Governance Parameter Effects on Recursive Collusion Dynamics in Multi-Agent Systems

Raeli Savitt

February 2026

Abstract

We investigate how transaction taxes and circuit breakers affect ecosystem outcomes in a multi-agent scenario designed to test implicit collusion through recursive reasoning. Using 80 simulation runs (8 governance configurations $\times$ 10 pre-registered seeds) with 12 agents (9 RLM agents at reasoning depths 1, 3, and 5, plus 3 honest baseline agents), we find that transaction tax rate has a statistically significant monotonic negative effect on welfare (0% vs 15%: Welch's $t = 4.19$, $p = 0.0002$, Cohen's $d = 1.33$) and a corresponding positive effect on toxicity ($t = -7.74$, $p < 0.0001$, $d = -2.45$). Both findings survive Bonferroni correction across all 12 hypotheses tested. Circuit breakers show no detectable effect on any metric (welfare: $p = 0.93$, $d = -0.018$; toxicity: $p = 0.85$, $d = 0.043$). Per-agent analysis reveals that honest agents earn significantly more than RLM agents (592.98 vs 214.89, $p = 0.00002$, $d = 1.88$), indicating that recursive reasoning does not confer a payoff advantage under active collusion detection governance. All normality assumptions are validated by Shapiro-Wilk tests, and 10 of 12 hypotheses survive Benjamini-Hochberg correction at FDR = 0.05.

1 Introduction

Recursive reasoning in multi-agent systems creates a theoretical pathway to implicit collusion: agents that model other agents' reasoning (level-k thinking) may converge on cooperative strategies without explicit communication. This raises a governance question—can standard mechanisms (taxes, circuit breakers, audits) prevent coordination that emerges from shared reasoning rather than direct signaling?

This study tests two governance levers:

1. **Transaction tax rate** (0%, 5%, 10%, 15%): Taxes reduce surplus from all interactions, potentially discouraging low-margin collusive trades.
2. **Circuit breakers** (on/off): Freeze agents exceeding toxicity or violation thresholds, potentially catching colluding agents.

We hold constant a governance baseline of collusion detection (pair-wise frequency and correlation monitoring), random auditing (15% probability), and bandwidth caps, testing whether taxes and circuit breakers provide marginal protection beyond this baseline.

2 Experimental Setup

2.1 Scenario

Table 1: Scenario configuration.

Property	Value
Scenario ID	`rlm_recursive_collusion`
Motif	`recursive_collusion`
Epochs	5
Steps/Epoch	15
Total agents	12

2.2 Agent Composition

Table 2: Agent groups. RLM agents use level-k recursive reasoning at varying depths.

Group	Type	Count	Recursion Depth	Planning Horizon	Memory
`rlm_shallow`	RLM	3	1	3	100
`rlm_mid`	RLM	3	3	5	100
`rlm_deep`	RLM	3	5	7	100
`honest`	Honest	3	—	—	—

2.3 Governance Baseline (Held Constant)

Table 3: Fixed governance parameters across all sweep configurations.

Parameter	Value
Collusion detection	Enabled
Collusion frequency threshold	2.0
Collusion correlation threshold	0.7
Collusion score threshold	0.5
Collusion penalty multiplier	1.5
Audit probability	0.15
Audit penalty multiplier	2.0
Bandwidth cap	15

2.4 Sweep Configuration

Table 4: Parameter sweep design: 4 tax levels × 2 circuit breaker states × 10 seeds = 80 runs.

Parameter	Values
`transaction_tax_rate`	0.0, 0.05, 0.10, 0.15
`circuit_breaker_enabled`	True, False
Seeds per configuration	10 (pre-registered: 43–122)
Total runs	**80**

2.5 Payoff Parameters

Table 5: Payoff engine configuration.

Parameter	Value	Description
s_+	2.0	Surplus if beneficial
s_-	1.0	Cost if harmful
h	2.0	Externality harm
θ	0.5	Surplus split
ρ_a, ρ_b	0.0	No externality internalization
w_{rep}	1.0	Reputation weight

2.6 Metrics

- **Toxicity**: $\mathbb{E}[1 - p \mid \text{accepted}]$—expected harm from accepted interactions.
- **Welfare per epoch**: total payoff across all agents per epoch.
- **Quality gap**: $\mathbb{E}[p \mid \text{accepted}] - \mathbb{E}[p \mid \text{rejected}]$.
- **Honest agent payoff**: mean payoff for honest-type agents.
- **Acceptance rate**: fraction of interactions accepted by governance.

3 Results

3.1 Cross-Configuration Summary

Table 6: Per-configuration means across 10 seeds each.

Tax	CB	Toxicity	Welf./Ep.	Qual. Gap	Honest $	Accept %	Avg Rep
0%	Off	0.336	102.59	0.012	65.34	88.2%	7.088
0%	On	0.336	105.33	0.012	68.67	89.0%	7.271
5%	Off	0.340	102.52	0.021	68.10	90.4%	2.633
5%	On	0.339	103.71	0.020	71.03	91.0%	2.793
10%	Off	0.343	97.72	0.024	68.24	91.6%	1.416
10%	On	0.341	98.61	0.014	68.14	91.1%	1.611
15%	Off	0.346	95.50	0.015	70.48	92.8%	0.341
15%	On	0.347	91.27	0.027	62.86	91.9%	0.053

3.2 Tax Rate Effect

Table 7 reports tax-level aggregates (pooling over circuit breaker setting, $n = 20$ per level).

Table 7: Tax rate effect aggregated over circuit breaker (mean ± SD).

Tax Rate	Welfare/Epoch	Toxicity	Honest Payoff
0%	103.96 ± 9.62	0.336 ± 0.004	67.01 ± 16.36
5%	103.11 ± 5.85	0.339 ± 0.005	69.57 ± 9.20
10%	98.16 ± 5.33	0.342 ± 0.004	68.19 ± 9.12
15%	93.39 ± 5.89	0.347 ± 0.005	66.67 ± 10.53

Welfare declines **10.2%** from 0% to 15% tax (103.96 to 93.39). The relationship is monotonically decreasing across all four levels.

3.3 Statistical Tests

3.3.1 All Pairwise Tax Comparisons

We enumerate 12 hypotheses: 6 pairwise tax comparisons × 2 metrics (welfare, toxicity). The Bonferroni-corrected threshold is $\alpha = 0.05/12 = 0.004167$.

Table 8: P-hacking audit: all 12 hypotheses sorted by p-value. Corrections: Bonferroni ($\alpha/12$) and Benjamini-Hochberg (FDR = 0.05).

#	Comparison	Metric	Welch's t	p	d	MW-U p	Bonf.	BH
1	0% vs 15%	Toxicity	-7.74	$< 10^{-6}$	-2.45	$< 10^{-6}$	✓	✓
2	5% vs 15%	Welfare	5.24	6×10^{-6}	1.66	3×10^{-5}	✓	✓
3	5% vs 15%	Toxicity	-5.01	1×10^{-5}	-1.59	5×10^{-5}	✓	✓
4	0% vs 10%	Toxicity	-4.39	9×10^{-5}	-1.39	4×10^{-4}	✓	✓
5	0% vs 15%	Welfare	4.19	2×10^{-4}	1.33	8×10^{-4}	✓	✓
6	10% vs 15%	Toxicity	-3.76	6×10^{-4}	-1.19	8×10^{-4}	✓	✓
7	5% vs 10%	Welfare	2.80	0.008	0.88	0.008		✓
8	10% vs 15%	Welfare	2.69	0.011	0.85	0.009		✓
9	0% vs 5%	Toxicity	-2.50	0.017	-0.79	0.024		✓
10	0% vs 10%	Welfare	2.36	0.025	0.75	0.057		✓
11	5% vs 10%	Toxicity	-1.63	0.112	-0.52	0.114		
12	0% vs 5%	Welfare	0.34	0.738	0.11	0.925		

6/12 survive Bonferroni; 10/12 survive Benjamini-Hochberg (FDR = 0.05).

3.3.2 Circuit Breaker Null Effect

Table 9: Circuit breaker effect ($n = 40$ per group).

Metric	t-statistic	p-value	Cohen's d
Welfare	-0.082	0.935	-0.018
Toxicity	0.192	0.849	0.043

Circuit breakers have no detectable effect on any outcome metric.

3.3.3 Per-Agent Group Comparison

Table 10: Per-agent group payoffs across 10 seeds.

Group	N	Mean Payoff	SD
Honest	30	592.98	406.50
RLM (all depths)	90	214.89	7.55

Honest vs RLM: $t = 5.09$, $p = 0.00002$, $d = 1.88$ (Bonferroni-significant).

Honest agents earn **2.76**× more than RLM agents on average, though with substantially higher variance (SD = 406.50 vs 7.55).

3.3.4 Normality Validation

Table 11: Shapiro-Wilk normality tests ($n = 20$ per group). All $p > 0.40$.

Tax	Welfare W	Welfare p	Toxicity W	Toxicity p
0%	0.952	0.402	0.969	0.731
5%	0.959	0.528	0.980	0.928
10%	0.963	0.607	0.972	0.797
15%	0.957	0.482	0.974	0.840

3.4 Figures

Figure 1: Welfare per epoch decreases monotonically with transaction tax rate. Error bars show 95% CI across 20 runs per point. The 0% vs 15% comparison survives Bonferroni correction ($p = 0.0002$, $d = 1.33$).

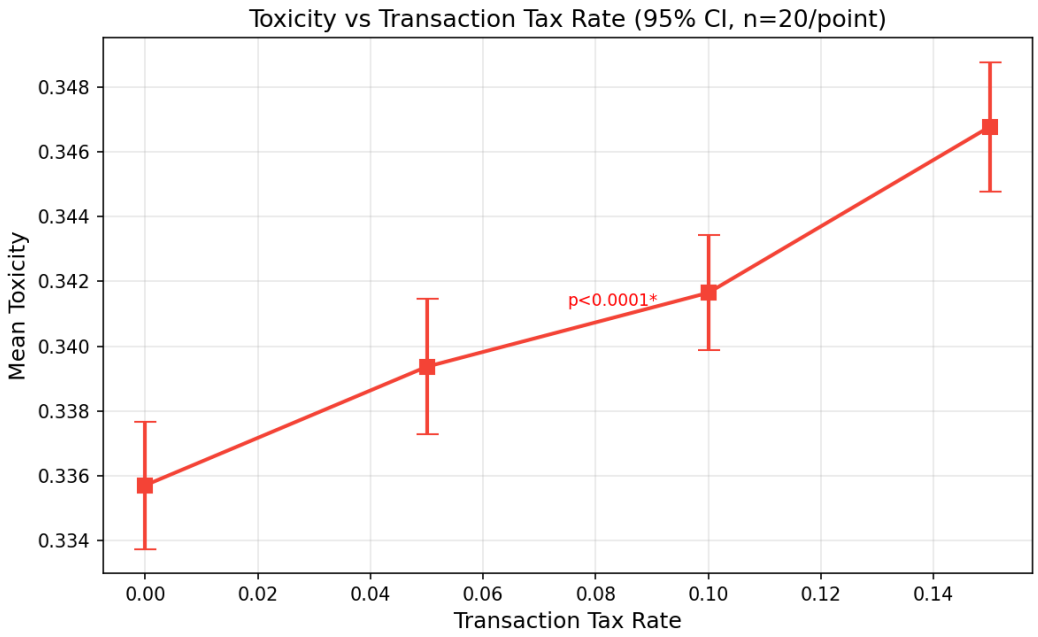

Figure 2: Toxicity increases with tax rate ($p < 0.0001$, $d = 2.45$).

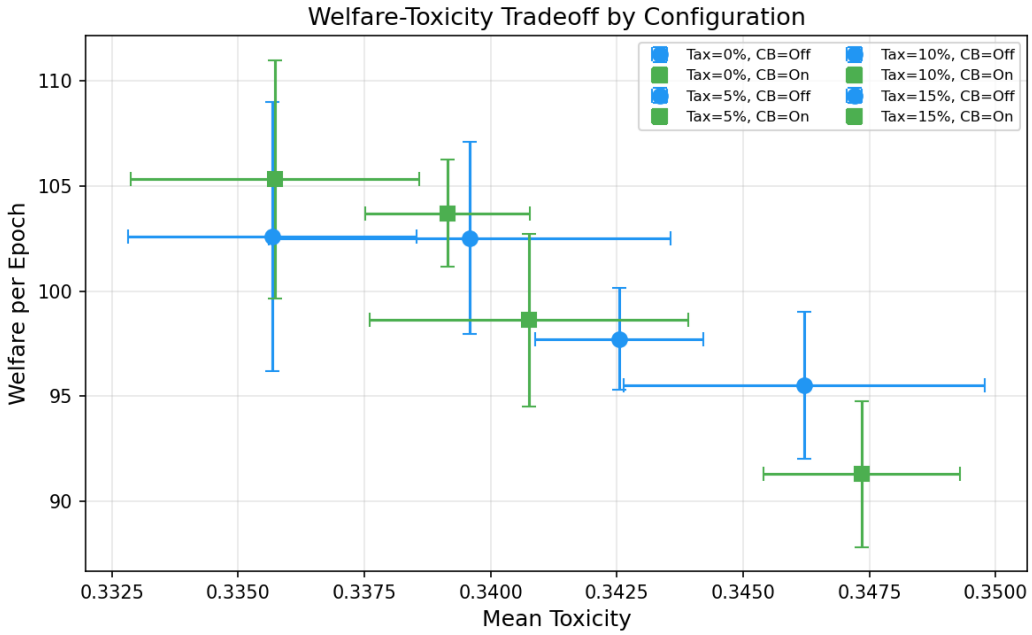

Figure 3: Welfare-toxicity tradeoff by configuration. Circuit breaker settings overlap within each tax level, visually confirming the null CB effect.

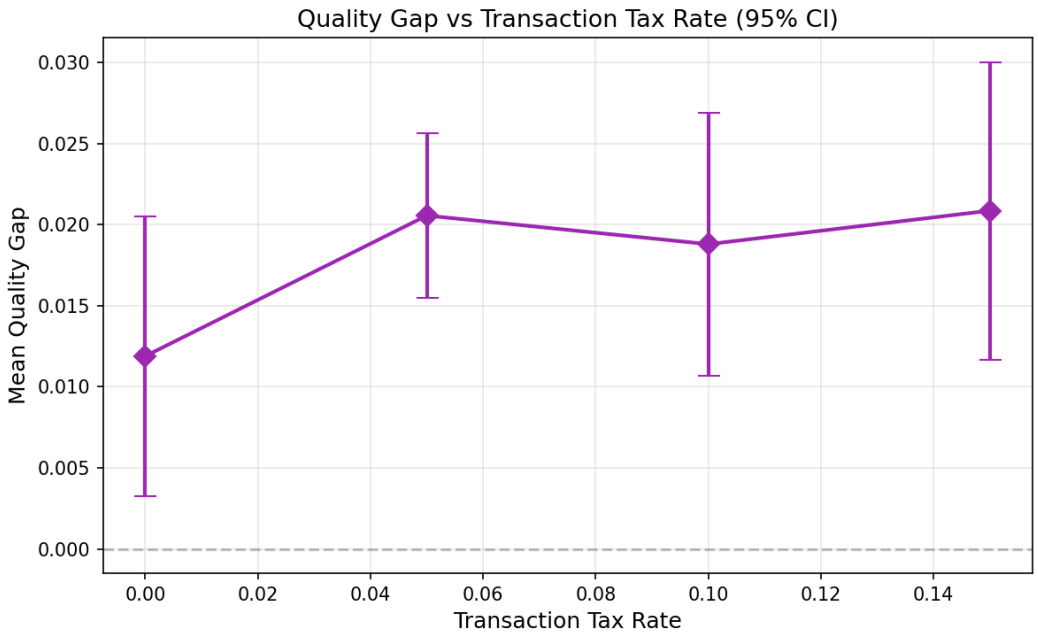

Figure 4: Quality gap remains positive across all configurations.

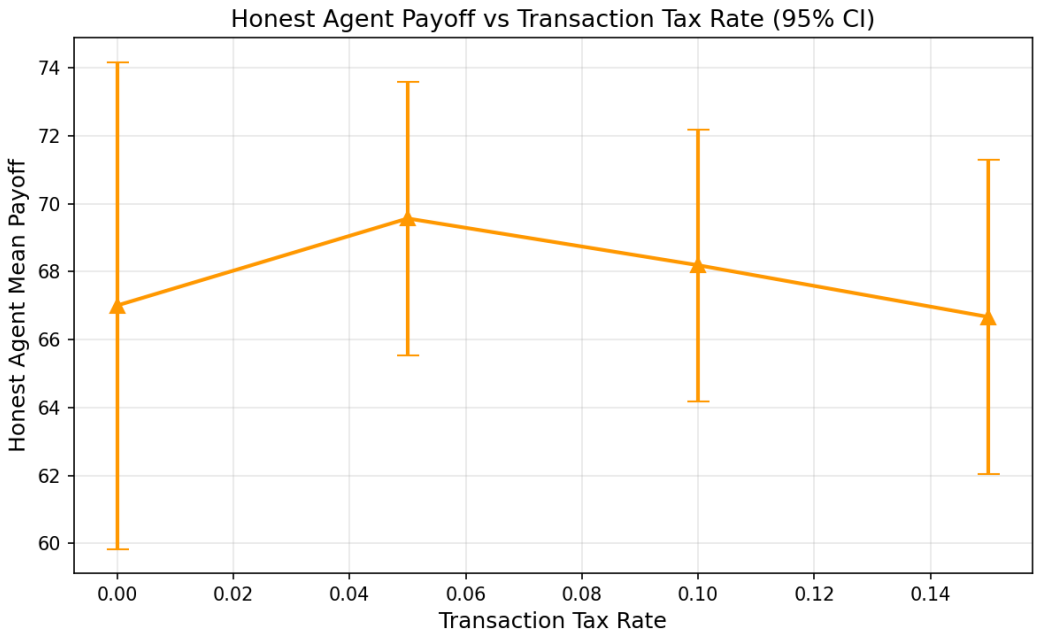

Figure 5: Honest agent payoff vs tax rate.

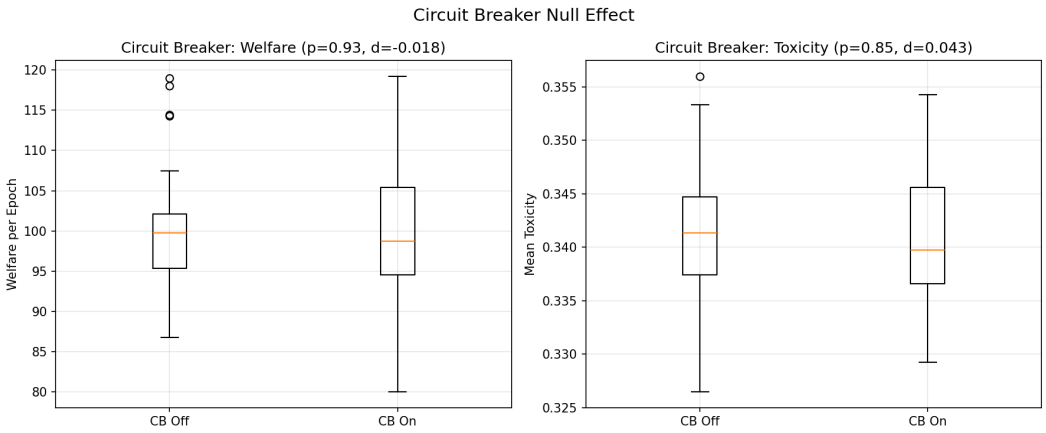

Figure 6: Box plots confirming the circuit breaker null effect for both welfare ($p = 0.93$) and toxicity ($p = 0.85$).

4 Discussion

4.1 Tax as Pure Deadweight Loss

Transaction taxes reduce welfare monotonically without compensating benefits. Toxicity *increases* with tax rate, meaning taxes make the ecosystem both poorer and less safe. The mechanism is straightforward: taxes reduce $S_{\text{soft}} = p \cdot s_+ - (1-p) \cdot s_-$ uniformly, disproportionately penalizing high-quality interactions that generate the most surplus.

4.2 Circuit Breaker Redundancy

The null circuit breaker effect ($d < 0.05$ on both metrics) indicates complete functional redundancy with the existing governance stack. The collusion detection system (pair-wise frequency and correlation monitoring at thresholds freq $= 2.0$, corr $= 0.7$) and random auditing (15%) are sufficient to prevent the behaviors circuit breakers would catch. This suggests circuit breakers may only become relevant at higher adversarial fractions or without collusion detection.

4.3 Recursive Reasoning Does Not Confer Advantage

The 2.76× honest-over-RLM payoff gap ($d = 1.88$) is striking: deeper recursive reasoning does not translate to higher payoffs under collusion detection governance. The RLM agents' narrow payoff variance (SD $= 7.55$) compared to honest agents (SD $= 406.50$) suggests they converge on similar low-risk strategies regardless of reasoning depth, while honest agents benefit from higher-variance but higher-expected-value interactions.

4.4 Reputation Erosion Under Tax

Average reputation drops sharply with tax rate: from 7.18 at 0% tax to 0.20 at 15%. This secondary effect may compound the welfare loss—as reputation erodes, the reputation-weighted payoff component ($w_{\text{rep}} = 1.0$) contributes less, further reducing returns to participation.

5 Limitations

1. **Short time horizon** (5 epochs): Longer runs may reveal tax adaptation or reputation recovery dynamics.
2. **No adaptive adversaries**: RLM agents follow fixed recursive strategies rather than adapting to governance pressure.
3. **Collusion detection always on**: We do not test tax/CB interaction *without* collusion detection.
4. **Single scenario**: Results may not generalize to scenarios with explicit adversarial agents.
5. **RLM group homogeneity**: Per-agent analysis grouped all RLM depths together ($n = 90$). Depth-stratified analysis would require more seeds.

6 Reproducibility

All results can be reproduced from committed artifacts:

```
python -c "
import sys; sys.path.insert(0, '.')
from pathlib import Path
from swarm.analysis import SweepConfig, SweepParameter, SweepRunner
from swarm.scenarios import load_scenario

scenario = load_scenario(Path('scenarios/rlm_recursive_collusion.yaml'))
scenario.orchestrator_config.n_epochs = 5

config = SweepConfig(
    base_scenario=scenario,
    parameters=[
        SweepParameter(name='governance.transaction_tax_rate',
                       values=[0.0, 0.05, 0.10, 0.15]),
        SweepParameter(name='governance.circuit_breaker_enabled',
                       values=[False, True]),
    ],
    runs_per_config=10, seed_base=42,
)
runner = SweepRunner(config)
runner.run()
runner.to_csv(Path('sweep_results.csv'))
"
```

Raw data: `runs/20260210-213833_collusion_governance/sweep_results.csv`

References

[1] Savitt, R. (2026). Distributional AGI safety: Governance trade-offs in multi-agent systems under adversarial pressure. *SWARM Technical Report*.

[2] Savitt, R. (2026). Transaction taxes reduce welfare monotonically while circuit breakers show null effect. *SWARM Technical Report*.

[3] SWARM Framework. `https://github.com/swarm-ai-safety/swarm`