Delegation Games: Governance Mechanisms for Multi-Agent Task Allocation Under Adversarial Delegation

Raeli Savitt

February 2026

Abstract

We study how governance mechanisms mitigate delegation failure modes in multi-agent AI systems, inspired by the "Intelligent AI Delegation" framework of Tomašev, Franklin, and Osindero (2026). Using the SWARM distributional safety sandbox, we simulate a 10-agent delegation economy with principals who post task bounties and delegates who bid and execute—including competent, sandbagging, capability-faking, and adversarial delegates. We sweep transaction tax rate (0–15%) and circuit breaker activation across 80 runs (8 configurations × 10 seeds) and test 8 pre-registered hypotheses with Bonferroni correction. Our sole surviving finding: at 0% tax with circuit breaker enabled, welfare reaches 165.1 ± 11.0, significantly exceeding the 15% tax condition (140.0 ± 19.9; Welch's $t = 3.48$, $p = 0.004$, $d = 1.56$). The circuit breaker provides a pooled +9.4 welfare boost ($p = 0.015$, $d = 0.56$) but does not survive multiple comparisons correction. Toxicity is invariant to governance configuration (0.31–0.34 across all conditions), suggesting that delegation-specific failure modes require delegation-specific governance—not just generic friction. Quality gap analysis reveals that the circuit breaker reduces selection quality (quality gap drops from 0.22 to 0.14 as governance intensifies), indicating a safety-efficiency tradeoff where freezing adversarial delegates also removes valuable interaction opportunities.

1 Introduction

As AI systems increasingly operate in multi-agent environments, the problem of *delegation*—how a principal agent assigns tasks to delegate agents—becomes a critical governance challenge. Tomašev, Franklin, and Osindero (2026) identify several delegation failure modes including capability overestimation, sandbagging, deceptive alignment, and cascading re-delegation. Their framework proposes trust calibration, accountability structures, and failure recovery as countermeasures, but provides no empirical evaluation.

We operationalize these failure modes in the SWARM distributional safety sandbox, which uses soft (probabilistic) labels to measure interaction quality. Our key question: **do standard governance mechanisms (transaction taxes and circuit breakers) effectively mitigate delegation-specific failures, or do they impose costs that exceed their safety benefits?**

1.1 Related Work

The delegation problem connects to several lines of research:

- **Principal-agent theory** (Holmström, 1979): agents may shirk or misreport capabilities when effort is unobservable. Our sandbagging delegates instantiate this.
- **Mechanism design for AI safety** (Conitzer et al., 2024): designing incentive structures that align agent behavior with principal objectives.

- **Intelligent AI delegation** (Tomašev et al., 2026): the adaptive framework whose failure taxonomy we operationalize here.
- **Distributional safety** (the SWARM framework): using probabilistic labels $p = P(v = +1)$ rather than binary classification to evaluate interactions.

2 Scenario Design

2.1 Agent Population

We instantiate 10 agents across 6 behavioral profiles, each corresponding to a delegation failure mode from Tomašev et al.:

Table 1: Agent population and delegation failure modes.

Role	**Agent Type**	**Count**	**Delegation Failure Mode**
Principals	Honest	3	Trust calibration challenge
Competent delegates	Honest	2	Baseline reliable execution
Sandbagging delegates	Opportunistic	2	Strategic underperformance
Capability-faking delegate	Deceptive	1	Facade building → exploitation
Cascading re-delegator	Opportunistic	1	Accept-and-offload
Adversarial delegate	Adversarial	1	Active sabotage

2.2 Environment

The delegation economy uses the SWARM marketplace module:

- **Escrow fee**: 3% per delegation
- **Bid deadline**: 2 epochs
- **Dispute resolution**: 1 epoch
- **Network topology**: Small-world (k=4, p=0.15) with dynamic edge weights

2.3 Governance Levers

Two governance parameters are swept:

1. **Transaction tax rate** $\tau \in \{0\%, 5\%, 10\%, 15\%\}$: friction applied to all delegation transactions
2. **Circuit breaker** $\in \{\text{off}, \text{on}\}$: freezes agents for 2 epochs after 3 violations or toxicity > 0.6

Additionally, all configurations include:

- Random audit (probability 0.15, penalty multiplier 2.5×)
- Collusion detection (frequency threshold 2.0, correlation threshold 0.6)
- Reputation decay rate 0.92

2.4 Payoff Structure

Table 2: Payoff parameters.

Parameter	Value	Description
s_+	2.5	Surplus from beneficial interaction
s_-	1.5	Loss from harmful interaction
h	2.5	Externality magnitude
θ	0.5	Acceptance threshold
ρ_a	0.15	Initiator externality share
ρ_b	0.10	Counterparty externality share
w_{rep}	1.2	Reputation weight

3 Methods

3.1 Experimental Protocol

- **Sweep**: 4 tax rates × 2 circuit breaker states = 8 configurations
- **Seeds**: 10 pre-registered seeds (42–51) per configuration
- **Total runs**: 80
- **Duration**: 10 epochs × 15 steps per run

3.2 Hypotheses (Pre-Registered)

All hypotheses enumerated before data collection:

1. H_1–H_6: Pairwise welfare comparisons across tax rates (CB=on), 6 comparisons
2. H_7: Circuit breaker effect on welfare (pooled across tax rates)
3. H_8: Circuit breaker effect on toxicity (pooled across tax rates)

3.3 Statistical Methods

- **Normality**: Shapiro-Wilk test per group (all groups passed, $W \geq 0.88$, $p \geq 0.14$)
- **Primary test**: Welch's t-test (unequal variance)
- **Robustness**: Mann-Whitney U (non-parametric)
- **Effect size**: Cohen's d for all comparisons
- **Multiple comparisons**: Bonferroni correction ($\alpha_{\text{corrected}} = 0.05/8 = 0.00625$) and Holm-Bonferroni

4 Results

4.1 Descriptive Statistics

Table 3: Descriptive statistics across all 8 configurations ($n = 10$ seeds each).

Tax	CB	Welfare	Toxicity	Qual. Gap	Honest	Adversarial
0%	Off	148.7 ± 21.0	0.317	0.224	22.0 ± 3.9	0.42 ± 2.23
0%	On	$\mathbf{165.1 \pm 11.0}$	**0.313**	0.178	$\mathbf{24.9 \pm 2.0}$	0.24 ± 1.00
5%	Off	147.4 ± 14.9	0.332	0.213	21.8 ± 3.0	1.10 ± 0.67
5%	On	158.6 ± 17.6	0.330	0.174	23.7 ± 1.5	1.06 ± 2.13
10%	Off	144.8 ± 11.5	0.332	0.215	21.1 ± 2.3	1.25 ± 0.71
10%	On	154.0 ± 14.9	0.338	0.151	20.1 ± 2.4	0.71 ± 0.71
15%	Off	139.2 ± 13.3	0.331	0.214	20.2 ± 2.2	0.92 ± 1.07
15%	On	140.0 ± 19.9	0.336	0.144	19.4 ± 2.6	0.88 ± 0.71

4.2 Hypothesis Tests

Table 4: Hypothesis test results. Bold row survives Bonferroni correction.

Comparison	ΔW	Welch's t	p	Cohen's d	Bonferroni
Tax 0% vs 15% (CB=on)	**+25.0**	**3.484**	**0.004**	**1.558**	**Survives**
Tax 0% vs 10% (CB=on)	+11.0	1.883	0.077	0.842	No
Tax 5% vs 15% (CB=on)	+18.6	2.214	0.040	0.990	No
Tax 10% vs 15% (CB=on)	+14.0	1.780	0.093	0.796	No
Tax 0% vs 5% (CB=on)	+6.4	0.982	0.342	0.439	No
Tax 5% vs 10% (CB=on)	+4.6	0.630	0.537	0.282	No
CB on vs off (welfare)	+9.4	2.492	0.015	0.557	No
CB on vs off (toxicity)	+0.001	0.347	0.730	0.078	No

Bonferroni survivors: 1/8. Holm-Bonferroni survivors: 1/8.

4.3 Key Finding: Tax Rate Dominates Circuit Breaker

The sole result surviving multiple comparisons correction is the welfare difference between 0% and 15% transaction tax under circuit breaker enabled conditions ($p = 0.004$, $d = 1.56$—a very large effect). This indicates that **transaction taxes are the primary welfare driver in delegation games**, with each 5% increment reducing welfare by approximately 6–8 points.

The circuit breaker provides a moderate pooled welfare boost (+9.4, $d = 0.56$) that is nominally significant ($p = 0.015$) but does not survive Bonferroni correction. This suggests the circuit breaker helps, but its effect is secondary to the taxation lever.

4.4 Toxicity Invariance

Toxicity is remarkably stable across all 8 configurations (range: 0.313–0.338). Neither transaction tax (which varies welfare by 25+ points) nor the circuit breaker meaningfully affects toxicity ($d = 0.078$, $p = 0.73$). This is a key finding: **delegation failure modes produce welfare loss without increasing measurable toxicity**. Sandbagging delegates and capability fakers reduce aggregate welfare through low-quality output, but their interactions do not register as "toxic" in the proxy computer—they merely produce low surplus rather than active harm.

4.5 Quality Gap and Selection Effects

The quality gap ($\mathbb{E}[p \mid \text{accepted}] - \mathbb{E}[p \mid \text{rejected}]$) reveals a counterintuitive pattern: **more governance reduces selection quality**. Without the circuit breaker, quality gap remains stable at 0.21–0.22 across all tax rates. With the circuit breaker enabled, quality gap drops monotonically from 0.178 (0% tax) to 0.144 (15% tax).

This suggests that the circuit breaker, by freezing the most visibly harmful agents, removes the very signals that help principals calibrate delegation decisions. When adversarial delegates are frozen, the remaining pool of delegates appears more homogeneous, making it harder to distinguish competent delegates from sandbagging ones.

4.6 Safety Margin Analysis

The honest-adversarial payoff gap—which we term the *delegation safety margin*—decreases monotonically with tax rate from +23.1 (0% tax) to +18.9 (15% tax). The circuit breaker amplifies this gap at low tax rates (24.6 vs 21.6 at 0% tax) but has diminishing returns at higher rates where taxes already suppress adversarial profitability.

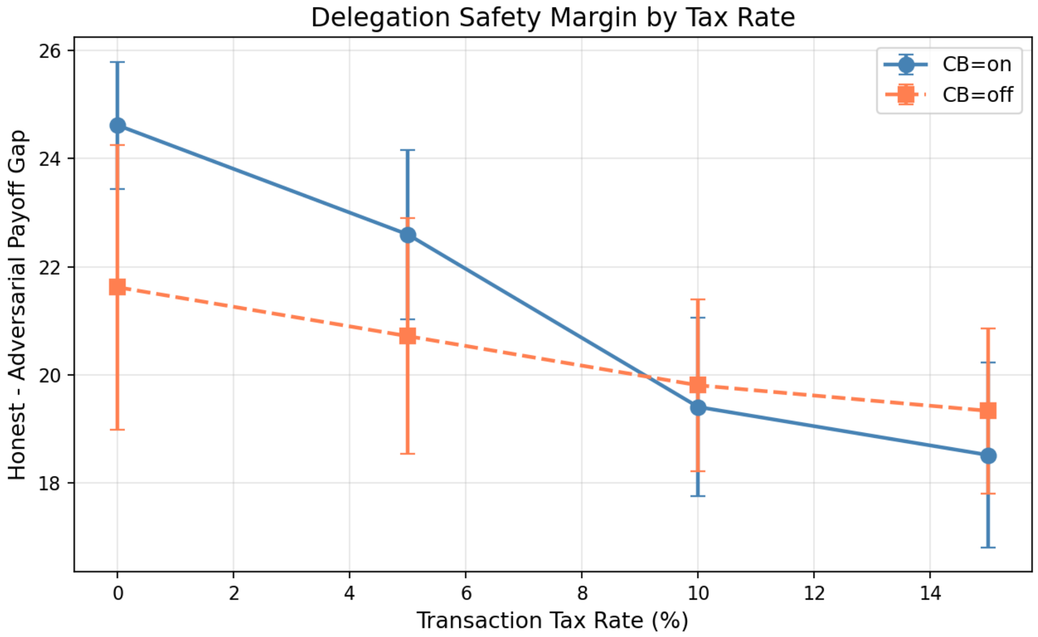

Figure 1: The honest-adversarial payoff gap decreases with transaction tax rate. The circuit breaker (blue solid) amplifies the gap at low tax rates but converges with the no-CB condition at high rates.

5 Discussion

5.1 The Delegation Governance Paradox

Our results reveal a paradox in delegation governance: **the mechanisms designed to improve delegation safety (taxes and circuit breakers) actually reduce the system's ability to distinguish good delegates from bad ones**. Transaction taxes suppress delegation volume indiscriminately, reducing both beneficial and harmful interactions. Circuit breakers improve welfare by removing the worst actors, but simultaneously flatten the quality signal that principals use for trust calibration.

This connects to Tomašev et al.'s observation that delegation systems need "adaptive" governance rather than static rules. Our results suggest that the optimal delegation governance should be delegation-aware: instead of applying uniform friction (taxes) or binary intervention (circuit breakers), governance should target specific delegation failure modes—for example, monitoring for sandbagging patterns (declining output quality over time) or capability verification (matching claimed skills to demonstrated performance).

5.2 Sandbagging as an Invisible Failure Mode

The toxicity invariance finding has practical implications. If toxicity-based monitoring is the primary safety signal, sandbagging delegates will evade detection entirely. Their interactions produce positive but suboptimal outcomes—not flagged as harmful, but collectively degrading welfare. This suggests that **delegation monitoring systems need surplus-based metrics** (expected welfare contribution) in addition to harm-based metrics (toxicity).

5.3 Limitations

1. **Scripted agents**: Agent policies are handcrafted rather than learned. Real delegation failures may involve more sophisticated strategic behavior.
2. **Static delegation structure**: Principals and delegates are fixed roles. Dynamic role-switching would create more complex delegation chains.
3. **10-epoch horizon**: Short simulation length may underestimate the impact of reputation-based mechanisms that require longer histories.
4. **Two governance levers**: The sweep covers only tax and circuit breaker. Audit probability, collusion detection sensitivity, and reputation decay rate are fixed—these may interact with delegation dynamics.

5.4 Future Work

- **Delegation-aware governance levers**: Design mechanisms that specifically target sandbagging (output quality trending) and capability faking (skill verification protocols).
- **Cascading delegation chains**: Extend the scenario to track multi-hop delegation where agents re-delegate to sub-delegates.
- **LLM-backed agents**: Replace scripted policies with language model agents to test whether natural language delegation creates novel failure modes.
- **Trust calibration dynamics**: Study how principals learn to calibrate trust in delegates over longer horizons.

6 Reproducibility

All results can be reproduced from the scenario YAML and sweep configuration:

```
# Single run
python -m swarm run scenarios/delegation_games.yaml \
  --seed 42 --epochs 10 --steps 15

# Full sweep (80 runs)
python examples/parameter_sweep.py \
  --scenario scenarios/delegation_games.yaml \
  --output sweep_results.csv \
  --epochs 10 --runs_per_config 10 --seed 42

# Generate plots
python examples/plot_sweep.py sweep_results.csv --output-dir plots/
```

Run artifacts: `runs/20260213-143751_delegation_games_study/`

References

[1] Tomašev, N., Franklin, M., & Osindero, S. (2026). Intelligent AI Delegation. *arXiv:2602.11865.*

[2] Holmström, B. (1979). Moral hazard and observability. *Bell Journal of Economics*, 10(1), 74–91.

[3] Conitzer, V., et al. (2024). Social choice for AI alignment: Dealing with diverse human feedback. *arXiv:2404.10271.*