma)$ again:

$$\begin{aligned} W_R''(\gamma) &= \frac{\omega k^*}{1-\alpha}\frac{d}{d\gamma}\left[(1-\gamma)\gamma^{-\eta}\right] \\ &= \frac{\omega k^*}{1-\alpha}\left[-\gamma^{-\eta} - \eta(1-\gamma)\gamma^{-\eta-1}\right] \\ &= \frac{\omega k^*}{1-\alpha} \cdot \gamma^{-\eta-1}[(\eta-1)\gamma - \eta]\,. \end{aligned}$$

For $\gamma \in (0,1)$ and $\eta > 2$, $(\eta-1)\gamma \leq \eta - 1 < \eta$, so the bracketed term is strictly negative. Hence $W_R''(\gamma) < 0$ for all $\gamma \in (0,1)$, establishing strict concavity.

Boundary behavior. As $\gamma \to 0^+$, $\gamma^{-\eta} \to +\infty$ and $(1-\gamma) \to 1$, so $W_R'(\gamma) \to +\infty$. At $\gamma = 1$, $(1-\gamma) = 0$, so $W_R'(1) = 0$. Moreover, $W_R(1) = W(\hat{k}(1)) = W(k^*)$ since $\hat{k}(1) = k^*$. □

A.4. Proof of Theorem 7 (Optimal Liability Regulation)

Part (a): Existence and uniqueness. Let $R(\gamma) = W_R(\gamma) - C(\gamma - \gamma_0)$. Since W_R is strictly concave (Appendix A.3) and $-C$ is concave (C being convex), R is strictly concave on $[\gamma_0, 1]$. The interval $[\gamma_0, 1]$ is compact and R is continuous, so a maximizer exists; strict concavity implies uniqueness. The maximizer may lie at a boundary of $[\gamma_0, 1]$.

Part (b): Interior solution and FOC. We have $R'(\gamma) = W_R'(\gamma) - C'(\gamma - \gamma_0)$. At $\gamma = \gamma_0$, $R'(\gamma_0) = W_R'(\gamma_0) - C'(0) > 0$ by assumption, so $\gamma^* > \gamma_0$. At $\gamma = 1$, $R'(1) = W_R'(1) - C'(1 - \gamma_0) = -C'(1-\gamma_0) < 0$ since $C' > 0$, so $\gamma^* < 1$. Since R is strictly concave with $R'(\gamma_0) > 0$ and $R'(1) < 0$, the equation $R'(\gamma^*) = 0$ has a unique solution in $(\gamma_0, 1)$.

Using the expression for $W_R'(\gamma)$ derived in Appendix A.3, the first-order condition $R'(\gamma^*) = 0$ becomes

$$\frac{\omega k^*}{1-\alpha}(1-\gamma^*)(\gamma^*)^{-\eta} = C'(\gamma^* - \gamma_0),$$

where $\eta = (2-\alpha)/(1-\alpha)$. □

Part (c): Comparative statics. Write the first-order condition as $G(\gamma^*, \theta) \equiv W_R'(\gamma^*) - C'(\gamma^* - \gamma_0) = 0$. By the implicit function theorem,

$$\frac{\partial \gamma^*}{\partial \theta} = -\frac{\partial G/\partial \theta}{\partial G/\partial \gamma^*}.$$

Since $\partial G/\partial \gamma^* = W_R''(\gamma^*) - C''(\gamma^* - \gamma_0) < 0$ (W_R strictly concave, C convex), the sign of $\partial \gamma^*/\partial \theta$ coincides with the sign of $\partial G/\partial \theta$.

Writing $W_R'(\gamma) = \Lambda \cdot (1-\gamma)\gamma^{-\eta}$ with scalar coefficient $\Lambda \equiv \omega k^*/(1-\alpha)$, and using $k^* = (\alpha a m^{1-\alpha}/\omega)^{1/(1-\alpha)}$, we obtain

$$\Lambda = \frac{\omega}{1-\alpha}\left(\frac{\alpha a m^{1-\alpha}}{\omega}\right)^{\frac{1}{1-\alpha}} = \frac{(\alpha a)^{\frac{1}{1-\alpha}} \cdot m \cdot \omega^{-\frac{\alpha}{1-\alpha}}}{1-\alpha}.$$

(i) For $C(\Delta) = c_0 \cdot f(\Delta)$, $\partial G/\partial c_0 = -f'(\gamma^* - \gamma_0) < 0$, so $\partial\gamma^*/\partial c_0 < 0$.
(ii) Since $\Lambda \propto m$, $\partial\Lambda/\partial m = \Lambda/m > 0$, so $\partial W_R'/\partial m > 0$ and $\partial\gamma^*/\partial m > 0$.
(iii) Since $\Lambda \propto a^{1/(1-\alpha)}$, $\partial\Lambda/\partial a > 0$, so $\partial W_R'/\partial a > 0$ and $\partial\gamma^*/\partial a > 0$.
(iv) Since $\Lambda \propto \omega^{-\alpha/(1-\alpha)}$ and the exponent $-\alpha/(1-\alpha) < 0$, $\partial\Lambda/\partial\omega < 0$, so $\partial W_R'/\partial\omega < 0$ and $\partial\gamma^*/\partial\omega < 0$.

Economic intuition: When the handoff cost ω is large, the first-best delegation depth k^* is itself small ($k^* \propto \omega^{-1/(1-\alpha)}$). The magnitude of over-fragmentation $\hat{k} - k^*$ is also proportional to $\omega^{-1/(1-\alpha)}$, so the absolute extent of excessive fragmentation shrinks as ω increases, reducing the marginal benefit of regulatory correction. □

A.5. Derivation of Corollary 8 (Low-Cost Regime)

Optimal liability approximation. Under linear enforcement cost $C(\Delta) = c_0\Delta$, the first-order condition is

$$\frac{\omega k^*}{1-\alpha}(1-\gamma^*)(\gamma^*)^{-\eta} = c_0.$$

Assuming γ^* is close to 1, let $\epsilon = 1-\gamma^*$. A Taylor expansion $(\gamma^*)^{-\eta} = (1-\epsilon)^{-\eta} \approx 1 + \eta\epsilon$ gives

$$\frac{\omega k^*}{1-\alpha} \cdot \epsilon \cdot (1+\eta\epsilon) \approx c_0.$$

Retaining only terms of order ϵ,

$$\epsilon \approx \frac{c_0(1-\alpha)}{\omega k^*},$$

that is, $\gamma^* \approx 1 - c_0(1-\alpha)/(\omega k^*)$. This approximation is valid when $c_0 \ll \omega k^*/(1-\alpha)$. □

Quadratic welfare loss. Since $W_R(1) = W(k^*)$ and $W_R'(1) = 0$ (Appendix A.3),

$$\Delta W = W(k^*) - W_R(\gamma^*) = \int_{\gamma^*}^{1} W_R'(\gamma)\, d\gamma.$$

In the low-cost regime where γ^* is close to 1, we Taylor-expand $W_R'(\gamma)$ around $\gamma = 1$. Since $W_R'(1) = 0$,

$$W_R'(\gamma) \approx W_R''(1)(\gamma - 1) = |W_R''(1)|(1-\gamma),$$

where the last equality uses $W_R''(1) < 0$ and $\gamma - 1 < 0$. Substituting $\gamma = 1$ into the general expression for the second derivative from Appendix A.3,

$$W_R''(1) = \frac{\omega k^*}{1-\alpha}[(\eta - 1) - \eta] = -\frac{\omega k^*}{1-\alpha}.$$

Hence $|W_R''(1)| = \omega k^*/(1-\alpha)$, and

$$\Delta W \approx \int_{\gamma^*}^{1} |W_R''(1)|(1-\gamma)\, d\gamma = \frac{|W_R''(1)|}{2}(1-\gamma^*)^2$$
$$= \frac{\omega k^*}{2(1-\alpha)} \left(\frac{c_0(1-\alpha)}{\omega k^*}\right)^2 = \frac{c_0^2(1-\alpha)}{2\,\omega\, k^*}.$$

The welfare loss scales quadratically with the enforcement cost c_0. □

A.6. Extension to Tree Workflows

Our results depend on the workflow only through *aggregativeness*: each boundary player's marginal value $R(k(s)) - \gamma\omega$ depends on the profile $\mathbf{s}$ only through the agent count $k(s)$. We show that this property, and the results that follow from it, extend from the line to any tree.

Proposition 9 (Tree Workflows) *Let the workflow be any connected graph whose undirected version is a tree, with boundary players indexed by its edges and homogeneous developers. Then $k(s) = 1 + \sum_e s_e$ for every profile s, the boundary-insertion game is an exact potential game with the potential of Proposition 1, and Theorem 5 and Corollary 6 hold verbatim.*

Proof In a tree every edge is a bridge, so removing a set F of edges yields exactly $|F| + 1$ connected components. Writing $F(s) = \{e : s_e = 1\}$ for the boundaries inserted under profile s and identifying each agent with a connected component of the cut graph, the agent count is $k(s) = |F(s)| + 1 = 1 + \sum_e s_e$. The marginal value of inserting boundary e is then $R(k(s)) - \gamma\omega$ with $R(k) = \alpha a m^{1-\alpha} k^{\alpha-1}$, exactly as in Eq. (2). Under homogeneity the aggregate output is $am^{1-\alpha}k^{\alpha}$ and the welfare $W(k)$ is unchanged, so the discrete Rosenthal potential, the deviation identity $\Delta u_\ell = \Delta\Phi_\gamma^{\text{disc}}$ (Appendix A.2), and the continuous potential Φ_γ all apply without modification. The equilibrium $\hat{k}(\gamma)$ in (3), the over-fragmentation conclusion, and the scale-free loss $\mathcal{L}(\gamma)$ are thus identical. ■

As on the line, this uses the continuous relaxation and the homogeneous equal-mass partition; a connected equal-mass partition into k pieces need not exist for every integer k on a given tree (for example, a star), so the same integer-k caveat applies.

The tree structure is necessary. If the undirected graph contains a cycle, cutting an edge need not increase the number of components, so the count $k(s)$ is no longer $1 + \sum_e s_e$ and a boundary's marginal value depends on which other boundaries are cut. In the reconverging workflow $1 \to \{2,3\} \to 4$ with edges $(1,2), (1,3), (2,4), (3,4)$, removing $(1,2)$ leaves the graph connected and the agent count unchanged, so aggregativeness fails and the exact-potential construction no longer applies.

A.7. On the Homogeneity Assumption

We distinguish two roles. *Developers* are the strategic players who choose the partition; *agents* are the units that execute the resulting subtasks, whose capability differences are absorbed into the production term $a|S_i|^{1-\alpha}$. Heterogeneity among agents (for instance, a strong orchestrator with weaker workers) is therefore already accommodated at the level of the production function. Homogeneity plays a more limited role than it may appear. The exact-potential structure (Proposition 1) does not

rest on symmetric players: Rosenthal-type potential games admit heterogeneous players, and what the construction requires is aggregativeness, namely that each boundary's marginal value depend on the profile only through the agent count $k(s)$. What homogeneity delivers is the symmetric partition $|S_i| = m/k$, under which the aggregate output is $am^{1-\alpha}k^{\alpha}$ and the welfare loss $\mathcal{L}(\gamma)$ in Corollary 6 is scale-free. Relaxing it to developer-specific handoff costs or unequal scopes would preserve the over-fragmentation force: each developer still internalizes only a fraction γ of the marginal handoff cost, so private incentives favor finer partitions than the planner would choose whenever $\gamma < 1$. It would, however, replace the closed-form $\mathcal{L}$ with a developer-dependent expression and, under partition-dependent payoffs, may forgo the exact potential and the single-equilibrium guarantee.

A.8. Robustness to Convex Coordination Costs

The model uses a linear handoff cost ωk for tractability. Empirically, coordination quality can degrade faster than linearly as a chain lengthens: benchmark evaluations report degradation in coordination quality as chains lengthen and iterations accumulate [17], which motivates a convex coordination cost $c(k)$ with $c'(k) > 0$ and $c''(k) > 0$. The over-fragmentation direction is robust to this change. With welfare $W(k) = am^{1-\alpha}k^{\alpha} - c(k)$, the social optimum solves $\alpha am^{1-\alpha}k^{\alpha-1} = c'(k)$, while a developer internalizing a fraction γ of the marginal coordination cost equates $\alpha am^{1-\alpha}k^{\alpha-1} = \gamma\, c'(k)$. Because the marginal benefit $R(k) = \alpha am^{1-\alpha}k^{\alpha-1}$ is strictly decreasing in k, any $\gamma < 1$ yields an equilibrium depth above the optimum, so $\hat{k}(\gamma) > k^*$ for every increasing marginal cost, linear or convex. Convexity changes the curvature of the first-order conditions but not their ordering, so the bias persists; its effect on the absolute size of the wedge $\hat{k}(\gamma) - k^*$ depends on the curvature of c and is not monotone in general. What a convex $c(k)$ does change is the closed form: the scale-free loss $\mathcal{L}(\gamma)$ of Corollary 6 relies on the power-law structure and need not survive, and partition-dependent coordination costs may forgo the exact potential.

A.9. Toward Empirical Calibration

Our quantitative results use a single synthetic parameter point, and the framework has not been confronted with agentic-workflow data. We outline how it could be calibrated and what such data could and could not test. Multi-agent benchmarks provide a natural starting point. In particular, [17] evaluate LLM agent teams under several coordination protocols, of which the *chain* protocol, where each agent passes its output to the next, corresponds to our sequential workflow. Their agent-count ablation is qualitatively consistent with the primitives of our welfare model $W(k) = am^{1-\alpha}k^{\alpha} - \omega k$: as the number of agents grows, per-agent milestone contribution falls and the coordination score gains diminish after an initial improvement, matching the concavity of the specialization term against a rising coordination cost. The specialization gain α could in principle be estimated from how aggregate task quality scales with the number of agents at fixed scope, and ω from the incremental coordination cost of an added handoff. Two limits should be stated plainly. First, such a benchmark fixes the number of agents exogenously and has no liability instrument or partition-choice game, so it can identify the primitives of $W(k)$ but cannot directly test the equilibrium prediction $\hat{k}(\gamma) > k^*$, which requires developers who bear a share γ of handoff costs and choose the partition. Second, a single benchmark with saturating quality scores and few distinct team-size settings does not support a precise estimate of α; we therefore report the correspondence as qualitative rather than as a

fitted calibration. Beyond requiring such developers, testing over-fragmentation directly would also require an independent estimate of the socially optimal depth k^* as a comparison point.

Appendix B. Pigouvian Tax Equivalence

We state the equivalence noted in Section 3.3 formally.

Proposition 10 (Pigouvian Tax Equivalence) *Consider a regime in which the regulator imposes a per-handoff levy $t \in [0,\ (1-\gamma_0)\omega]$. The developer's effective handoff cost becomes $\gamma_0\omega + t$, and the equilibrium delegation depth is*

$$\hat{k}(t) = \left(\frac{\alpha a m^{1-\alpha}}{\gamma_0\omega + t}\right)^{\frac{1}{1-\alpha}}.$$

Defining the effective liability share as $\gamma_{\text{eff}} = \gamma_0 + t/\omega$, we have $\gamma_{\text{eff}} \in [\gamma_0, 1]$ and $\hat{k}(t) = \hat{k}(\gamma_{\text{eff}})$.

That is, raising the liability share from γ_0 to γ is equivalent to imposing a per-handoff levy of $t = (\gamma - \gamma_0)\omega$. In particular, $t^* = (\gamma^* - \gamma_0)\omega$ is the optimal levy corresponding to the solution of the regulator's problem (4), consistent with the first-order condition of Theorem 7.

Proof Substituting $\gamma_{\text{eff}} = \gamma_0 + t/\omega$ into the equilibrium delegation-depth formula (3) gives the stated equality. ■

Appendix C. Supplementary Numerical Results

This appendix presents supplementary numerical results that complement the analysis in Section 5. All figures use the baseline parameters $m = 10$, $\alpha = 0.3$, $a = \omega = 1.0$, $\gamma_0 = 0.1$.

C.1. Quadratic Welfare-Loss Scaling

Corollary 8 establishes that the residual welfare loss under optimal regulation scales as $\Delta W \propto c_0^2$ in the low-cost regime. Figure 3 provides numerical verification of this result.

The figure plots $\Delta W = W(k^*) - W_R(\gamma^*)$ against the enforcement cost c_0 on a log-log scale over the range $c_0 \in [10^{-3}, 10^3]$. The exact solution (solid blue) is computed by numerically solving the first-order condition (7) for each value of c_0 and evaluating $W_R(\gamma^*)$. The leading-term approximation $c_0^2(1-\alpha)/(2\omega k^*)$ (dashed red) is overlaid for comparison.

In the low-cost regime ($c_0 \ll \omega k^*/(1-\alpha)$), the two curves are nearly indistinguishable. A linear fit on the log-log scale yields a slope of ≈ 1.98, close to the theoretical value of 2.0. Deviations become visible only for $c_0 \gtrsim 1$, where higher-order terms in the Taylor expansion of $W_R'(\gamma)$ become non-negligible.

The quadratic scaling has a concrete policy implication: halving the enforcement cost c_0 reduces the residual welfare loss by roughly a factor of four. This suggests that investments in enforcement infrastructure, such as audit logging, compliance tooling, and standardized handoff protocols, can yield more-than-proportional reductions in residual welfare loss. Here c_0 is a reduced-form parameter capturing the marginal cost of raising the effective liability share; the result is consistent with recent regulatory trends requiring log retention and transparency for high-risk AI systems [6].

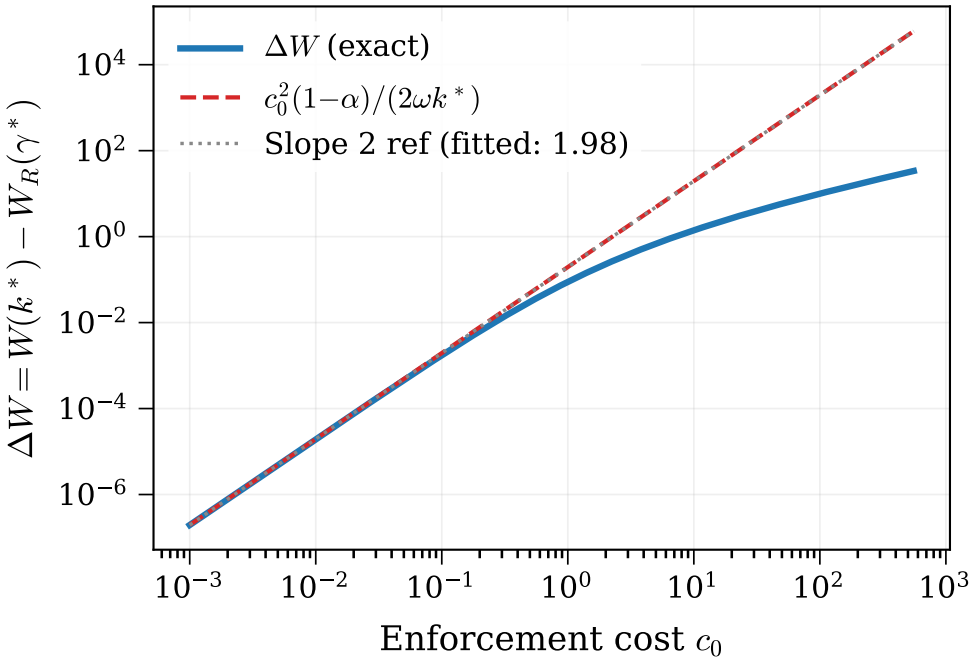

Figure 3: Welfare loss $\Delta W = W(k^*) - W_R(\gamma^*)$ versus enforcement cost c_0 (log-log scale). The exact solution (solid blue) closely matches the leading-term approximation $c_0^2(1-\alpha)/(2\omega k^*)$ (dashed red) in the low-cost regime, with a fitted slope of ≈ 1.98 (theoretical value 2.0; Corollary 8). Parameters: $m = 10$, $\alpha = 0.3$, $a = \omega = 1.0$, $\gamma_0 = 0.1$.

C.2. Approximation Accuracy of Optimal Regulation

Corollary 8 provides a closed-form approximation $\gamma^* \approx 1 - c_0(1-\alpha)/(\omega k^*)$ that is valid when $c_0 \ll \omega k^*/(1-\alpha)$. Figure 4 assesses the accuracy of this approximation across the full range of enforcement costs.

The figure plots the optimal liability share γ^* as a function of the normalized enforcement cost $\rho \equiv c_0/\Lambda_0$, where $\Lambda_0 \equiv \omega k^*/(1-\alpha)$. This normalization is chosen so that the first-order condition (7) under linear cost reduces to $(1-\gamma^*)(\gamma^*)^{-\eta} = \rho$, making ρ the natural dimensionless measure of enforcement cost.

Three regimes are visible in the figure:

- *Low-cost regime* ($\rho \ll 1$): $\gamma^* \to 1$, corresponding to near-complete liability internalization. The low-cost approximation (dashed orange) is accurate in this region.
- *Intermediate regime*: γ^* decreases smoothly, and the low-cost approximation begins to deviate from the exact solution.
- *High-cost regime* ($\rho \to \rho_{\max} \equiv (1-\gamma_0)\gamma_0^{-\eta}$): γ^* approaches the baseline γ_0. For $\rho \geq \rho_{\max}$, the marginal welfare gain $W_R'(\gamma_0)$ is dominated by the marginal enforcement cost c_0, so additional regulation is not cost-effective and the constrained optimum is $\gamma^* = \gamma_0$.

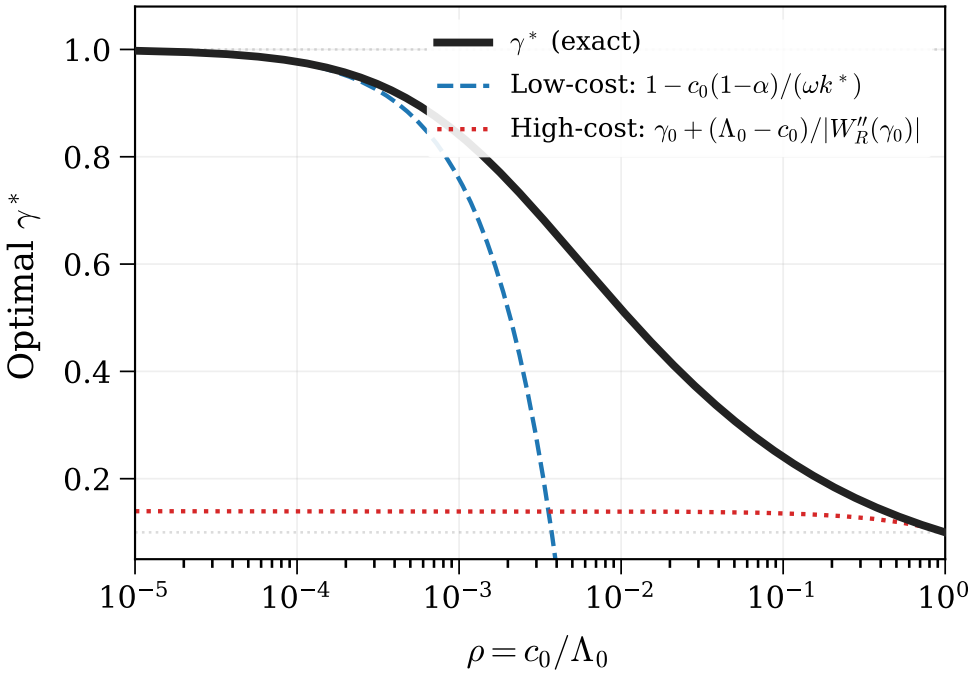

Figure 4: Optimal liability share γ^* as a function of the normalized enforcement cost $\rho \equiv c_0/\Lambda_0$. The low-cost approximation (dashed orange; Corollary 8) is accurate for small ρ. As ρ increases toward $\rho_{\max} = (1-\gamma_0)\gamma_0^{-\eta}$, γ^* approaches the baseline liability level γ_0. Parameters: $\alpha = 0.3$, $a = \omega = 1.0$, $\gamma_0 = 0.1$.

The transition from the low-cost to the high-cost regime is governed by the exponent $\eta = (2-\alpha)/(1-\alpha)$, which controls the curvature of $W_R'(\gamma)$ near $\gamma = 1$. For $\alpha = 0.3$, $\eta \approx 2.43$, so the marginal welfare gain decays rapidly as γ moves away from 1, creating a sharp transition in the $\gamma^*(\rho)$ curve.

C.3. Scale-Free Structure of Welfare Loss

The relative welfare loss $\mathcal{L}(\gamma)$ from Corollary 6 depends only on γ and α, and is independent of the workflow size m, agent productivity a, and handoff cost ω. This scale-free property has several implications worth elaborating.

First, $\mathcal{L}$ serves as a *design diagnostic* that allows direct comparison of over-delegation severity across workflows of different scales. A liability share of $\gamma = 0.4$ under $\alpha = 0.3$ yields $\mathcal{L} \approx 0.47$ regardless of whether the workflow has $m = 5$ or $m = 500$ subtasks, or whether agent productivity is high or low. This means that the *proportional* welfare loss from over-fragmentation is a universal function of the governance parameters alone.

Second, the scale-free property clarifies the relationship to the classical Price of Anarchy (PoA). Since $\mathcal{L} = 1 - 1/\text{PoA}$, Corollary 6 provides a closed-form characterization of the PoA under partial liability internalization. Unlike smoothness-based PoA bounds [11], which yield worst-case guarantees over broad classes of games, our expression is *exact* for the handoff-externality model.

The PoA ratio $W(k^*)/W(\hat{k}(\gamma))$ is defined only when $W(\hat{k}) > 0$, which requires $\gamma > \alpha$. By contrast, $\mathcal{L}$ remains well-defined on the full domain $(0, 1]$: when γ is sufficiently small, $W(\hat{k}(\gamma))$ becomes negative (the coordination costs exceed the output), and $\mathcal{L}$ exceeds one.

Third, $\mathcal{L}$ grows rapidly in the region where α is large and γ is small. Workflows with strong specialization gains (large α) are especially sensitive to weak liability internalization: the marginal benefit of each additional agent is high, encouraging excessive fragmentation when handoff costs are only partially borne. Under the baseline $\alpha = 0.3$ and $\gamma_0 = 0.1$, for example, $\mathcal{L} \approx 8.66$, indicating that the equilibrium welfare is not merely below the first-best but in fact negative: the coordination costs of the excessively fragmented workflow overwhelm the aggregate output. At the optimal $\gamma^* = 0.863$, the residual welfare loss falls to $\mathcal{L} \approx 0.007$, demonstrating that even partial regulatory correction can recover nearly all of the first-best welfare.

C.4. Counterintuitive Comparative Static: $\partial\gamma^*/\partial\omega < 0$

The sign $\partial\gamma^*/\partial\omega < 0$ in Theorem 7(c) may appear counterintuitive: one might expect that higher handoff costs would call for stricter regulation to curb fragmentation. The opposite holds because ω rescales the problem rather than measuring how harmful a failure is; the effect operates through the *scale* of the problem, not the *severity* of the externality.

When ω is large, the first-best delegation depth k^* is itself small ($k^* \propto \omega^{-1/(1-\alpha)}$), because the planner also responds to high handoff costs by using fewer agents. The equilibrium overshoot $\hat{k}(\gamma) - k^* = k^*(\gamma^{-1/(1-\alpha)} - 1)$ is proportional to k^*, so the *absolute* magnitude of over-fragmentation shrinks as ω increases. Concretely, writing $\Lambda \equiv \omega k^*/(1-\alpha)$, we have $\Lambda \propto \omega^{-\alpha/(1-\alpha)}$ with a negative exponent. Since $W_R'(\gamma) = \Lambda \cdot (1-\gamma)\gamma^{-\eta}$, a higher ω reduces the marginal welfare gain from raising γ at every point, shifting the optimal balance toward less costly regulation.

This can be summarized as follows: high handoff costs make the first-best itself frugal, so there is less over-fragmentation to correct and less marginal benefit from regulatory intervention. The result is consistent with the intuition that regulation is most valuable when the gap between private and social incentives is large *in absolute terms*: this occurs when ω is moderate, not when it is extreme.