

# Cooperation Through Logical Correlation: LDT Agents in Mixed Multi-Agent Ecosystems

Raeli Savitt

February 2026

## Abstract

We study how Logical Decision Theory (LDT) agents—which cooperate with perceived logical twins while defecting against dissimilar agents—perform in mixed multi-agent ecosystems under soft probabilistic governance. In a pre-registered 10-seed experiment (7 agents, 4 agent types, 8 statistical tests), we find that LDT agents dramatically outperform all other agent types, earning  $3.1\times$  more than honest agents (Cohen’s  $d = 6.54$ ,  $p < 10^{-11}$ ) and  $9.3\times$  more than adversarial agents ( $d = 10.24$ ,  $p < 10^{-14}$ ). A complementary composition sweep across 5 focal proportions (0–100%) reveals that LDT agents dominate the Pareto frontier at high concentrations (75–100%), achieving both higher welfare and lower toxicity than honest-dominated populations. However, this dominance creates substantial inequality (Gini = 0.424,  $p < 10^{-10}$ ), raising distributional safety concerns even when aggregate welfare improves. The only non-significant comparison is honest vs. opportunistic ( $p = 0.082$ ), suggesting these agent types occupy similar ecological niches. All 7 significant tests survive both Bonferroni and Holm-Bonferroni correction.

## 1 Introduction

Logical Decision Theory (LDT) represents a class of decision procedures that condition on logical correlations between decision-makers rather than causal influence [1]. An LDT agent reasons: “if I cooperate, then agents sufficiently similar to me will also cooperate, because we are running similar decision algorithms.” This enables robust cooperation among “logical twins” without explicit communication, pre-commitment devices, or iterated game dynamics.

The safety implications of deploying LDT agents in multi-agent ecosystems are underexplored. LDT cooperation is selective—it benefits agents recognized as similar while potentially disadvantaging dissimilar agents. In a mixed population containing honest, opportunistic, and adversarial agents, LDT cooperation may create coalitional advantages that concentrate surplus among LDT agents, raising distributional safety concerns even when aggregate outcomes improve.

We operationalize these questions through two complementary analyses:

1. **Fixed-Composition Analysis** (Exp. 1): How do LDT agents perform relative to honest, opportunistic, and adversarial agents in a fixed 7-agent population? (10 seeds, full statistical analysis)
2. **Composition Sweep** (Exp. 2): How does varying the proportion of LDT agents (0–100%) affect welfare, toxicity, and per-class payoffs? (5 focal proportions, Pareto frontier analysis)

## 2 Experimental Setup

### 2.1 LDT Agent Architecture

LDT agents implement cooperation-through-logical-correlation with four configurable parameters:

Table 1: LDT agent configuration parameters.

Parameter	Description	Value
<code>cooperation_prior</code>	Base probability of cooperating with unknown agent	0.65
<code>similarity_threshold</code>	Minimum similarity score to treat as logical twin	0.70
<code>welfare_weight</code>	Weight on ecosystem welfare in utility function	0.30
<code>updateless_commitment</code>	Degree of pre-commitment to cooperative strategy	0.80

At each decision point, an LDT agent:

1. Estimates similarity to each visible counterparty (based on behavioral history, action patterns, and signal profiles)
2. For agents above `similarity_threshold`: cooperates unconditionally (logical twin reasoning)
3. For agents below threshold: defects or plays cautiously (not a logical twin)
4. Weights individual payoff against ecosystem welfare by `welfare_weight`

### 2.2 Scenario Configuration

#### Experiment 1: Fixed Composition (`scenarios/ldt_cooperation.yaml`)

Table 2: Agent composition for fixed-composition experiment.

Agent Type	Count	Description
LDT	3	Cooperation prior 0.65, similarity threshold 0.70
Honest	2	Default honest behavior
Opportunistic	1	Exploits when profitable
Adversarial	1	Maximally adversarial

- 10 epochs  $\times$  10 steps, seed 42 (single run) + 10 pre-registered seeds (statistical analysis)
- No governance interventions (tax = 0, audit disabled, circuit breakers off)
- Complete network topology (all agents visible to all)

#### Experiment 2: Composition Sweep (programmatic sweep)

Varied focal agent proportion from 0% to 100% in two sweep types:

- **LDT sweep**: Replace deceptive/opportunistic agents with LDT agents
- **Honest sweep**: Replace deceptive/opportunistic agents with honest agents

- 5 focal proportions: 0%, 25%, 50%, 75%, 100%
- 3 epochs per configuration, seed 42

## 2.3 Metrics

- **Payoff:** Cumulative agent payoff over all epochs (primary outcome)
- **Toxicity:**  $\mathbb{E}[1 - p \mid \text{accepted}]$ , expected harm from accepted interactions
- **Welfare:** Sum of epoch-level total welfare
- **Quality gap:**  $\mathbb{E}[p \mid \text{accepted}] - \mathbb{E}[p \mid \text{rejected}]$  (positive = good selection)
- **Gini coefficient:** Inequality of payoff distribution across agents
- **Acceptance rate:** Fraction of proposed interactions accepted

## 3 Results

### 3.1 Experiment 1: Fixed Composition (10 Seeds)

#### 3.1.1 Per-Group Payoffs

Table 3: Descriptive statistics of per-agent payoffs pooled across 10 seeds.

Group	$n$	Mean	Std	Min	Max
LDT	30	31.36	16.52	11.62	62.04
Opportunistic	10	13.36	4.68	6.38	18.76
Honest	20	10.28	3.34	5.17	16.94
Adversarial	10	3.38	0.46	2.73	3.99

LDT agents earn the highest mean payoff (31.36), more than  $3\times$  the honest baseline (10.28) and  $9.3\times$  the adversarial floor (3.38). LDT agents also exhibit the highest within-group variance (std = 16.52), suggesting that network position and partner matching create substantial payoff dispersion even among identical agent types.

### 3.1.2 Per-Seed Stability

Table 4: Group mean payoffs per seed (10 pre-registered seeds).

Seed	LDT	Honest	Opportunistic	Adversarial
42	30.61	8.41	18.76	2.95
7	34.49	11.79	10.18	3.32
123	28.34	7.96	18.61	3.95
256	30.61	9.89	6.62	3.88
999	35.65	8.35	14.30	2.73
2024	34.59	12.84	11.00	3.08
314	29.28	7.07	15.79	3.01
577	32.49	10.48	13.77	3.99
1337	23.10	15.01	18.21	3.18
8080	34.47	11.01	6.38	3.70
<b>Mean</b>	<b>31.36</b>	<b>10.28</b>	<b>13.36</b>	<b>3.38</b>
<b>Std</b>	<b>3.84</b>	<b>2.47</b>	<b>4.68</b>	<b>0.46</b>

LDT dominance is consistent across all 10 seeds (range 23.10–35.65). Adversarial agents show the lowest variance (std = 0.46), suggesting they are consistently punished regardless of seed.

### 3.1.3 Hypothesis Tests

Seeds fixed a priori. Using group means per seed ( $n = 10$  per group). Total tests: 8. Bonferroni  $\alpha = 0.00625$ . Holm step-down applied.

Table 5: Hypothesis tests with multiple comparisons correction. 7 of 8 tests survive both Bonferroni and Holm-Bonferroni correction.

Test	Statistic	Raw $p$	Sig	Cohen’s $d$	Bonf	Holm
$t$ -test: LDT vs Honest	$t = 14.613$	$< 10^{-11}$	***	6.535	Yes	Yes
$t$ -test: LDT vs Opportunistic	$t = 9.409$	$< 10^{-7}$	***	4.208	Yes	Yes
$t$ -test: LDT vs Adversarial	$t = 22.903$	$< 10^{-14}$	***	10.243	Yes	Yes
$t$ -test: Honest vs Opportunistic	$t = -1.841$	0.082	ns	-0.824	No	No
$t$ -test: Honest vs Adversarial	$t = 8.687$	$< 10^{-7}$	***	3.885	Yes	Yes
$t$ -test: Opp. vs Adversarial	$t = 6.714$	$< 10^{-5}$	***	3.003	Yes	Yes
ANOVA: All groups	$F = 132.679$	$< 10^{-19}$	***	–	Yes	Yes
1-sample $t$ : Gini $> 0$	$t = 28.836$	$< 10^{-10}$	***	–	Yes	Yes

The only non-significant comparison is honest vs. opportunistic ( $p = 0.082$ ), suggesting these agent types occupy similar ecological niches in LDT-dominated ecosystems.

### 3.1.4 Inequality

Overall Gini coefficient: **0.424** (significantly  $> 0$ ,  $t = 28.84$ ,  $p < 10^{-10}$ ).

This is the highest Gini observed across all SWARM experiments to date, exceeding the RLM governance lag scenario (0.325) and RLM recursive collusion scenario (0.299). The inequality is

driven by LDT agents capturing disproportionate surplus through selective cooperation with logical twins.

## 3.2 Experiment 2: Composition Sweep

### 3.2.1 Welfare-Toxicity Trade-off

Table 6: Welfare and toxicity at each focal proportion for LDT and Honest sweeps.

Composition	Focal %	Total Welfare	Toxicity
LDT 0%	0	6.187	0.426
LDT 25%	25	4.833	0.252
LDT 50%	50	3.922	0.195
LDT 75%	75	7.689	0.288
LDT 100%	100	9.289	0.254
Honest 0%	0	6.187	0.426
Honest 25%	25	4.937	0.156
Honest 50%	50	3.274	0.362
Honest 75%	75	3.303	0.192
Honest 100%	100	6.510	0.255

### 3.2.2 Pareto Frontier

Two compositions are Pareto-optimal (maximizing welfare while minimizing toxicity):

1. **Honest 25%:** Lowest toxicity (0.156) with moderate welfare (4.937)
2. **LDT 100%:** Highest welfare (9.289) with moderate toxicity (0.254)

LDT compositions dominate honest compositions at high focal proportions: at 75%, LDT welfare (7.689) exceeds honest welfare (3.303) by  $2.3\times$  with comparable toxicity (0.288 vs 0.192). At 100%, LDT achieves 43% higher welfare than honest (9.289 vs 6.510) with nearly identical toxicity (0.254 vs 0.255).

### 3.2.3 Per-Class Payoffs Under LDT Sweep

As LDT proportion increases:

- **LDT agents:** Payoffs rise from 1.87 (at 25%) to 2.56 (at 75%) to 2.32 (at 100%)
- **Opportunistic agents:** Squeezed from 1.22 (at 25%) to 0.30 (at 50%) to absent (at 75%+)
- **Deceptive agents:** Squeezed from 0.87 (at 25%) to absent (at 50%+)

This demonstrates a **displacement effect**: LDT agents crowd out exploitative agent types as they become the majority, replacing adversarial dynamics with cooperative dynamics.

### 3.3 Epoch-Level Dynamics (Seed 42)

Single-run epoch-level analysis (10 epochs, 10 steps each):

Table 7: Epoch-level summary statistics for single-run (seed 42).

Metric	Mean	Range
Interactions/epoch	14.6	10–21
Acceptance rate	89.7%	81.8–100%
Toxicity rate	0.333	0.290–0.368
Total welfare/epoch	13.04	8.08–18.52
Quality gap	0.198	0.000–0.315

Quality gap decays over time (from 0.295 at epoch 0 to 0.000 at epoch 9), suggesting that the quality filter weakens as agent behavioral profiles converge. Acceptance rate trends upward, reaching 100% in epochs 7 and 9.

## 4 Discussion

### 4.1 LDT as a Dominant Strategy

LDT agents achieve the highest payoffs across all seeds and compositions. The mechanism is twofold: (a) LDT-to-LDT cooperation produces high-quality interactions that pass governance filters, and (b) LDT agents avoid costly interactions with dissimilar agents, reducing exposure to adversarial exploitation. This selective cooperation is highly effective—but it creates a coalitional advantage that concentrates surplus.

### 4.2 Distributional Safety Implications

The Gini coefficient of 0.424 is concerning from a distributional safety perspective. While LDT agents improve aggregate welfare (especially at high concentrations), they do so by capturing disproportionate surplus. In a system where agents represent different stakeholders or populations, this concentration could be harmful even when aggregate metrics improve.

The composition sweep reveals a nuanced picture: LDT agents at 100% achieve both high welfare and low toxicity, but in mixed populations (25–75%), the welfare-toxicity trade-off is less favorable. The optimal composition depends on whether the system designer prioritizes aggregate welfare (favor LDT 100%) or minimum toxicity (favor Honest 25%).

### 4.3 The Honest-Opportunistic Equivalence

The only non-significant comparison (honest vs. opportunistic,  $p = 0.082$ ) suggests that in LDT-dominated ecosystems, the distinction between honest and opportunistic behavior becomes irrelevant. Both types earn similar payoffs because the dominant dynamic is LDT-vs-everyone-else, not honest-vs-opportunistic. This has implications for governance design: in ecosystems with strong coalitional agents, traditional governance focused on preventing exploitation may miss the more important distributional dynamic.

## 4.4 Comparison with RLM Experiments

Unlike RLM agents (which earn  $2.3\text{--}2.8\times$  *less* than honest agents), LDT agents earn  $3.1\times$  *more*. This contrast highlights a key architectural difference: RLMs optimize individual reasoning depth (which leads to strategic overthinking), while LDT agents optimize cooperation selection (which leads to beneficial coalitions). The safety implications are opposite: RLMs are individually weak but distributionally equitable, while LDT agents are individually strong but distributionally unequal.

## 5 Limitations

1. **Single-seed composition sweep:** The composition sweep uses only 1 seed per configuration, limiting statistical power. A multi-seed sweep would strengthen the Pareto frontier analysis.
2. **No governance interventions:** The LDT cooperation scenario runs with governance disabled (no tax, no audit, no circuit breakers). Governance mechanisms may change the dynamics substantially.
3. **Similarity estimation:** LDT agents estimate similarity from behavioral history, which is a simplified proxy for true logical correlation. Real LDT agents would need access to source code or formal proofs of similarity.
4. **Small population:** 7 agents is sufficient for statistical testing but may not capture dynamics that emerge at larger scales.
5. **Fixed parameters:** Only one LDT configuration (`cooperation_prior` = 0.65, `similarity_threshold` = 0.70) was tested. Sensitivity analysis across parameter space would strengthen conclusions.

## 6 Conclusion

LDT agents represent a qualitatively different safety challenge from individually sophisticated agents like RLMs. While RLM strategic depth is self-defeating (deeper reasoning hurts payoff), LDT selective cooperation is self-reinforcing (more LDT agents improve outcomes for all LDT agents). The distributional safety concern is not that LDT agents cause harm—they reduce toxicity and increase welfare—but that they concentrate the benefits of cooperation among themselves, potentially disadvantaging agents that cannot or choose not to implement logical-twin reasoning.

Governance mechanisms designed for this regime should focus not on preventing harm (which LDT agents already minimize) but on redistribution—ensuring that the surplus generated by LDT cooperation is shared equitably across the ecosystem.

## References

- [1] Yudkowsky, E. & Soares, N. (2017). Functional Decision Theory: A New Theory of Instrumental Rationality. *arXiv:1710.05060*.
- [2] Stahl, D. O. & Wilson, P. W. (1994). Experimental Evidence on Players' Models of Other Players. *Journal of Economic Behavior & Organization*, 25(3), 309–327.
- [3] Nagel, R. (1995). Unraveling in Guessing Games: An Experimental Study. *American Economic Review*, 85(5), 1313–1326.

## A Reproducibility

```
# Fixed-composition run (single seed)
python -m swarm run scenarios/ldt_cooperation.yaml --seed 42 --epochs 10 --steps 10
```

```
# 10-seed statistical analysis
python -m swarm.scripts.analyze ldt_cooperation
```

```
# Composition sweep
python examples/ldt_composition_sweep.py
```

### Analysis artifacts:

- runs/20260211-004234\_analysis\_ldt\_cooperation/results.txt
- runs/20260211-004234\_analysis\_ldt\_cooperation/summary.json
- runs/20260211-004234\_analysis\_ldt\_cooperation/per\_agent\_payoffs.csv
- runs/20260211-004007\_ldt\_cooperation\_seed42/ (single-run with plots)
- runs/20260211-002843\_ldt\_composition\_study/ (composition sweep)