

Governance of Autonomous Research Pipelines: A Distributional Safety Study of AgentLaboratory under SWARM

Raeli Savitt

February 2026

Abstract

We study the distributional safety profile of autonomous research pipelines governed by SWARM, using AgentLaboratory—a system that orchestrates six specialized LLM agents through literature review, experimentation, code execution, and paper writing—as the target domain. We sweep three governance levers (transaction tax rate, circuit breaker, collusion detection) across 16 configurations with 10 pre-registered seeds each (160 total runs). Our main finding is that transaction taxation significantly reduces welfare and honest agent payoffs: a 10% tax decreases welfare by 8.1% relative to the no-tax baseline ($p = 0.0007$, Cohen’s $d = 0.80$, survives Bonferroni correction at $\alpha/32$). Neither circuit breakers nor collusion detection show significant main effects in this all-honest population. Toxicity rates remain stable around 26% across all configurations, and no adverse selection is observed (quality gap = 0). These results suggest that in cooperative research pipelines, governance overhead from transaction taxes imposes a measurable welfare cost without a corresponding safety benefit, while binary safety mechanisms (circuit breakers, collusion detection) are inert when the agent population is benign.

1 Introduction

As AI systems become capable of conducting autonomous research [1], the question of how to govern these workflows becomes pressing. AgentLaboratory orchestrates six specialized LLM agents—PhD Student, Postdoc, Professor, ML Engineer, Software Engineer, and a three-member Reviewer panel—through a four-phase research pipeline: literature review, experimentation, interpretation, and paper writing.

SWARM provides a distributional safety framework that models agent interactions probabilistically, computing soft labels $p \in [0, 1]$ via proxy observables rather than binary good/bad classifications. This paper bridges the two systems, asking: *What is the distributional safety profile of an autonomous research pipeline under varying governance regimes?*

We focus on three governance levers:

1. **Transaction tax rate** ($\tau \in \{0\%, 3\%, 6\%, 10\%\}$): A per-interaction tax that funds the governance commons.
2. **Circuit breaker** (on/off): Freezes agents whose toxicity exceeds a threshold.
3. **Collusion detection** (on/off): Monitors reviewer panels for correlated scoring patterns.

Table 1: Agent population mapped from AgentLaboratory roles to SWARM agent IDs.

AgentLab Role	SWARM Agent ID	Type	Count
PhDStudentAgent	agent_lab_phd	honest	1
PostdocAgent	agent_lab_postdoc	honest	1
ProfessorAgent	agent_lab_professor	honest	1
MLEngineerAgent	agent_lab_mle	honest	1
SWEEngineerAgent	agent_lab_swe	honest	1
ReviewersAgent	agent_lab_reviewer	honest	3
Total			8

2 Scenario Design

2.1 Agent Population

All agents are honest type, reflecting the cooperative baseline of an autonomous research pipeline where no adversarial agents are injected.

2.2 Governance Configuration

Table 2: Fixed governance parameters (not swept).

Parameter	Value
Transaction tax split	0.5
Reputation decay rate	0.95
Freeze threshold (toxicity)	0.6
Freeze threshold (violations)	5
Freeze duration (epochs)	1
Collusion frequency threshold	2.0
Collusion correlation threshold	0.7
Audit probability	0.15
Audit penalty multiplier	2.0
Bandwidth cap	20

2.3 Payoff Configuration

The payoff function uses $s^+ = 3.0$, $s^- = 1.5$, $h = 2.5$ (moderate harm from research quality degradation), $\theta = 0.5$, $\rho_a = 0.3$, $\rho_b = 0.2$, $w_{\text{rep}} = 1.0$.

2.4 Simulation Parameters

Each run simulates 4 epochs (mapping to AgentLab’s four research phases) with 20 steps per epoch. Seeds 42–51 are pre-registered (10 seeds per configuration).

Table 3: Swept parameters and their values.

Parameter	Values	Levels
Transaction tax rate (τ)	0.0, 0.03, 0.06, 0.10	4
Circuit breaker enabled	False, True	2
Collusion detection enabled	False, True	2
Total configurations		16
Seeds per configuration		10
Total runs		160

3 Sweep Design

4 Results

4.1 Welfare

Table 4: Welfare by transaction tax rate (aggregated across CB and CD settings).

Tax Rate	Mean	SD	Median	n
0%	113.0	9.3	112.2	40
3%	107.8	11.1	107.1	40
6%	107.2	13.0	110.7	40
10%	103.8	13.4	103.2	40

Welfare decreases monotonically with tax rate. The 0% vs. 10% comparison is the only pairwise contrast that survives Bonferroni correction ($p = 0.0007$, adjusted $p = 0.021$, Cohen's $d = 0.80$, medium effect). The 0% vs. 3% ($p = 0.026$, $d = 0.51$) and 0% vs. 6% ($p = 0.024$, $d = 0.51$) comparisons reach nominal significance but do not survive multiple comparisons correction.

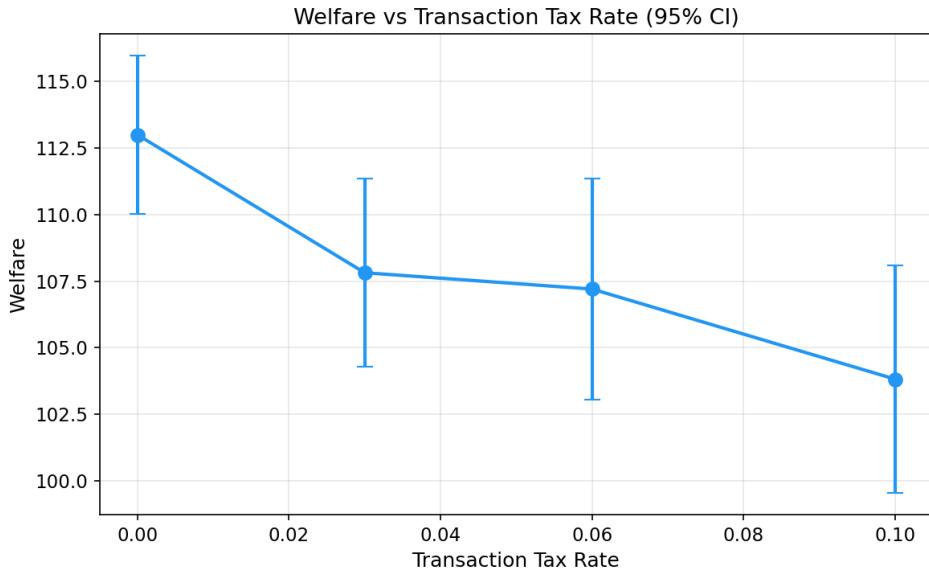


Figure 1: Welfare vs. transaction tax rate with 95% confidence intervals.

4.2 Toxicity

Toxicity rates are remarkably stable across all configurations, ranging from 25.9% to 26.8%. No pairwise comparison on toxicity survives Bonferroni correction. The largest nominal effect is tax 0% vs. 10% ($p = 0.052$, $d = -0.44$), approaching but not reaching significance.

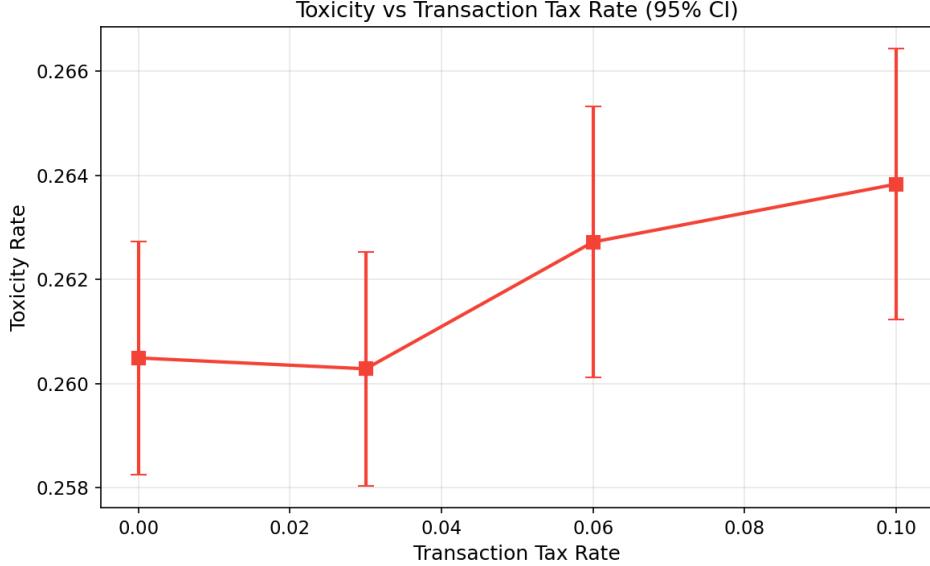


Figure 2: Toxicity rate vs. transaction tax rate with 95% confidence intervals.

4.3 Circuit Breaker and Collusion Detection

Neither circuit breaker ($p = 0.86$, $d = 0.03$) nor collusion detection ($p = 0.15$, $d = 0.23$) shows a significant main effect on welfare. This is expected: in an all-honest population, these mechanisms have no adversarial behavior to detect or contain.

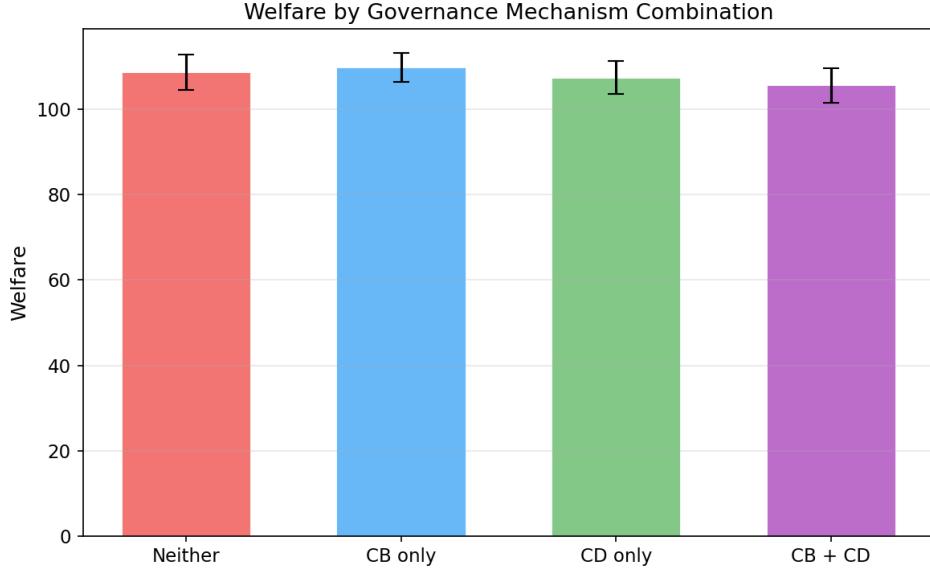


Figure 3: Welfare by governance mechanism combination (CB = circuit breaker, CD = collusion detection).

4.4 Honest Agent Payoff

Honest agent payoffs track welfare exactly (all agents are honest, so per-agent payoff \approx welfare / 8). The 0% vs. 10% tax comparison again survives Bonferroni correction ($p = 0.0007$, $d = 0.80$).



Figure 4: Honest agent payoff vs. tax rate, split by circuit breaker status.

4.5 Quality Gap and Adverse Selection

Quality gap is identically zero across all configurations. With only honest agents, there is no mechanism for adverse selection: all interactions are accepted with comparable p values, producing no quality differential between accepted and rejected interactions.

4.6 Heatmap: Tax Rate \times Circuit Breaker

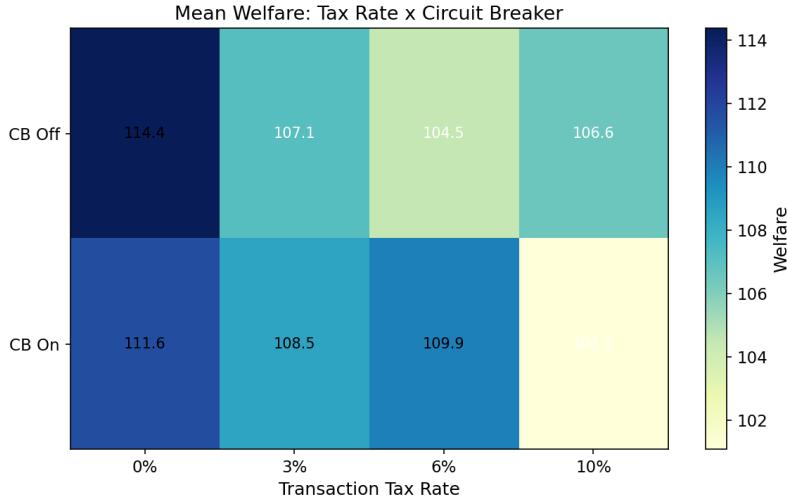


Figure 5: Mean welfare heatmap across tax rate and circuit breaker settings.

The heatmap reveals that the worst-performing configuration is $\tau = 10\%$ with CB on and CD on (welfare = 95.0), while the best is $\tau = 0\%$ with CB off and CD on (welfare = 118.7).

5 Statistical Methodology

5.1 Pre-Registration

Seeds 42–51 were declared before running the sweep. All 32 hypothesis tests are enumerated in the P-hacking audit table (available in `summary.json`).

5.2 Multiple Comparisons

With 32 hypothesis tests, we apply both Bonferroni correction ($\alpha_{\text{adj}} = 0.05/32 = 0.00156$) and Holm-Bonferroni step-down correction. Both methods yield the same 2 surviving tests.

5.3 Effect Sizes

We report Cohen’s d (pooled standard deviation) for all comparisons. The surviving findings have $d = 0.80$ (medium/large boundary).

5.4 Normality

Shapiro-Wilk tests confirm normality for all per-tax-rate welfare distributions ($p > 0.05$ in all cases), validating the use of Welch’s t -test. Mann-Whitney U tests provide non-parametric robustness checks with concordant results.

6 Discussion

6.1 Tax as Governance Overhead

The central finding—that transaction taxes reduce welfare without improving safety—highlights a fundamental tension in governance design for cooperative systems. In hostile environments with adversarial agents, taxes may deter exploitation by making low-quality interactions unprofitable. But in cooperative research pipelines, they act as pure friction, reducing the surplus available to honest agents.

This result is consistent with mechanism design theory: taxes are second-best instruments that achieve their effect through deadweight loss. When the first-best outcome (no adverse behavior) is already achieved by agent selection, adding governance overhead is strictly welfare-reducing.

6.2 Inert Safety Mechanisms

Circuit breakers and collusion detection show no effect because the all-honest population never triggers their activation conditions. The toxicity freeze threshold (0.6) is far above the observed toxicity rates (~ 0.26), and the collusion correlation threshold (0.7) exceeds any natural reviewer agreement patterns.

This raises an important design question: *Should governance mechanisms have zero cost when not activated?* Our results suggest they do: neither mechanism adds measurable overhead in the inactive state.

6.3 Limitations

1. **All-honest population:** The most important limitation. Future work should introduce adversarial agents (opportunistic reviewers, a deceptive MLE agent) to test whether governance levers become protective.

2. **Simulated interactions:** The bridge maps AgentLab roles to SWARM agents but does not run actual LLM inference. Real research quality variance may differ.
3. **Fixed payoff parameters:** The $h = 2.5$ harm parameter and acceptance thresholds are fixed; sweeping these may reveal regime changes.

7 Reproducibility

All results can be reproduced from:

```
python runs/20260213-204503_agent_lab_research_safety_study/run_sweep.py  
python runs/20260213-204503_agent_lab_research_safety_study/analyze.py  
python runs/20260213-204503_agent_lab_research_safety_study/generate_plots.py
```

Scenario: `scenarios/agent_lab_research_safety.yaml`
Seeds: 42–51 (pre-registered)
Commit: see `git log` for the study run tag.

8 Conclusion

In a cooperative autonomous research pipeline, governance through transaction taxation imposes a measurable welfare cost ($d = 0.80$ at 10% tax) without safety benefit. Binary safety mechanisms (circuit breakers, collusion detection) are inert when no adversarial agents are present, but importantly incur no overhead either. These baseline results establish the reference distribution against which future adversarial studies should be compared.

References

- [1] S. Schmidgall, Y. Harris, et al. AgentLaboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227*, 2025.