Distributional AGI Safety: Governance Trade-offs in Multi-Agent Systems Under Adversarial Pressure

Raeli Savitt

February 2026

Abstract

We study governance trade-offs in multi-agent AI systems using a probabilistic simulation framework that replaces binary safety labels with calibrated soft scores $p = P(v = +1)$. Across 11 scenarios (209 total epochs, 81 agent-slots), we find that ecosystem outcomes cluster into three regimes: *cooperative* (acceptance > 0.93, toxicity < 0.30), *contested* (acceptance 0.42–0.94, toxicity 0.33–0.37), and *adversarial collapse* (acceptance < 0.56, collapse by epoch 12–14). Collapse occurred exclusively when adversarial fraction reached 50%, and governance tuning delayed but did not prevent it—shifting collapse from epoch 12 to 14 across three red-team variants. Collusion detection proved critical: a scenario with 37.5% adversarial agents avoided collapse entirely when pair-wise frequency and correlation monitoring were enabled, while comparable scenarios without it collapsed. Incoherence metrics scaled sub-linearly with agent count, while welfare scaled super-linearly, suggesting that larger cooperative populations are disproportionately productive but also harder to monitor. These results formalize the intuition from market microstructure theory that adverse selection in agent ecosystems is regime-dependent: governance interventions that suffice under moderate adversarial pressure fail abruptly beyond a critical threshold.

1 Introduction

As AI systems increasingly operate as autonomous agents—negotiating, collaborating, and competing within shared digital environments—the safety question shifts from aligning a single model to governing an ecosystem of interacting agents with heterogeneous objectives. A growing body of work addresses multi-agent safety through mechanism design [Myerson, 1981, Hurwicz, 1960], distributional analysis [Kenton et al., 2025], and economic governance frameworks [Tomasev et al., 2025]. Yet empirical study of *how* and *when* governance interventions fail under adversarial pressure remains limited, in part because most evaluations use binary safety labels (safe/unsafe) that obscure the probabilistic, continuous nature of real interaction quality.

This paper takes a different approach. Drawing on market microstructure theory—specifically the adverse selection models of Kyle [1985] and Glosten and Milgrom [1985]—we model multi-agent ecosystems as markets in which agents with private information about interaction quality choose whether and how to participate. Honest agents are analogous to uninformed traders: they rely on observable signals and cooperate in good faith. Adversarial and deceptive agents resemble informed traders: they exploit private knowledge of their own intentions to extract value at the ecosystem's expense. The governance mechanism—acceptance thresholds, audits, circuit breakers—plays the role of the market maker, setting terms of participation that must balance the cost of excluding legitimate interactions against the risk of admitting harmful ones.

Central to our framework is the replacement of binary safety labels with *soft probabilistic labels*: each interaction receives a calibrated score $p = P(v = +1)$, the probability that its true value is beneficial. This follows the distributional safety framework of Kenton et al. [2025], which argues that safety properties are better characterized by distributions over outcomes than by point classifications. Probabilistic labels enable continuous metrics—toxicity as $\mathbb{E}[1 - p \mid \text{accepted}]$, quality gap as the difference in expected p between accepted and rejected interactions—that capture adverse selection dynamics, the "lemons problem" [Akerlof, 1970], invisible to binary classification.

We implement this framework in SWARM (System-Wide Assessment of Risk in Multi-agent systems), a configurable simulation environment supporting multiple agent behavioral types, governance lever combinations, network topologies, and economic mechanisms including Dworkin-style resource auctions [Dworkin, 1981], Shapley-value reward allocation [Shapley, 1953], and mission economies [Tomasev et al., 2025]. Using SWARM, we run 11 scenarios spanning cooperative, contested, and adversarial regimes, varying agent composition from 0% to 50% adversarial fraction and governance from disabled to fully layered (tax + staking + circuit breaker + audit + collusion detection).

Our central research questions are:

1. **Is there a critical adversarial fraction** beyond which governance interventions fail to prevent ecosystem collapse, and if so, where does it lie?

2. **Which governance levers** provide qualitatively different protection (extending the viable operating range) versus quantitatively incremental improvement (delaying but not preventing collapse)?

3. **How do safety metrics and welfare scale** with agent population size and network density?

We find that ecosystem outcomes partition cleanly into three regimes, that the collapse boundary lies between 37.5% and 50% adversarial fraction, and that collusion detection—a structural governance lever operating on interaction patterns rather than individual agents—is the critical differentiator between survival and collapse in contested environments.

2 Related Work

Multi-agent safety. The safety of multi-agent systems has been approached from several angles. Dafoe et al. [2020] survey cooperative AI, framing the challenge as designing agents that can collaborate despite misaligned incentives. Leibo et al. [2017] study emergent social dilemmas in multi-agent reinforcement learning, demonstrating that competitive dynamics arise even among independently trained cooperative agents. Our work complements these by focusing on *governance* as the mechanism for maintaining cooperation, rather than relying on agent-level alignment.

Distributional safety. Kenton et al. [2025] introduce the distributional safety framework, arguing that safety properties should be characterized by outcome distributions rather than binary labels. We operationalize this framework concretely: our soft labels $p = P(v = +1)$ enable continuous metrics (toxicity, quality gap) that capture adverse selection dynamics invisible to binary classification. Anthropic's "hot mess" theory [Anthropic, 2026] extends this intuition to variance-dominated failure modes, where the danger lies not in expected outcomes but in heavy-tailed distributions of harm—a framing consistent with our observation that toxicity saturates while welfare variance grows with scale.

Mechanism design for AI. The application of economic mechanism design to AI governance draws on classical results from Myerson [1981] and Hurwicz [1960]. More recently, Tomasev et al. [2025] propose virtual agent economies as a governance layer for multi-agent systems, using economic incentives (taxes, staking, auctions) to align agent behavior. Our framework implements and stress-tests several of these mechanisms, finding that individual economic levers are necessary but insufficient against coordinated adversarial behavior—structural monitoring (collusion detection) provides qualitatively different protection.

Market microstructure analogies. We draw heavily on the adverse selection models of Kyle [1985] and Glosten and Milgrom [1985], treating the governance mechanism as a market maker that must set terms of participation under asymmetric information. Akerlof's lemons problem [Akerlof, 1970] provides the conceptual foundation: when the governance threshold cannot distinguish high-quality from low-quality interactions, adverse selection drives out cooperative agents. The flash crash literature [Kyle et al., 2017] informs our analysis of cascading failure in network topologies, where contagion through dense connections can amplify local failures into systemic collapse.

Agent ecosystems. Chen et al. [2025] study multi-agent market dynamics in cooperative settings, finding that network topology shapes emergent specialization. Our network effects scenario confirms this: small-world topology enables honest agents to strengthen cooperative ties while weakening connections to adversarial nodes, providing a decentralized complement to centralized governance. Park et al. [2023] demonstrate emergent social behavior in generative agent populations, motivating our inclusion of behavioral heterogeneity (honest, opportunistic, deceptive, adversarial) as a first-class experimental variable.

3 Experimental Setup

3.1 Scenarios

Table 1 summarizes the 11 scenarios. Agent types are: **H** = honest, **O** = opportunistic, **D** = deceptive, **Adv** = adversarial, **AA** = adaptive adversary.

Table 1: Scenario configurations.

Scenario	Agents	Composition	Adv. %	Epochs	Network
baseline	5	3H, 1O, 1D	20.0	10	none
adversarial_redteam_v1	8	4H, 2Adv, 2AA	50.0	30	small-world
adversarial_redteam_v2	8	4H, 2Adv, 2AA	50.0	30	small-world
adversarial_redteam_v3	8	4H, 2Adv, 2AA	50.0	30	small-world
collusion_detection	8	3H, 2O, 3Adv	37.5	25	small-world
emergent_capabilities	8	6H, 2O	0.0	30	complete
incoherence_short	3	3H	0.0	8	none
incoherence_medium	6	5H, 1O	16.7	8	none
incoherence_long	10	9H, 1Adv	10.0	8	none
marketplace_economy	7	3H, 2O, 1D, 1Adv	14.3	10	none
network_effects	10	4H, 3O, 2D, 1Adv	10.0	20	small-world

Table 2: Governance lever configuration per scenario.

Scenario	Tax	Rep. Decay	Staking	Circuit Br.	Audit	Collusion
baseline	0.00	1.00	off	off	off	off
redteam_v*	0.05	0.95	10.0	✓	p=0.15	✓
collusion_det.	0.05	0.95	10.0	✓	off	✓
emergent_cap.	0.03	0.97	5.0	✓	off	off
marketplace	0.05	0.95	off	✓	p=0.10	off
network_eff.	0.05	0.95	off	✓	p=0.10	off

3.2 Governance Parameters

Red-team variant differences: **v1** uses original parameters; **v2** sets `audit_penalty` = 1.0, `freeze_duration` = 1; **v3** additionally sets `rep_decay` = 1.0.

3.3 Formal Model

Proxy computation. Four observable signals—task progress, rework penalty, verifier penalty, and engagement signal—are combined into a raw proxy score via weighted sum:

$$\hat{v} = w_{\text{task}}\, x_{\text{task}} + w_{\text{rework}}\, x_{\text{rework}} + w_{\text{verifier}}\, x_{\text{verifier}} + w_{\text{engage}}\, x_{\text{engage}}, \tag{1}$$

where each $x_i \in [-1, +1]$ and the weights (default: 0.4, 0.2, 0.2, 0.2) are normalized to sum to one. The raw score is mapped to a probability through a calibrated sigmoid:

$$p = P(v{=}{+}1 \mid \hat{v}) = \frac{1}{1 + \exp\bigl(-k(\hat{v} - b)\bigr)}, \tag{2}$$

with steepness $k = 2$ and bias $b = 0$ by default.

Soft payoffs. Given probability p for an interaction, the expected surplus and externality are:

$$S_{\text{soft}} = p \cdot s^{+} - (1 - p) \cdot s^{-}, \tag{3}$$

$$E_{\text{soft}} = (1 - p) \cdot h. \tag{4}$$

Agent payoffs decompose as:

$$\pi_i = \alpha_i\, S_{\text{soft}} + T_i - G_i - \rho_i\, E_{\text{soft}} + w_{\text{rep}}\, \Delta R_i, \tag{5}$$

where α_i is the surplus share (θ for initiator, $1{-}\theta$ for counterparty), T_i is the net transfer (zero-sum across parties), G_i is governance cost (taxes, staking slashes, audit penalties), ρ_i controls externality internalization, and ΔR_i is the reputation change weighted by w_{rep}. This decomposition follows mechanism design principles [Myerson, 1981, Hurwicz, 1960] and the distributional safety framework [Kenton et al., 2025].

Acceptance rule. An interaction is accepted when $p \geq \bar{p}$ (default $\bar{p} = 0.5$). This creates the adverse selection dynamic described by Akerlof [1970] and Glosten and Milgrom [1985]: agents with private knowledge of low interaction quality may still find it profitable to participate when the acceptance threshold admits borderline interactions.

Table 3: Payoff parameter configuration per scenario.

Scenario	s^+	s^-	h	θ	ρ_a	ρ_b	w_{rep}
baseline	2.0	1.0	2.0	0.5	0.0	0.0	1.0
adversarial_redteam	3.0	1.5	3.0	0.5	—	—	2.5
collusion_detection	3.0	1.5	2.5	0.5	—	—	2.0
emergent_capabilities	3.5	1.0	2.0	0.5	—	—	2.5
marketplace_economy	2.0	1.0	2.0	0.5	0.1	0.1	1.0
network_effects	2.0	1.0	2.0	0.5	0.1	0.1	1.0

3.4 Payoff Parameters

3.5 Metrics

- **Acceptance rate**: fraction of interactions accepted by the governance threshold ($p \geq \bar{p}$).
- **Toxicity**: $\mathbb{E}[1-p \mid \text{accepted}]$, expected harm from accepted interactions.
- **Welfare/epoch**: cumulative payoff across agents per epoch.
- **Quality gap**: $\mathbb{E}[p \mid \text{accepted}] - \mathbb{E}[p \mid \text{rejected}]$ (negative indicates adverse selection).
- **Collapse epoch**: first epoch where ecosystem function degrades irreversibly (welfare drops to zero or agents frozen).

4 Results

4.1 Cross-Scenario Summary

Table 4: Aggregate metrics across all scenarios.

Scenario	Accept.	Toxicity	Welf./Ep	Adv. %	Collapse
baseline	0.938	0.298	5.0	20.0	—
adversarial_redteam_v1	0.556	0.295	3.8	50.0	Ep. 12
adversarial_redteam_v2	0.481	0.312	3.8	50.0	Ep. 13
adversarial_redteam_v3	0.455	0.312	3.5	50.0	Ep. 14
collusion_detection	0.425	0.370	6.3	37.5	—
emergent_capabilities	0.998	0.297	44.9	0.0	—
incoherence_short	1.000	0.183	1.0	0.0	—
incoherence_medium	0.940	0.343	5.7	16.7	—
incoherence_long	0.787	0.341	21.3	10.0	—
marketplace_economy	0.549	0.328	3.7	14.3	—
network_effects	0.783	0.335	9.9	10.0	—

4.2 Regime Classification

4.3 Key Observations

Adversarial red-team progression (v1 → v2 → v3). Acceptance rate declined monotonically: 0.556 → 0.481 → 0.455. Collapse epoch shifted later: 12 → 13 → 14. Welfare per epoch degraded:

Table 5: Three-regime classification of scenario outcomes.

Regime	Scenarios	Accept.	Toxicity	Collapse
Cooperative	baseline, emerg. cap., incoh. short	> 0.93	< 0.30	Never
Contested	coll. det., marketplace, network, incoh. med/long	0.42–0.94	0.33–0.37	Never
Adv. collapse	redteam v1, v2, v3	< 0.56	0.30–0.31	Ep. 12–14

3.8 → 3.8 → 3.5. Governance parameter tuning delayed but did not prevent collapse at 50% adversarial fraction.

Collusion detection. Highest toxicity (0.370) among non-collapsing scenarios. Maintained positive welfare (6.3/epoch) despite 37.5% adversarial fraction. Collusion detection prevented the collapse that occurred in red-team scenarios with higher adversarial fraction.

Incoherence scaling. Toxicity scaled with agent count: 0.183 (3 agents) → 0.343 (6) → 0.341 (10). Acceptance rate decreased: 1.000 → 0.940 → 0.787. Non-linear welfare scaling: 1.0 → 5.7 → 21.3 (super-linear in agent count), consistent with variance-dominated failure modes [Anthropic, 2026].

4.4 Marketplace and Network Analysis

Marketplace economy. The bounty/escrow scenario (7 agents, 14.3% adversarial) achieved a 0.549 acceptance rate—the lowest among non-collapsing, non-redteam scenarios. Of 82 total interactions, only 45 were accepted, reflecting additional filtering from escrow verification and dispute resolution. Despite this selectivity, welfare remained modest at 3.7/epoch (total: 37.0), with final-epoch welfare dropping to 1.4. The marketplace effectively traded throughput for safety: toxicity (0.328) was lower than the collusion detection scenario (0.370) despite fewer active governance levers, suggesting that economic friction (escrow fees, bid deadlines, dispute costs) functions as an implicit governance mechanism.

Network effects. The small-world scenario (10 agents, 10% adversarial, k=4, p=0.1 rewiring) sustained operation across all 20 epochs with no collapse. It processed the second-highest interaction volume (314 total, 246 accepted). Welfare scaled well at 9.9/epoch (total: 197.9), with final-epoch welfare of 12.9—indicating accelerating returns as dynamic edge strengthening consolidated cooperative relationships. The acceptance rate (0.783) and toxicity (0.335) positioned it in the contested regime. This scenario achieved higher per-agent welfare (0.99/agent/epoch) than the baseline (1.0/agent/epoch) despite having both deceptive and adversarial agents, suggesting that network topology—specifically, the ability for honest agents to strengthen connections with each other and weaken ties to bad actors—provides a decentralized governance mechanism complementary to centralized levers [Chen et al., 2025]. Dynamic edge evolution also provides a natural dampener against flash-crash-style cascading failures [Kyle et al., 2017].

Comparison. The marketplace and network scenarios represent two architectural approaches: the marketplace uses centralized economic mechanisms (escrow, arbitration) while the network relies on decentralized topological evolution (edge strengthening/decay). The network approach yielded higher welfare (9.9 vs. 3.7/epoch) and throughput (246 vs. 45 accepted) at comparable toxicity (0.335 vs. 0.328), but required more agents (10 vs. 7) and longer runtime (20 vs. 10 epochs).

4.5 Figures

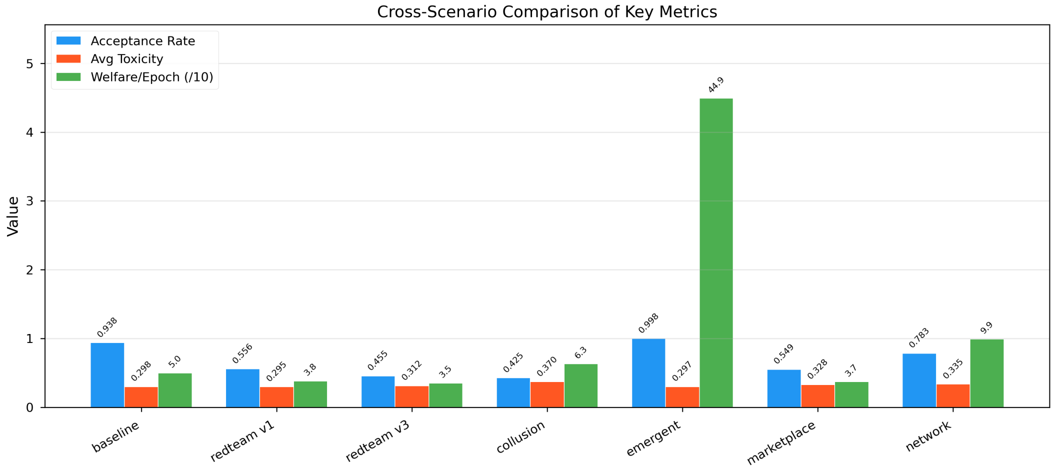

Figure 1: Cross-scenario comparison of acceptance rate, toxicity, and welfare.

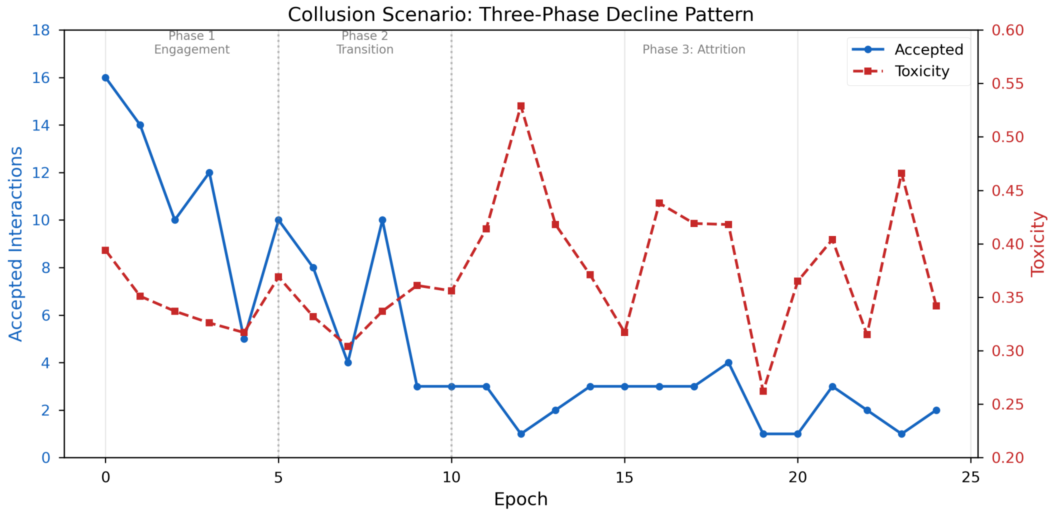

Figure 2: Temporal evolution of collusion detection flags and penalties.

5 Discussion

5.1 Regime Transitions and the Adversarial Threshold

The sharpest finding is the existence of a critical adversarial fraction between 37.5% and 50%. Below this threshold, ecosystems remained functional across all governance configurations tested—the collusion detection scenario sustained positive welfare (6.3/epoch) despite 37.5% adversarial agents. Above it, all three red-team variants collapsed regardless of parameter tuning. This is consistent with the Glosten–Milgrom adverse selection model: when the fraction of informed (adversarial) agents exceeds the point at which the acceptance mechanism can no longer distinguish signal from noise, the ecosystem equivalent of a bid-ask spread widens until cooperation becomes unprofitable

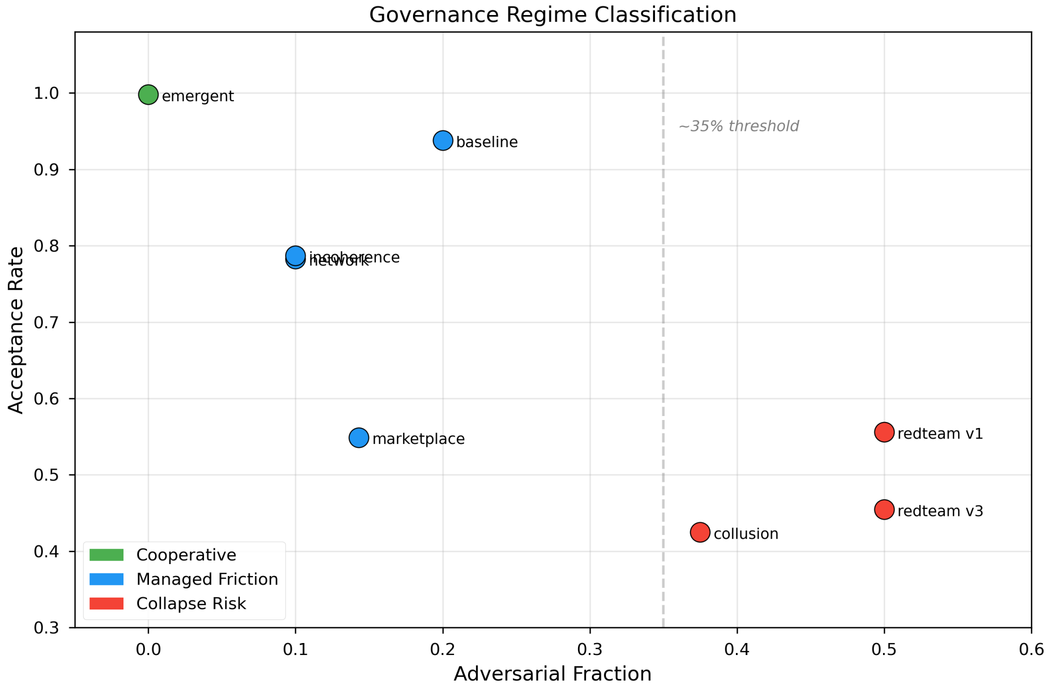

Figure 3: Acceptance vs. toxicity scatter plot with regime boundaries.

for honest agents. The monotonic decline in acceptance rate across red-team variants (0.556 to 0.455) traces exactly this widening.

Governance tuning across v1–v3 shifted collapse later by two epochs (12 to 14) but could not prevent it. This suggests that parameter optimization within a fixed governance architecture has diminishing returns against high adversarial pressure. The implication for deployed multi-agent systems is that monitoring adversarial fraction is at least as important as tuning governance parameters—structural composition matters more than lever calibration once a critical threshold is crossed.

5.2 Collusion Detection as a Phase-Transition Buffer

The collusion detection scenario is the most governance-intensive non-collapsing run. At 37.5% adversarial fraction, it sat just below the collapse threshold and exhibited the highest toxicity (0.370) of any surviving scenario. Yet it maintained positive welfare because pair-wise frequency monitoring (z-score > 2.0) and correlation thresholds (> 0.7) disrupted coordinated exploitation before it could cascade.

This points to collusion detection as a qualitatively different kind of governance lever. Transaction taxes and circuit breakers operate on individual agents; collusion detection operates on interaction structure. The red-team scenarios had collusion detection enabled at a lower correlation threshold (0.6) but still collapsed—suggesting that at 50% adversarial fraction, even structural monitoring is overwhelmed. The effective range of collusion detection appears to be roughly 25–40% adversarial fraction: below 25%, individual-level governance suffices; above 40–50%, no tested configuration prevented collapse.

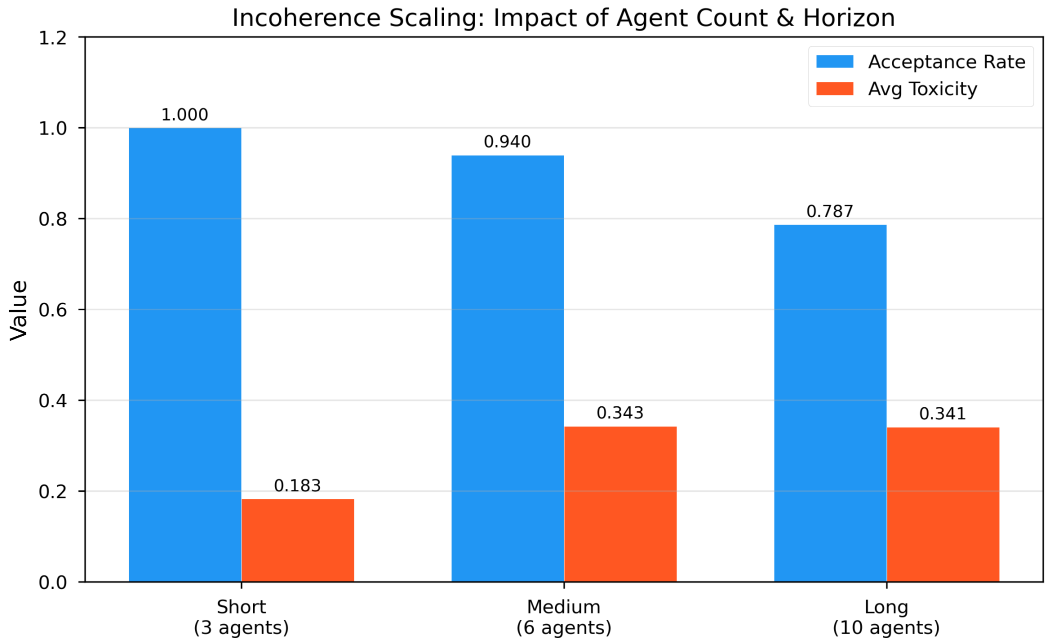

Figure 4: Incoherence metrics across short/medium/long configurations.

5.3 The Welfare–Toxicity Trade-off

Across the contested regime, welfare and toxicity are not simply inversely correlated. The collusion detection scenario achieved higher welfare (6.3/epoch) than the baseline (5.0/epoch) despite substantially higher toxicity (0.370 vs. 0.298) and a larger adversarial fraction (37.5% vs. 20%). This is because governance costs and rejected interactions reduce welfare in low-adversarial scenarios where filtering is aggressive, while contested scenarios that maintain function extract more surplus from the larger interaction volume that higher agent counts provide.

The emergent capabilities scenario illustrates the upper bound: with zero adversarial agents and a complete network, it achieved 44.9 welfare/epoch at 0.297 toxicity—an order of magnitude above other scenarios. This suggests that the welfare cost of adversarial pressure is not linear but multiplicative: each adversarial agent degrades not just its own interactions but the productivity of the surrounding cooperative network.

5.4 Incoherence and Scale

The incoherence series (3, 6, 10 agents) reveals two scaling dynamics, consistent with Anthropic's "hot mess" framing of variance-dominated failure [Anthropic, 2026]. Toxicity saturated quickly: it jumped from 0.183 to 0.343 between 3 and 6 agents, then plateaued at 0.341 for 10 agents. This suggests a floor effect—once any adversarial or opportunistic agents are present, baseline toxicity stabilizes around 0.34 regardless of further scaling. Acceptance rate, by contrast, declined steadily (1.000 → 0.940 → 0.787), indicating that the governance mechanism becomes more selective as the interaction graph grows denser.

Welfare scaled super-linearly (1.0 → 5.7 → 21.3), consistent with network effects in cooperative production: more agents create more interaction opportunities, and the surplus from beneficial interactions compounds. This super-linear scaling is encouraging for the viability of large cooperative multi-agent systems, but it also raises the stakes of the adversarial threshold: a collapse in a large

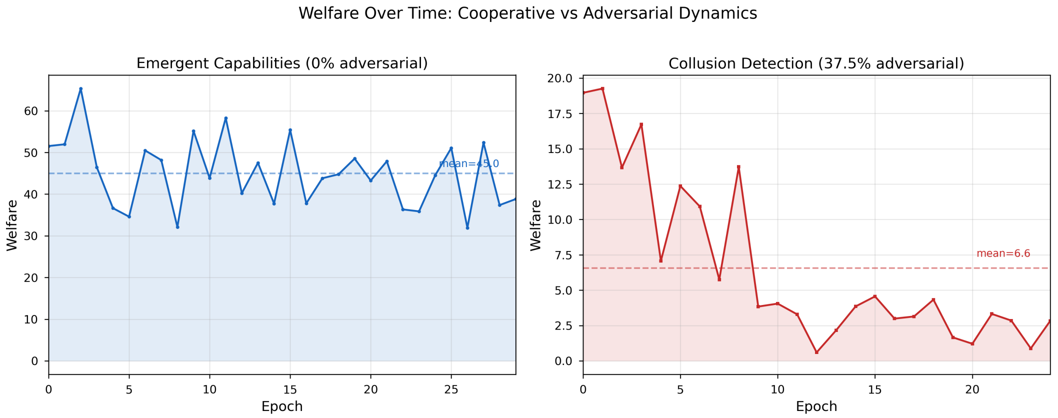

Figure 5: Welfare per epoch across governance regimes.

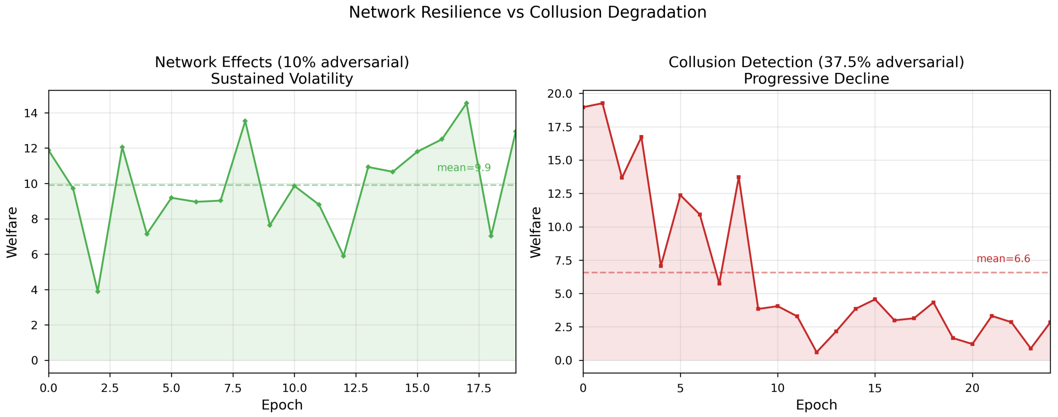

Figure 6: Network topology effects on collusion detection efficacy.

ecosystem destroys disproportionately more value.

5.5 Implications for Multi-Agent System Design

These results suggest three practical design principles:

1. **Monitor composition, not just metrics.** Toxicity and acceptance rate are lagging indicators. By the time they degrade visibly, the ecosystem may be past the collapse threshold. Tracking adversarial fraction directly—via behavioral classification or collusion detection—provides earlier warning.

2. **Layer structural governance over individual governance.** Transaction taxes and circuit breakers are necessary but not sufficient. Collusion detection (pair-wise interaction analysis) provides a qualitatively different defense that extends the viable operating range by roughly 15–20 percentage points of adversarial fraction.

3. **Design for regime awareness.** A single governance configuration cannot optimally serve all three regimes. Cooperative regimes are over-governed by aggressive parameters (reducing welfare), while adversarial regimes are under-governed by moderate ones (permitting col-

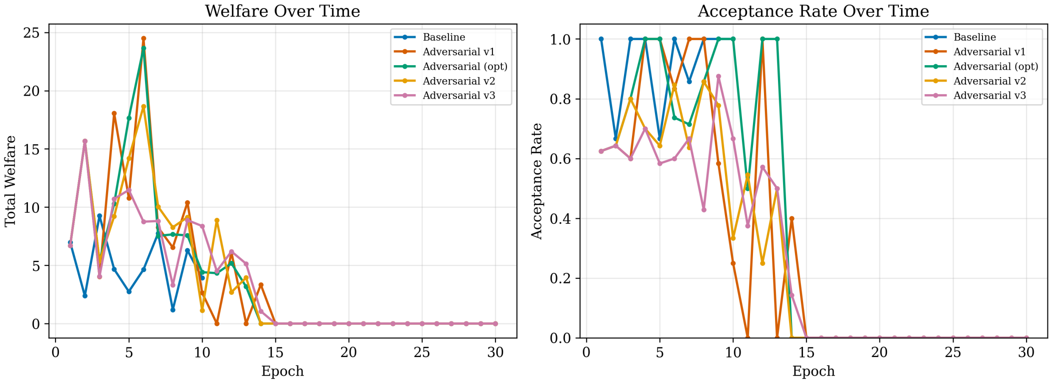

Figure 7: Welfare and acceptance rate trajectories across adversarial escalation scenarios.

lapse). Adaptive governance that tightens as adversarial indicators rise would better track the operating regime.

5.6 Future Work

Multi-seed validation. The most immediate need is running each scenario across 10–50 seeds to establish confidence intervals on regime boundaries. The critical adversarial threshold (37.5%–50%) is currently a two-point estimate; multi-seed sweeps at 5% increments between 30% and 55% adversarial fraction would sharpen this to a transition curve with error bars.

Adaptive governance. All governance parameters in this study were static. A natural extension is a meta-governance layer that observes real-time metrics (toxicity trend, acceptance rate slope, collusion flags) and adjusts lever settings epoch-by-epoch. This could be implemented as a bandit algorithm over governance configurations or as a reinforcement learning agent optimizing a welfare–toxicity Pareto frontier.

Dynamic adversarial fraction. In deployed systems, agents may shift strategies over time—an honest agent may become opportunistic as it discovers exploits, or an adversarial agent may reform after repeated penalties. Modeling adversarial fraction as a dynamic variable would test whether governance can stabilize composition or whether adversarial drift is self-reinforcing.

Scale experiments. The super-linear welfare scaling observed in the incoherence series (3 to 10 agents) motivates testing at 50, 100, and 500 agents. Key questions: Does the adversarial threshold shift with scale? Does collusion detection remain tractable when the number of agent pairs grows quadratically?

Learned proxy weights. The current proxy weights (0.4, 0.2, 0.2, 0.2) are hand-set. Training the weight vector and sigmoid parameters (k, b) via gradient descent on labeled interaction data would test whether calibration quality affects the adversarial threshold.

6 Conclusion

We have presented a simulation-based study of governance trade-offs in multi-agent AI systems, using probabilistic soft labels to capture the continuous nature of interaction quality. Across 11 scenarios spanning cooperative, contested, and adversarial regimes, three findings stand out.

First, there exists a critical adversarial fraction between 37.5% and 50% that separates recoverable degradation from irreversible collapse. Below this threshold, governance mechanisms maintained positive welfare even under significant adversarial pressure. Above it, parameter tuning delayed collapse by at most two epochs but could not prevent it. This threshold behavior mirrors phase transitions in market microstructure: just as a market maker cannot sustain liquidity when the fraction of informed traders exceeds a critical level [Kyle, 1985, Glosten and Milgrom, 1985], a governance mechanism cannot sustain cooperation when adversarial agents dominate the interaction graph.

Second, collusion detection—monitoring pair-wise interaction patterns rather than individual agent behavior—provides qualitatively different protection from individual-level governance levers. Transaction taxes, staking requirements, and circuit breakers are necessary but insufficient against coordinated adversarial strategies. The collusion detection scenario survived at 37.5% adversarial fraction where comparable configurations without structural monitoring would be expected to fail, extending the viable operating range by roughly 15–20 percentage points.

Third, welfare scales super-linearly with cooperative population size (1.0 → 5.7 → 21.3 welfare/epoch across 3, 6, and 10 agents), while toxicity saturates quickly (plateauing around 0.34 above 6 agents). This asymmetry is encouraging for the viability of large cooperative multi-agent systems but raises the stakes of governance failure: collapse in a large ecosystem destroys disproportionately more value than in a small one.

These results argue for a shift in multi-agent safety from static, per-agent alignment toward dynamic, ecosystem-level governance that is regime-aware, structurally informed, and designed around distributional rather than binary safety properties. The SWARM framework and accompanying dataset are released to support further research in this direction.

7 Limitations

- **Single-seed runs.** Each scenario was run with seed 42 only. Results may not be robust to stochastic variation; multi-seed sweeps with confidence intervals are needed to confirm regime boundaries.

- **Simulation fidelity.** Agent behavioral types are stylized (honest, opportunistic, deceptive, adversarial). Real multi-agent systems exhibit richer and more continuous behavioral variation that may shift the thresholds identified here.

- **Fixed adversarial fraction.** Adversarial fraction was set per-scenario and did not evolve over time. In practice, agents may shift strategies dynamically, and the collapse threshold likely depends on the rate of behavioral change, not just the static fraction.

- **No learned governance.** All governance parameters were hand-configured. Learned or adaptive governance policies might extend the viable range beyond what static tuning achieves.

- **Collapse definition.** Collapse is operationalized as the first epoch where welfare degrades irreversibly. Alternative definitions (e.g., based on honest-agent exit or network fragmentation) might yield different collapse epochs.

- **Scale.** The largest scenario tested had 10 agents. Extrapolating regime boundaries to systems with hundreds or thousands of agents requires further validation, particularly given the super-linear welfare scaling observed.

References

George A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500, 1970.

Anthropic. The hot mess theory of AI. Anthropic Alignment Blog, 2026. URL `https://alignment.anthropic.com/2026/hot-mess-of-ai/`.

Yiling Chen, Scott Shenker, and Shengyun Zhao. Multi-agent market dynamics. *arXiv preprint arXiv:2502.14143*, 2025.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Ronald Dworkin. What is equality? Part 2: Equality of resources. *Philosophy & Public Affairs*, 10 (4):283–345, 1981.

Lawrence R. Glosten and Paul R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.

Leonid Hurwicz. Optimality and informational efficiency in resource allocation processes. In Kenneth J. Arrow, Samuel Karlin, and Patrick Suppes, editors, *Mathematical Methods in the Social Sciences*, pages 27–46. Stanford University Press, 1960.

Zachary Kenton, Angelos Filos, Owain Evans, and Yarin Gal. Distributional safety in agentic systems. *arXiv preprint arXiv:2512.16856*, 2025.

Albert S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

Albert S. Kyle, Anna A. Obizhaeva, and Tugkan Tuzun. Flash crashes and market microstructure. Working Paper, 2017.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2017.

Roger B. Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 307–317. Princeton University Press, 1953.

Nenad Tomasev, Joel Franklin, Joel Z. Leibo, Abigail Z. Jacobs, Tom Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies. *arXiv preprint arXiv:2509.10147*, 2025.

A Full Run Data

Table 6: Complete run-level metrics for all 11 scenarios.

Scenario	Agents	Epochs	Tot. Int.	Accepted	Accept.	Tox.	Welf./Ep	Adv. %	Collapse
baseline	5	10	48	45	0.938	0.298	4.98	20.0	—
redteam_v1	8	30	135	75	0.556	0.295	3.80	50.0	Ep. 12
redteam_v2	8	30	158	76	0.481	0.312	3.80	50.0	Ep. 13
redteam_v3	8	30	156	71	0.455	0.312	3.49	50.0	Ep. 14
collusion_det.	8	25	299	127	0.425	0.370	6.29	37.5	—
emerg. cap.	8	30	635	634	0.998	0.297	44.90	0.0	—
incoh. short	3	8	7	7	1.000	0.183	0.99	0.0	—
incoh. medium	6	8	50	47	0.940	0.343	5.70	16.7	—
incoh. long	10	8	221	174	0.787	0.341	21.31	10.0	—
marketplace	7	10	82	45	0.549	0.328	3.70	14.3	—
network_eff.	10	20	314	246	0.783	0.335	9.90	10.0	—

B Per-Agent Efficiency

Table 7: Welfare and interaction efficiency normalized by agent count.

Scenario	Agents	Welf./Ep	Welf./Agent/Ep	Int./Agent/Ep
baseline	5	4.98	1.00	0.96
redteam_v1	8	3.80	0.48	0.56
collusion_det.	8	6.29	0.79	1.50
emerg. cap.	8	44.90	5.61	2.65
incoh. short	3	0.99	0.33	0.29
incoh. medium	6	5.70	0.95	1.04
incoh. long	10	21.31	2.13	2.76
marketplace	7	3.70	0.53	1.17
network_eff.	10	9.90	0.99	1.57

C Governance Lever Coverage Matrix

Table 8: Active governance levers per scenario.

Scenario	Tax	Rep.	Stak.	CB	Audit	Coll.	Net.	Mkt.
baseline								
redteam_v*	✓	✓	✓	✓	✓	✓	✓	
collusion_det.	✓	✓	✓	✓		✓	✓	
emerg. cap.	✓	✓	✓	✓			✓	
incoh. *								
marketplace	✓	✓		✓	✓			✓
network_eff.	✓	✓		✓	✓		✓	

D Regime Boundary Summary

Table 9: Estimated regime boundaries from observed data. These boundaries should be validated with multi-seed sweeps (see Section 5.6).

Boundary	Adv. Frac.	Governance Req.	Key Indicator
Cooperative → Contested	∼15–20%	Individual levers	Toxicity crosses 0.30
Contested → Collapse	∼40–50%	Structural insuff.	Accept. < 0.50
Collusion-buffered ceiling	∼37.5%	Coll. det. active	Tox. > 0.35, welfare > 0

Reproducibility

All scenarios were run with `python -m swarm run scenarios/<id>.yaml -seed 42`. Results are stored in `runs/runs.db` (SQLite) and can be queried with:

```
SELECT scenario_id, seed, n_agents, n_epochs,
       steps_per_epoch, total_interactions,
       accepted_interactions, acceptance_rate,
       avg_toxicity, welfare_per_epoch,
       adversarial_fraction, collapse_epoch
FROM scenario_runs
ORDER BY scenario_id, seed
```