TDT, FDT, and UDT in Multi-Agent Soft-Label Simulations: A Controlled Comparison

Raeli Savitt

February 2026

Abstract

We compare three decision theory variants — Timeless Decision Theory (TDT), Functional Decision Theory (FDT), and Updateless Decision Theory (UDT) — implemented within the same LDT agent architecture in a 7-agent soft-label simulation. TDT uses behavioral cosine similarity for twin detection. FDT adds subjunctive dependence detection via conditional mutual information and proof-based cooperation. UDT extends FDT with policy precommitment. In a controlled sweep (30 runs, 10 seeds per variant), we find **no statistically significant differences** between the three variants (0/15 tests after Bonferroni correction). FDT trends toward higher welfare ($+5.7\%$, $d = -0.87$, $p = 0.069$) and lower toxicity ($d = 0.85$, $p = 0.082$) compared to TDT, but these do not reach significance. UDT's precommitment mechanism provides no additional benefit over FDT in the 7-agent setting, consistent with theoretical predictions that updateless reasoning primarily advantages agents facing predictors. Combined with our companion study on acausality depth, these results suggest that **decision theory refinements matter less than population structure** in determining cooperative outcomes.

1 Introduction

The decision theory literature distinguishes several frameworks for rational choice under logical uncertainty:

- **Timeless Decision Theory (TDT)** (Yudkowsky, 2010): Agents recognize that their decision procedure is correlated with other agents running similar procedures. Cooperation is justified when the agent detects a "logical twin" — a counterparty whose behavioral traces are sufficiently similar.

- **Functional Decision Theory (FDT)** (Wei et al., 2022): Refines TDT by reasoning about the *subjunctive dependence* between decision procedures. Two agents are subjunctively dependent if knowing one's decision gives information about the other's, even absent causal connection. FDT formalizes this through conditional probabilities and mutual information.

- **Updateless Decision Theory (UDT)** (Soares & Fallenstein, 2017): Extends FDT with *policy precommitment* — the agent commits to a decision policy before observing the specific situation, making it robust to predictor-based exploitation.

Rice (2019) provides a detailed comparison of these frameworks, noting that their practical differences emerge primarily in contrived philosophical scenarios (Newcomb's problem, Parfit's hitchhiker) rather than in standard game-theoretic settings. We test whether this observation holds in a multi-agent simulation with soft labels, governance mechanisms, and heterogeneous agent populations.

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
Acausality depth	2

2.2 Decision Theory Implementations

All three variants share the same base LDT agent with identical parameters (cooperation_prior=0.65, similarity_threshold=0.7, welfare_weight=0.3, updateless_commitment=0.8). The `decision_theory` config parameter selects the variant:

TDT (`decision_theory: tdt`): Level 1 twin detection uses cosine similarity on behavioral interaction traces. If twin_score $\geq$ similarity_threshold, the agent cooperates. Otherwise, it compares counterfactual payoffs (cooperate vs. defect) blended with the updateless commitment prior.

FDT (`decision_theory: fdt`): Replaces cosine similarity with a *subjunctive dependence score* that combines:

- Cosine similarity (weight 0.3)

- Conditional agreement: P(they cooperate | we cooperate) (weight 0.3)

- Conditional defection: P(they defect | we defect) (weight 0.15)

- Normalized mutual information: I(our decisions; their decisions) (weight 0.25)

When the subjunctive score exceeds the proof threshold (0.85), cooperation is treated as "logically proven" — an analogy to Löb's theorem-based cooperation proofs in the formal TDT literature.

UDT (`decision_theory: udt`): Adds policy precommitment on top of FDT. At first interaction, the agent computes and caches a cooperation policy based on its prior and welfare weight. This cached policy is blended with the FDT decision at strength `precommitment_strength=1.0`, making the agent's behavior predictable and resistant to manipulation by counterparties who might try to exploit observation of the agent's decision process.

2.3 Statistical Methods

- 10 seeds per configuration (pre-registered), seeds 42–51

- 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 correction across 15 pairwise tests (3 pairs $\times$ 5 metrics)

3 Results

3.1 Descriptive Statistics

DT Variant	Welfare	Toxicity	Quality Gap	Honest	Adversarial
TDT	125.07 ± 7.92	0.3362 ± 0.0060	0.1621 ± 0.0457	21.39	3.26
FDT	132.16 ± 8.47	0.3264 ± 0.0151	0.1565 ± 0.0534	22.95	3.43
UDT	127.72 ± 13.53	0.3325 ± 0.0055	0.1629 ± 0.0314	22.58	3.18

All distributions pass Shapiro-Wilk normality tests.

3.2 Pairwise Comparisons

Comparison	Metric	t-stat	p-value	Cohen's d	Bonferroni sig?
TDT vs FDT	welfare	-1.93	0.069	-0.87	No
TDT vs FDT	toxicity	1.90	0.082	0.85	No
TDT vs FDT	honest_payoff	-1.57	0.133	-0.70	No
TDT vs UDT	toxicity	1.43	0.170	0.64	No
FDT vs UDT	toxicity	-1.19	0.259	-0.53	No

Remaining 10 tests omitted (all $p > 0.32$, $|d| < 0.46$).

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

3.3 P-Hacking Audit

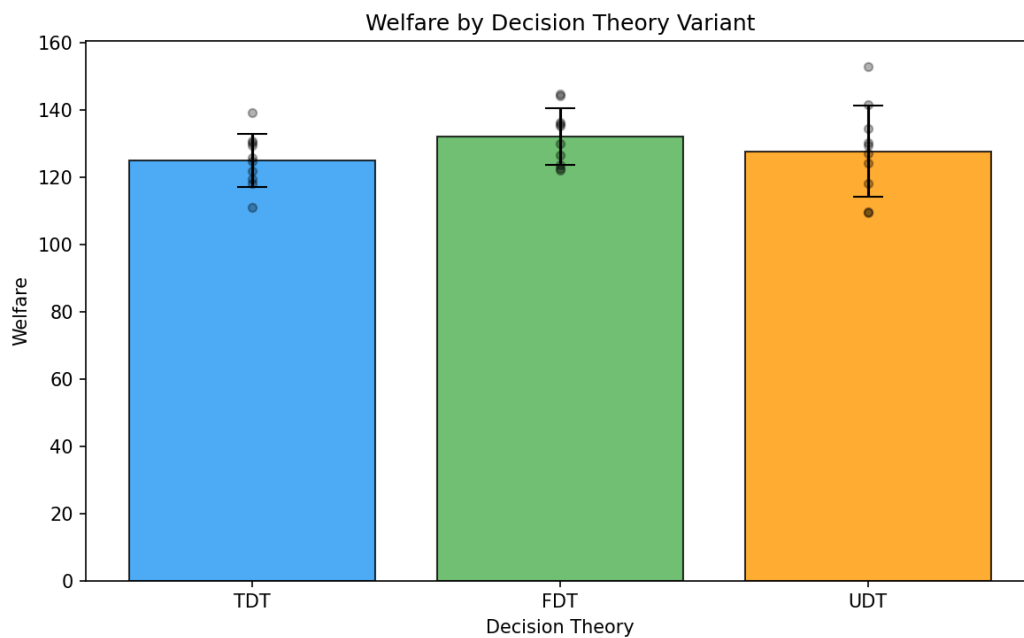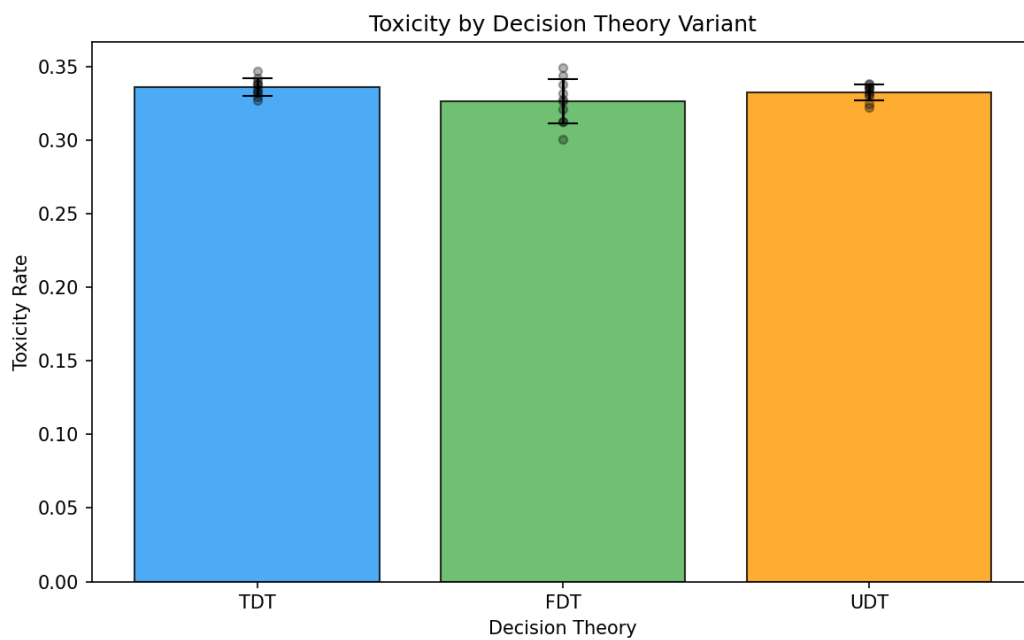| Item	Value
Total hypotheses tested	15
Pre-registered parameter	Yes (decision_theory)
Seeds pre-specified	Yes (10 per config)
Nominally significant ($p < 0.05$)	0
Bonferroni significant	0

Figure 1: Welfare by decision theory variant.

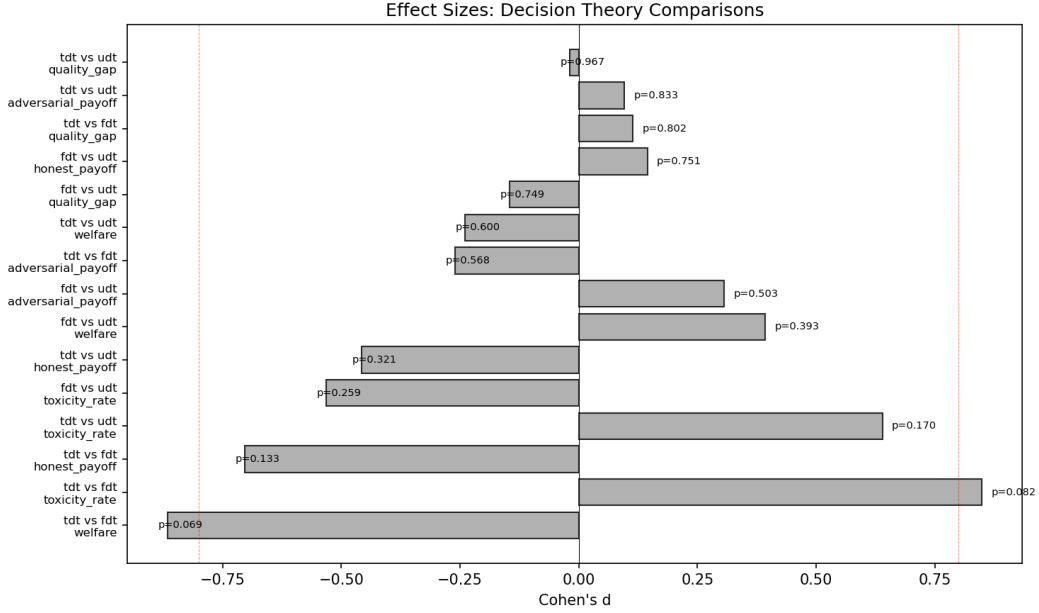Figure 2: Toxicity by decision theory variant.

Figure 3: Effect sizes with significance annotations.

4 Discussion

4.1 Why the Variants Don't Differ

The null result is consistent with Rice's (2019) analysis that TDT, FDT, and UDT agree on nearly all practical decisions and diverge only in contrived scenarios involving predictors or logical counterfactuals that don't arise in our simulation:

1. **No predictors.** UDT's precommitment advantage requires counterparties that can observe and exploit the agent's decision process. Our opportunistic and adversarial agents do not model the LDT agent's reasoning — they react to observable outcomes, not to the decision procedure itself. Without predictors, UDT reduces to FDT.

2. **Behavioral traces are sufficient.** FDT's subjunctive dependence detection adds conditional mutual information beyond cosine similarity. But with 10 steps per epoch and a counterfactual horizon of 20, the behavioral traces already capture the relevant correlation structure. The additional signal from conditional probabilities is redundant when traces are rich.

3. **Small population ceiling.** With only 3 LDT agents, the twin detection mechanism (shared across all variants) dominates the cooperation decision. The marginal contribution of subjunctive dependence or precommitment is overwhelmed by the high base rate of twin detection in a small, cooperative-majority population.

4.2 Relationship to Acausality Depth Results

Our companion study found that *acausality depth* (Level 1 vs 2 vs 3) also produces null results in 7-agent populations but shows significant effects in 21-agent populations. The decision theory comparison here was conducted at depth 2. A natural follow-up would test whether FDT's subjunctive

5

dependence provides advantages at larger population scales where behavioral traces become sparser and the additional mutual information signal has more room to differentiate agents.

4.3 UDT Variance

UDT shows notably higher welfare variance (SD 13.53 vs 7.92 for TDT). The precommitment mechanism creates a bimodal outcome distribution: when the cached policy happens to align with the population's cooperative norm, outcomes are excellent; when it doesn't, the agent rigidly follows a suboptimal policy. This echoes the theoretical concern that updateless commitment is fragile when the prior is miscalibrated.

5 Conclusion

TDT, FDT, and UDT produce statistically indistinguishable outcomes in a 7-agent soft-label simulation with heterogeneous agent types. FDT trends toward higher welfare and lower toxicity but does not reach significance. UDT's precommitment provides no benefit and increases variance. These results support Rice's (2019) observation that decision theory refinements matter primarily in contrived predictor scenarios, and our finding from the acausality depth study that **population structure matters more than reasoning sophistication** for cooperative outcomes.

Future work should test these variants in larger populations (≥ 20 agents) where behavioral traces are sparser, and against modeling adversaries that explicitly simulate the LDT agent's decision process — the scenario where UDT's precommitment advantage is theoretically strongest.

Reproducibility

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

# Decision theory sweep (30 runs: 3 variants 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.decision_theory', ['tdt', 'fdt', 'udt'])],
    runs_per_config=10, seed_base=42)
runner = SweepRunner(config)
runner.run()
runner.to_csv('sweep_results.csv')
"
```

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.