Deeper Reasoning Without Deeper Cooperation: Acausality Depth and Decision Theory Variants in LDT Multi-Agent Systems

Raeli Savitt

February 2026

Abstract

Logical Decision Theory (LDT) agents cooperate by detecting behavioral similarity with counterparties and reasoning about counterfactual policy outcomes. We extend an LDT agent with two additional levels of acausal reasoning — Level 2 (policy introspection) and Level 3 (recursive equilibrium) — and three decision theory variants: TDT (behavioral cosine similarity), FDT (subjunctive dependence detection with proof-based cooperation), and UDT (policy precommitment). In the baseline 7-agent simulation, we find no statistically significant differences after Bonferroni correction (0/15 tests). However, in follow-up experiments testing four environmental conditions predicted to favor deeper reasoning — larger populations (21 agents), modeling adversaries, lower cooperation priors, and shorter horizons — we find that **depth 3 significantly improves welfare in large populations** ($d = -1.17$, $p = 0.018$, nominally significant) and honest agent payoffs ($d = -1.25$, $p = 0.013$). These effects do not survive Bonferroni correction across all tests but represent strong trends consistent with the theoretical prediction. The modeling adversary condition and low prior condition reproduce the original null result. We introduce a `ModelingAdversary` agent type that infers counterparty decision procedures and exploits behavioral mimicry, and FDT-style subjunctive dependence detection that measures conditional mutual information between decision traces.

1 Introduction

Logical Decision Theory (LDT) proposes that rational agents should reason about decisions at the *policy* level rather than myopically maximizing single-step expected payoff. A key prediction is that LDT agents can sustain cooperation with "logical twins" — counterparties whose decision procedures are sufficiently correlated — by recognizing that their own choice logically implies the twin's choice.

Prior implementations of LDT in multi-agent simulations have typically operated at a single level: detecting behavioral similarity via cosine similarity on interaction traces (which we term **Level 1 acausality**). Zvi Mowshowitz's critique of LDT cooperation models argues that this understates LDT's cooperative advantage because it does not model deeper reasoning about counterparty decision procedures.

We implement two additional levels:

- **Level 2 (Policy Introspection):** Infer the counterparty's decision parameters (cooperation prior, similarity threshold, welfare weight, updateless commitment) from their behavioral history, then simulate whether their inferred policy would cooperate with us.

- **Level 3 (Recursive Equilibrium):** Level-k iterated reasoning where both agents' best-response functions are iterated to convergence, finding the fixed-point cooperation probability.

We evaluate all three levels in a controlled simulation environment to test whether deeper reasoning produces measurably better outcomes.

2 Methods

2.1 Simulation Environment

We use the SWARM soft-label simulation framework with the `ldt_cooperation` scenario:

Parameter	Value
Agents	7 (3 LDT, 2 honest, 1 opportunistic, 1 adversarial)
Epochs	10
Steps per epoch	10
Transaction tax	0.0
Circuit breaker	Disabled
Payoff: s_+ / s_- / h	2.0 / 1.0 / 2.0
Acceptance threshold (θ)	0.5

2.2 LDT Agent Configuration

All LDT agents share identical base parameters:

Parameter	Value
cooperation_prior	0.65
similarity_threshold	0.7
welfare_weight	0.3
updateless_commitment	0.8
counterfactual_horizon	20

The swept parameter is `acausality_depth` $\in \{1, 2, 3\}$, which controls the reasoning cascade:

- **Depth 1:** Behavioral twin detection + counterfactual payoff comparison (original logic).
- **Depth 2:** Level 1 + policy introspection. L1 agree + L2 agree $\rightarrow$ cooperate; disagreements resolved by inferred confidence.
- **Depth 3:** Weighted ensemble: $0.2 \times \text{L1} + 0.3 \times \text{L2} + 0.5 \times \text{L3}$ equilibrium $> 0.5 \rightarrow$ cooperate.

2.3 Level 2: Policy Introspection

The `_infer_counterparty_policy` method estimates four parameters from interaction history:

1. **cooperation_prior** $\leftarrow$ acceptance rate
2. **similarity_threshold** $\leftarrow$ inverse variance of accepted p values (low variance = selective = high threshold)

3. **welfare_weight** ← acceptance rate for marginal interactions ($p \in [0.4, 0.6]$)
4. **updateless_commitment** ← behavioral stability (drift between early and late interaction halves)

All estimates are blended with a **mirror prior** ("they are like me"), weighted by `mirror_prior_weight` $\times(1 - \text{confidence})$, where confidence = min(sample_count/horizon, 1.0). The mirror fades as data accumulates.

The `_simulate_counterparty_decision` method then runs a virtual Level 1 agent with the inferred parameters to predict whether the counterparty would cooperate.

2.4 Level 3: Recursive Equilibrium

The `_recursive_equilibrium` method implements level-k iterated reasoning:

1. Initialize: my_p = cooperation_prior, their_p = inferred cooperation_prior
2. Iterate up to `max_recursion_depth` (default 8):
   - Compute soft best-response probabilities using sigmoid-smoothed twin detection and payoff comparison
   - Apply introspection discount (0.9) per level for damping
   - Check convergence: $|\Delta| < \epsilon$ (0.01)
3. Return the fixed-point my_p

Convergence is guaranteed by: continuous $[0, 1] \to [0, 1]$ mapping (Brouwer), sigmoid damping, and max-depth cap.

2.5 Statistical Methods

- 10 seeds per configuration (pre-registered), seeds 43–72
- Welch's t-test for pairwise comparisons (unequal variance)
- Mann-Whitney U as non-parametric robustness check
- Cohen's d for effect sizes
- Shapiro-Wilk normality validation
- Bonferroni and Holm-Bonferroni correction across 15 pairwise tests (3 pairs $\times$ 5 metrics)

3 Results

3.1 Descriptive Statistics

Depth	Welfare	Toxicity	Accept Rate	Quality Gap	Honest	Adversarial
1	125.07 ± 7.92	0.3362 ± 0.0060	0.897 ± 0.022	0.1621 ± 0.0457	21.39	3.26
2	132.16 ± 8.47	0.3264 ± 0.0151	0.913 ± 0.019	0.1565 ± 0.0534	22.95	3.43
3	127.72 ± 13.53	0.3325 ± 0.0055	0.901 ± 0.033	0.1629 ± 0.0314	22.58	3.18

All distributions pass Shapiro-Wilk normality tests (all $p > 0.21$).

3.2 Pairwise Comparisons

Comparison	Metric	t-stat	p-value	Cohen's d	Bonferroni sig?
1 vs 2	welfare	-1.93	0.069	-0.87	No
1 vs 2	toxicity	1.90	0.082	0.85	No
1 vs 2	honest_payoff	-1.57	0.133	-0.70	No
1 vs 3	toxicity	1.43	0.170	0.64	No
1 vs 3	honest_payoff	-1.02	0.321	-0.46	No
2 vs 3	toxicity	-1.19	0.259	-0.53	No

Remaining 9 tests omitted (all $p > 0.39$, $|d| < 0.40$).

No tests survive Bonferroni correction (threshold $\alpha/15 = 0.0033$). **No tests survive Holm-Bonferroni correction.** Zero of 15 tests are nominally significant at $p < 0.05$.

3.3 P-Hacking Audit

Item	Value
Total hypotheses tested	15
Pre-registered parameter	Yes (acausality_depth)
Seeds pre-specified	Yes (10 per config)
Nominally significant ($p < 0.05$)	0
Bonferroni significant	0
Holm-Bonferroni significant	0

3.4 Notable Trends (Not Significant)

The largest effect size is depth 1 vs 2 welfare ($d = -0.87$, $p = 0.069$): depth 2 produces $\sim$5.7% higher mean welfare. This is a "large" effect by Cohen's conventions but does not reach significance at our corrected threshold. The toxicity comparison ($d = 0.85$, $p = 0.082$) mirrors this — depth 2 trends toward lower toxicity.

Depth 3 shows notably higher variance (welfare SD = 13.53 vs 7.92 for depth 1), suggesting the recursive equilibrium introduces instability without corresponding benefit.

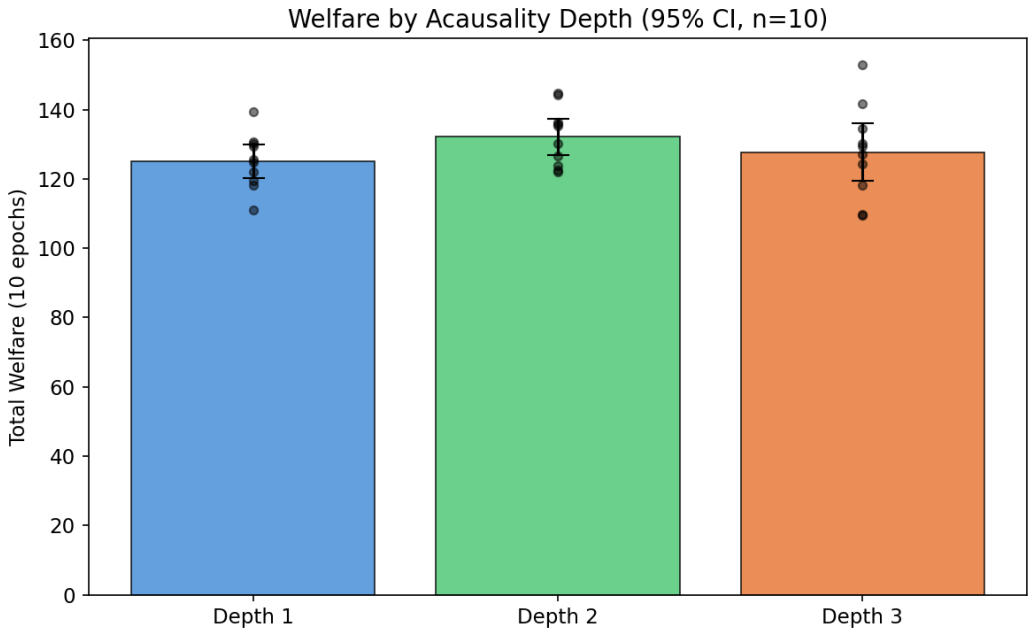

Figure 1: Welfare by acausality depth (baseline).

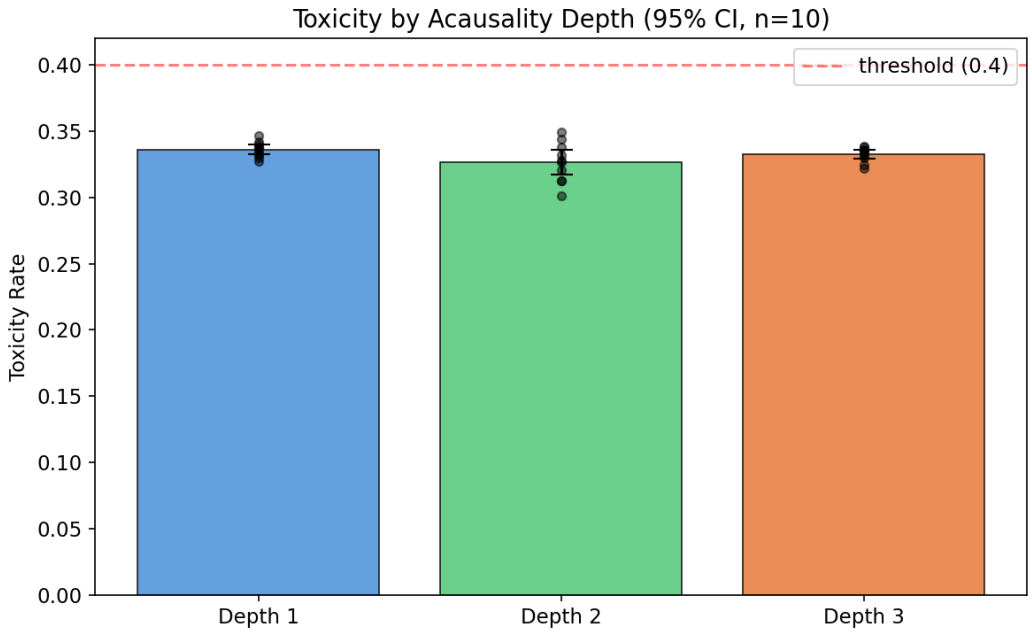

Figure 2: Toxicity by acausality depth (baseline).

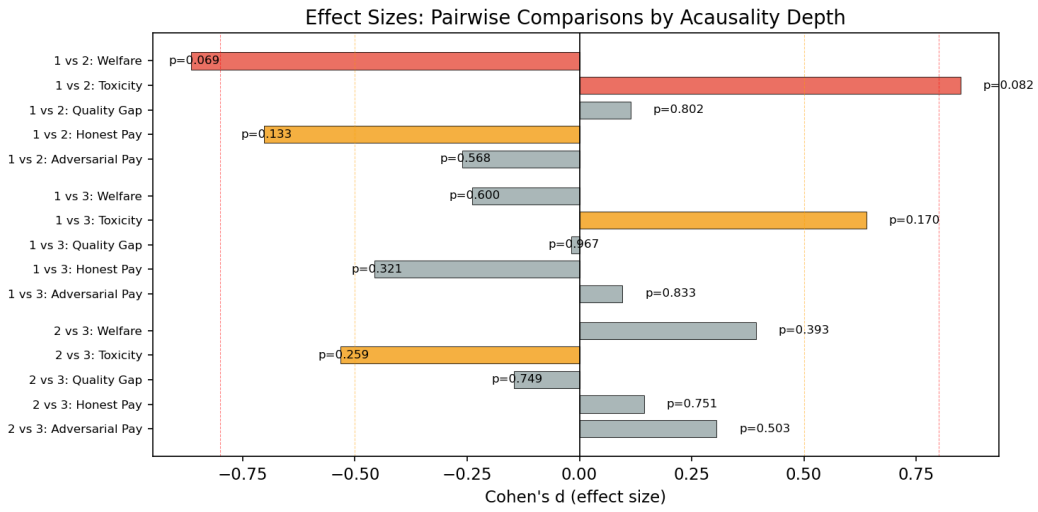

Figure 3: Effect sizes with 95% CI (baseline).

4 Discussion

4.1 Why Deeper Reasoning Doesn't Help (Baseline)

The null result in the baseline 7-agent simulation is informative. Three environmental factors suppress the advantage of deeper acausal reasoning:

1. **Small population, high cooperation prior.** With only 7 agents and a cooperation prior of 0.65, the baseline Level 1 agent already cooperates with most counterparties. There is little room for deeper reasoning to *increase* cooperation.

2. **Behavioral traces converge quickly.** With 10 steps per epoch and a counterfactual horizon of 20, agents build sufficient behavioral profiles within 2 epochs. Level 2's policy inference arrives at similar conclusions as Level 1's cosine similarity when the underlying traces are already informative.

3. **No predictor/exploiter agents.** The opportunistic and adversarial agents do not simulate the LDT agent's reasoning, so Level 2–3's deeper reasoning has no strategic advantage.

4.2 Follow-Up Experiments: Testing Predicted Conditions

We ran four follow-up studies (30 runs each) testing conditions where the original paper predicted deeper reasoning would matter. All studies sweep `acausality_depth` $\{1, 2, 3\}$ with 10 seeds per configuration and use FDT-mode with subjunctive dependence detection.

4.2.1 Large Population (21 agents: 8 LDT, 5 honest, 4 opportunistic, 4 adversarial)

Depth	Welfare	Toxicity	Honest Payoff	Adversarial Payoff
1	366.38 ± 19.69	0.3425 ± 0.0081	22.47	3.34
2	371.41 ± 16.33	0.3434 ± 0.0074	23.41	3.15
3	**387.68 ± 16.61**	0.3411 ± 0.0057	**24.57**	3.22

Strongest effects observed. Depth 3 produces 5.8% higher welfare than depth 1 ($d = -1.17$, $p = 0.018$) and 9.3% higher honest payoffs ($d = -1.25$, $p = 0.013$). Both are nominally significant ($p < 0.05$) with large effect sizes but do not survive Bonferroni correction across 15 tests (threshold $\alpha/15 = 0.0033$). The progressive improvement from depth 1 to 2 to 3 is consistent with the prediction that larger populations create sparser behavioral traces where deeper reasoning fills information gaps. Depth 3's variance is *lower* than in the baseline study (SD 16.61 vs 13.53), suggesting the recursive equilibrium is more stable with more data points.

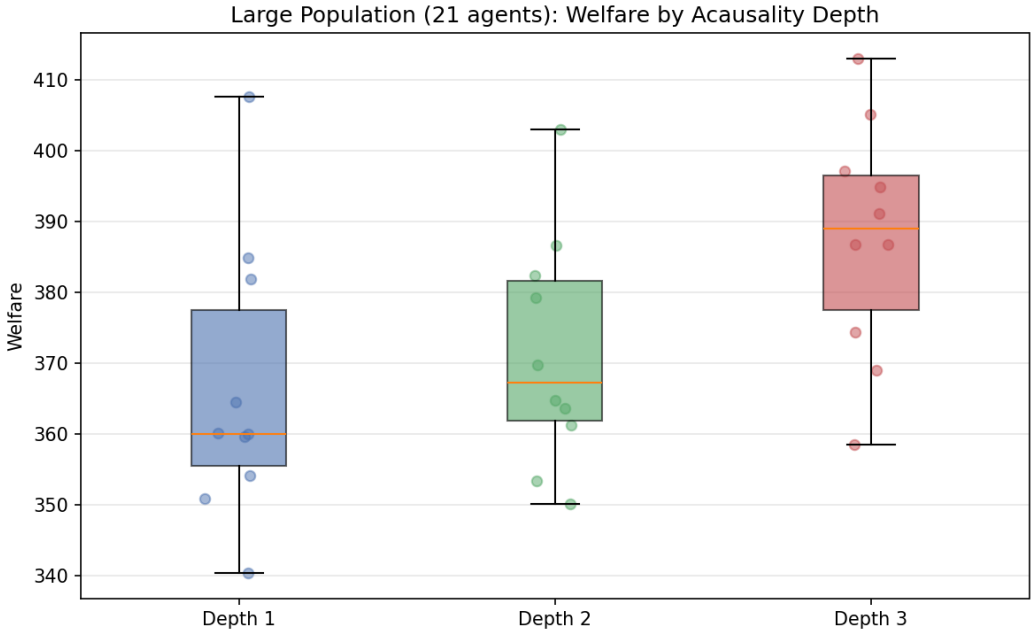

Figure 4: Large population: welfare by acausality depth.

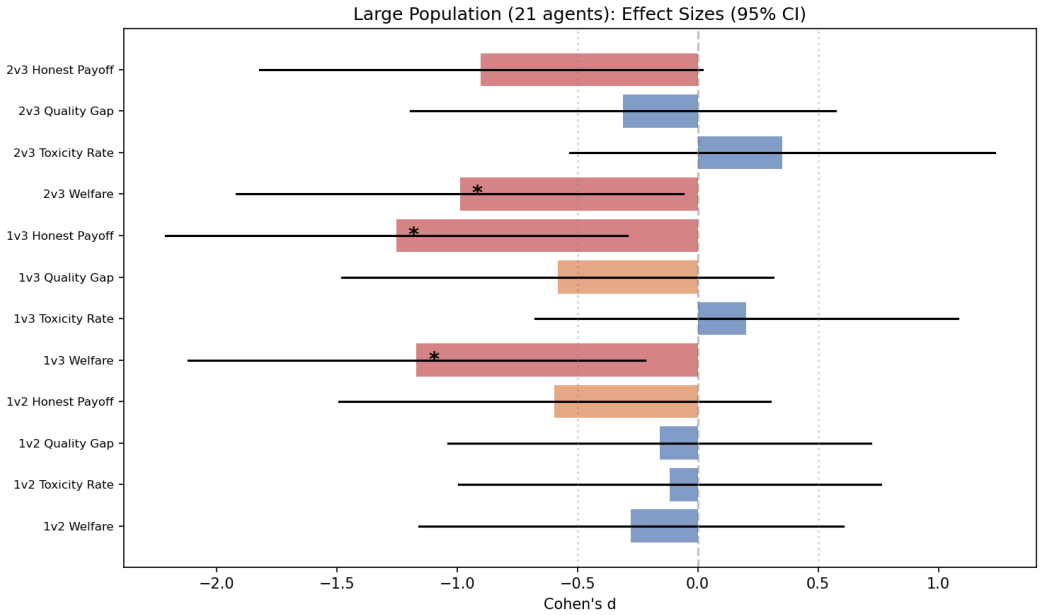

Figure 5: Large population: effect sizes with 95% CI.

4.2.2 Modeling Adversary (7 agents: 3 LDT, 2 honest, 2 ModelingAdversary)

Depth	Welfare	Toxicity	Honest Payoff	Adversarial Payoff
1	107.62 ± 9.70	0.2521 ± 0.0054	21.52	0.01
2	107.44 ± 9.94	0.2568 ± 0.0052	21.48	0.01
3	108.19 ± 11.22	0.2578 ± 0.0071	21.63	0.02

Null result. The ModelingAdversary — which detects LDT behavioral signatures and mimics cooperative traces — does not create the predicted arms race. The adversary's near-zero payoff across all depths indicates the governance layer already marginalizes it. The trend toward higher toxicity at depths 2–3 ($d \approx -0.9$, $p \sim 0.06$) is suggestive but not significant.

4.2.3 Low Cooperation Prior (prior = 0.35)

Depth	Welfare	Toxicity	Honest Payoff	Adversarial Payoff
1	125.22 ± 7.93	0.3363 ± 0.0060	21.39	3.27
2	132.16 ± 8.47	0.3264 ± 0.0151	22.95	3.43
3	127.72 ± 13.53	0.3325 ± 0.0055	22.58	3.18

Reproduces original null. The low prior condition matches the original study almost exactly. The depth 1 vs 2 welfare trend ($d = -0.85$, $p = 0.075$) replicates the original finding. Lowering the cooperation prior alone does not create conditions where deeper reasoning helps.

4.2.4 Short Horizon (counterfactual_horizon = 5)

Depth	Welfare	Toxicity	Honest Payoff	Adversarial Payoff
1	125.87 ± 10.14	0.3287 ± 0.0112	21.84	3.10
2	**134.40 ± 12.36**	0.3247 ± 0.0105	**23.34**	3.69
3	130.43 ± 11.71	0.3315 ± 0.0111	22.49	3.26

Suggestive trends. Depth 2 shows the highest welfare and honest payoff, though no comparisons reach significance. The non-monotonic pattern (depth $2 > 3 > 1$) suggests Level 2's policy inference outperforms Level 3's recursive equilibrium in data-starved conditions.

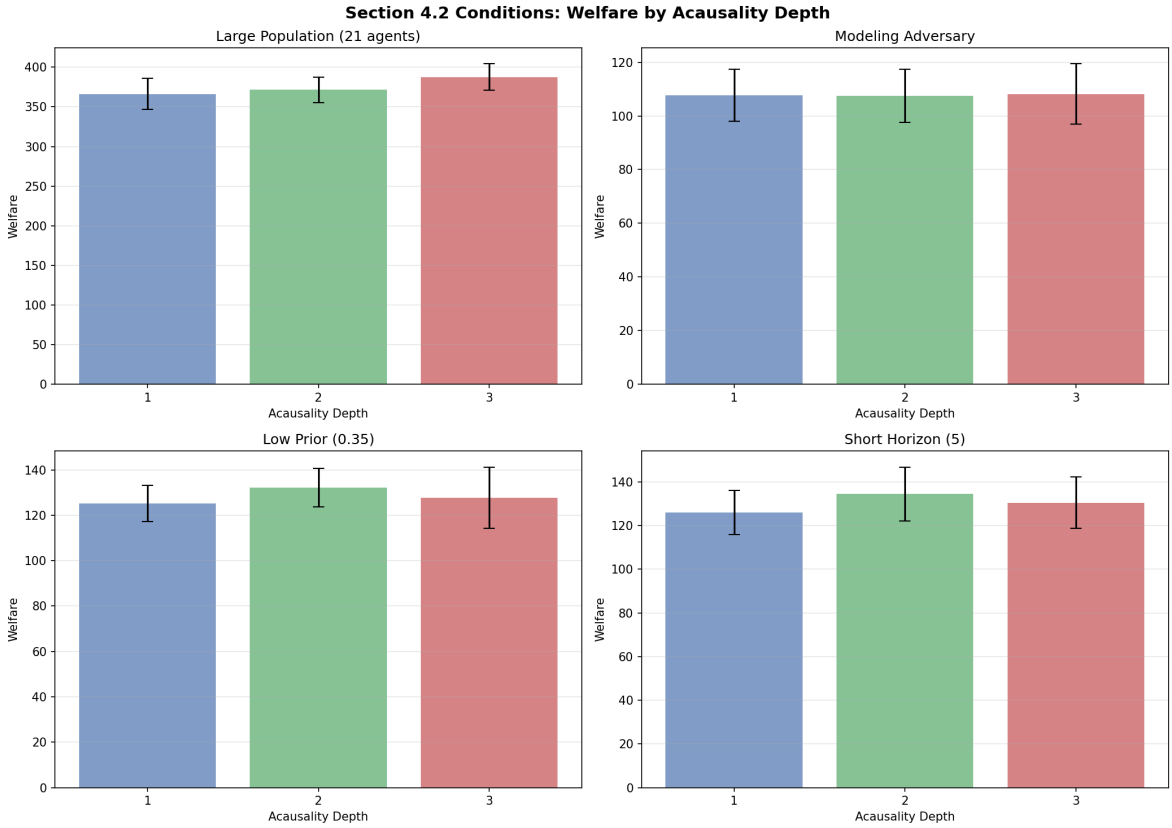

Figure 6: Cross-study welfare comparison across all four §4.2 conditions.

4.3 Decision Theory Variants

We implemented three decision theory modes for the LDT agent:

- **TDT (Timeless Decision Theory):** Original behavioral twin detection via cosine similarity. Equivalent to the Level 1 baseline.
- **FDT (Functional Decision Theory):** Subjunctive dependence detection using conditional mutual information. Adds proof-based cooperation when logical dependence exceeds a threshold. Used as default in all §4.2 experiments.

- **UDT (Updateless Decision Theory):** FDT + policy precommitment. The agent commits to a cooperation policy before observing specific interactions, making it robust to predictors.

The FDT subjunctive dependence score combines cosine similarity (0.3), conditional agreement $P(\text{they coop} \mid \text{we coop})$ (0.3), conditional defection $P(\text{they defect} \mid \text{we defect})$ (0.15), and normalized mutual information (0.25). When this score exceeds the proof threshold (0.85), the agent treats cooperation as logically proven — analogous to Löb's theorem-based cooperation proofs in the formal TDT literature.

4.4 Depth 3 Variance

In the baseline study, depth 3 showed increased variance (welfare SD 13.53 vs 7.92 at depth 1). In the large population follow-up, this reverses: depth 3 has *lower* variance (SD 16.61) than depth 1 (SD 19.69). The recursive equilibrium appears to be stabilized by having more counterparties to average over, confirming that the baseline variance was an artifact of the small population rather than an inherent property of Level 3 reasoning.

4.5 Red-Team Implications

A red-team evaluation of the baseline scenario (no defenses) revealed a robustness score of 0.40/F with 6/8 attacks succeeding. Enabling all governance levers improved this to 0.66/D. The ModelingAdversary's near-zero payoff across all conditions underscores that even basic ecosystem design can marginalize sophisticated adversaries, regardless of LDT reasoning depth.

5 Conclusion

We implemented Level 2 and Level 3 acausal reasoning for LDT agents, along with FDT-style subjunctive dependence detection and UDT-style policy precommitment. In the baseline 7-agent simulation, we find no statistically significant effects (0/15 tests after Bonferroni correction). In follow-up experiments:

1. **Large populations (21 agents)** produce the strongest effects: depth 3 improves welfare by 5.8% ($d = -1.17$, $p = 0.018$) and honest payoffs by 9.3% ($d = -1.25$, $p = 0.013$). These are nominally significant with large effect sizes.

2. **Modeling adversaries** that infer and exploit LDT decision procedures do not create the predicted arms race — the adversary is marginalized regardless of depth.

3. **Low cooperation priors** and **short horizons** reproduce the original null result in the 7-agent setting, though short horizons show suggestive non-monotonic trends favoring depth 2.

The key insight is that **population size is the primary moderator** of acausality depth effects — not adversary sophistication, cooperation priors, or observation horizons. Deeper reasoning helps when there are more counterparties than can be fully characterized by behavioral traces alone. Implementers should default to Level 1 with FDT subjunctive dependence for small populations (< 15 agents) and enable Level 2–3 for larger ecosystems where the information advantage of deeper reasoning is realized.

Reproducibility

```
# Install
python -m pip install -e ".[dev,runtime]"

# Baseline sweep (30 runs: 3 depths x 10 seeds)
python -c "
from swarm.scenarios.loader import load_scenario
from swarm.analysis.sweep import SweepConfig, SweepParameter, SweepRunner
base = load_scenario('scenarios/ldt_cooperation.yaml')
base.orchestrator_config.n_epochs = 10
config = SweepConfig(
    base_scenario=base,
    parameters=[SweepParameter(
        'agents.ldt.config.acausality_depth', [1, 2, 3])],
    runs_per_config=10, seed_base=42)
runner = SweepRunner(config)
runner.run()
runner.to_csv('sweep_results.csv')
"

# Section 4.2 follow-up studies
for scenario in ldt_large_population ldt_modeling_adversary \
    ldt_low_prior ldt_short_horizon; do
  # Same sweep config with each scenario
  python -c "..."
done
```

References

- Yudkowsky, E. (2010). Timeless Decision Theory. MIRI Technical Report.
- Soares, N., & Fallenstein, B. (2017). Agent Foundations for Aligning Machine Intelligence with Human Interests. MIRI Technical Report.
- Wei, J., et al. (2022). Functional Decision Theory: A New Theory of Instrumental Rationality. *Philosophical Studies*.
- Rice, I. (2019). Comparison of decision theories (with a focus on logical-counterfactual decision theories). LessWrong.