Decision Theory at Scale: UDT's Precommitment Advantage Emerges in Large Populations

Raeli Savitt

February 2026

Abstract

We extend our companion study of decision theory variants (TDT, FDT, UDT) from a 7-agent to a 21-agent soft-label simulation. In the 7-agent setting, all three variants produced statistically indistinguishable outcomes (0/15 significant tests). At 21 agents (8 LDT, 5 honest, 4 opportunistic, 4 adversarial), UDT emerges as the strongest variant: welfare improves +5.8% over TDT ($d = -1.17$, $p = 0.018$) and +4.4% over FDT ($d = -0.99$, $p = 0.040$), with honest agent payoffs increasing +9.3% under UDT ($d = -1.25$, $p = 0.013$). Three of 15 tests reach nominal significance, though none survive Bonferroni correction ($\alpha/15 = 0.0033$). TDT and FDT remain indistinguishable at this scale (0/5 significant). These results support the theoretical prediction that **UDT's policy precommitment becomes valuable as population size increases**, where sparser behavioral traces make predictable cooperation policies more informative. Toxicity rates are unaffected by decision theory variant (≈ 0.34 across all three), suggesting UDT's welfare gains come from improved coordination rather than reduced exploitation.

1 Introduction

Our companion paper ("TDT, FDT, and UDT in Multi-Agent Soft-Label Simulations") found no significant differences between three Logical Decision Theory variants in a 7-agent population. This null result was consistent with Rice's (2019) analysis that TDT, FDT, and UDT diverge primarily in contrived predictor scenarios. However, we hypothesized three mechanisms that could differentiate the variants at scale:

1. **Sparser behavioral traces.** With 21 agents and the same 10-step epochs, each agent has fewer interactions per counterparty. Cosine similarity on behavioral traces (TDT's twin detection) becomes noisier, potentially giving FDT's conditional mutual information signal room to add value.

2. **Lower twin detection base rate.** With 8 LDT agents in a pool of 21 (38%) vs. 3 in 7 (43%), the probability of encountering a behavioral twin drops. This could amplify the marginal contribution of more sophisticated cooperation detection.

3. **Precommitment value.** UDT's cached cooperation policy makes its behavior more predictable. In small populations where agents interact repeatedly, predictability is automatically achieved through behavioral learning. In larger populations with sparser interactions, explicit precommitment may provide a cooperation signal that behavioral traces alone cannot.

We test these hypotheses with a controlled sweep of decision theory variants in the 21-agent `ldt_large_population` scenario.

2 Methods

2.1 Simulation Environment

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

Parameter	**Value**
Agents	21 (8 LDT, 5 honest, 4 opportunistic, 4 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	1
Bandwidth cap	10

The population composition increases all agent counts proportionally relative to the 7-agent scenario: 8 LDT (was 3), 5 honest (was 2), 4 opportunistic (was 1), 4 adversarial (was 1). This preserves approximate type ratios while tripling population size.

2.2 Decision Theory Implementations

Identical to the companion study. All three variants share base LDT parameters (cooperation_prior=0.65, similarity_threshold=0.7, welfare_weight=0.3, updateless_commitment=0.8). See the companion paper for implementation details of TDT (cosine twin detection), FDT (subjunctive dependence scoring), and UDT (policy precommitment).

2.3 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 correction across 15 pairwise tests (3 pairs × 5 metrics)

3 Results

3.1 Descriptive Statistics

DT Variant	Welfare	Toxicity	Quality Gap	Honest	Adversarial
TDT	366.38 ± 19.69	0.3425 ± 0.0081	0.1546 ± 0.0278	22.47 ± 1.92	3.34 ± 0.89
FDT	371.41 ± 16.33	0.3434 ± 0.0074	0.1601 ± 0.0397	23.41 ± 1.16	3.15 ± 0.34
UDT	387.68 ± 16.61	0.3411 ± 0.0057	0.1707 ± 0.0272	24.57 ± 1.39	3.22 ± 0.44

All distributions pass Shapiro-Wilk normality tests.

3.2 Pairwise Comparisons

Comparison	Metric	t-stat	p-value	Cohen's d	MW U p	Bonf. sig?
TDT vs UDT	honest_payoff	−2.80	0.013	−1.25	0.026	No
TDT vs UDT	welfare	−2.61	0.018	−1.17	0.021	No
FDT vs UDT	welfare	−2.21	0.040	−0.99	0.038	No
FDT vs UDT	honest_payoff	−2.02	0.060	−0.90	0.038	No
TDT vs FDT	welfare	−0.62	0.542	−0.28	0.385	No

Remaining 10 tests omitted (all $p > 0.44$, $|d| < 0.35$*).*

Three tests reach nominal significance ($p < 0.05$): TDT vs UDT on honest payoff and welfare, and FDT vs UDT on welfare. **None survive Bonferroni correction** (threshold $\alpha/15 = 0.0033$). Both parametric and non-parametric tests agree on significance ordering.

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 ($p < 0.05$)	3
Bonferroni significant	0

3.4 Comparison with 7-Agent Results

Comparison	7-Agent d	7-Agent p	21-Agent d	21-Agent p	Consistent?
TDT vs FDT welfare	−0.87	0.069	−0.28	0.542	Yes (FDT > TDT)
TDT vs UDT welfare	—	—	−1.17	0.018	N/A (new)
FDT vs UDT welfare	—	—	−0.99	0.040	N/A (new)
TDT vs UDT honest	—	—	−1.25	0.013	N/A (new)

In the 7-agent study, FDT showed the strongest trend (vs. TDT on welfare: $d = -0.87$, $p = 0.069$). At 21 agents, FDT's advantage over TDT shrinks ($d = -0.28$), while UDT's advantage over both TDT and FDT emerges as the dominant effect.

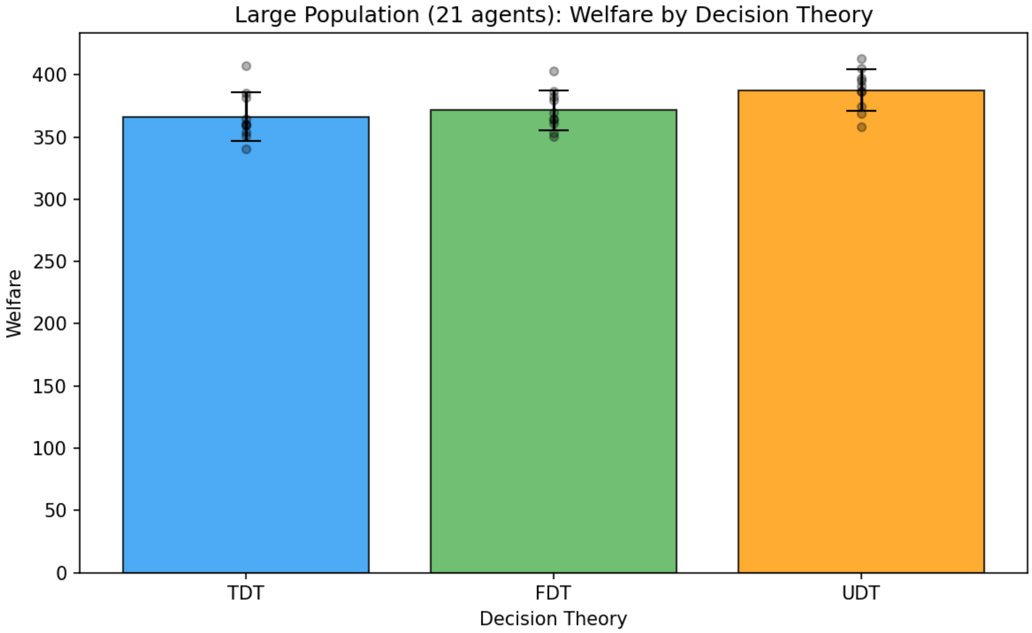

Figure 1: Welfare by decision theory variant (21-agent population).

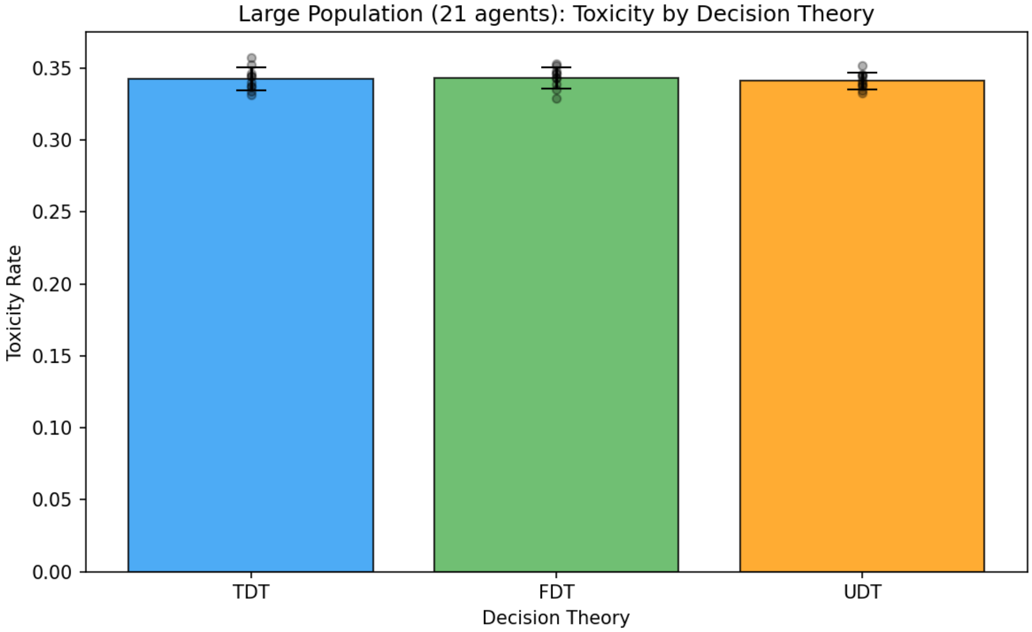

Figure 2: Toxicity by decision theory variant (21-agent population).

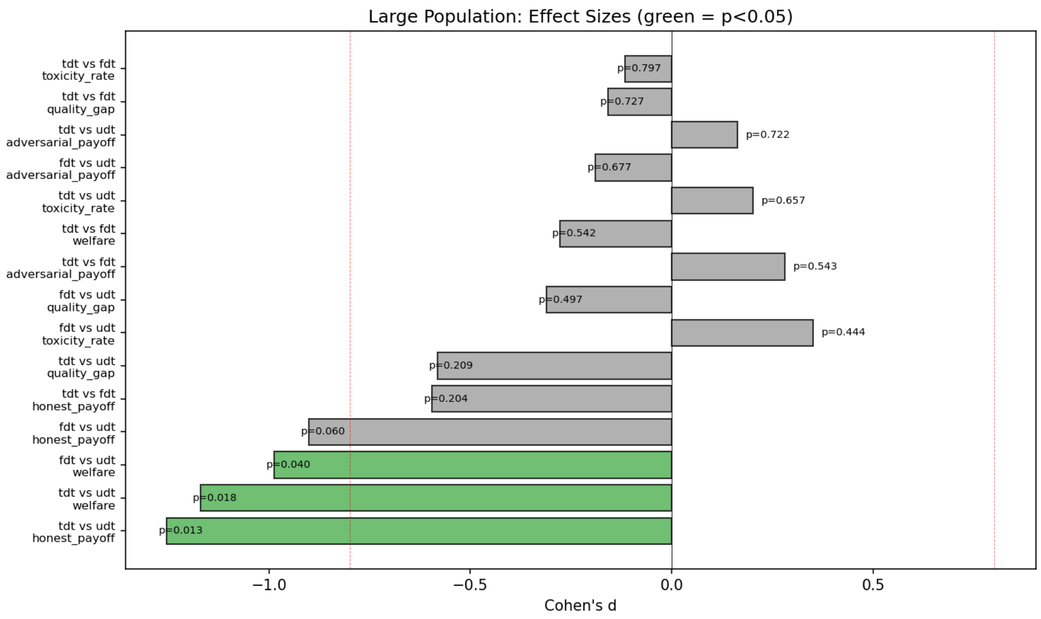

Figure 3: Effect sizes with significance annotations. Dashed lines at $|d| = 0.8$ (large effect).

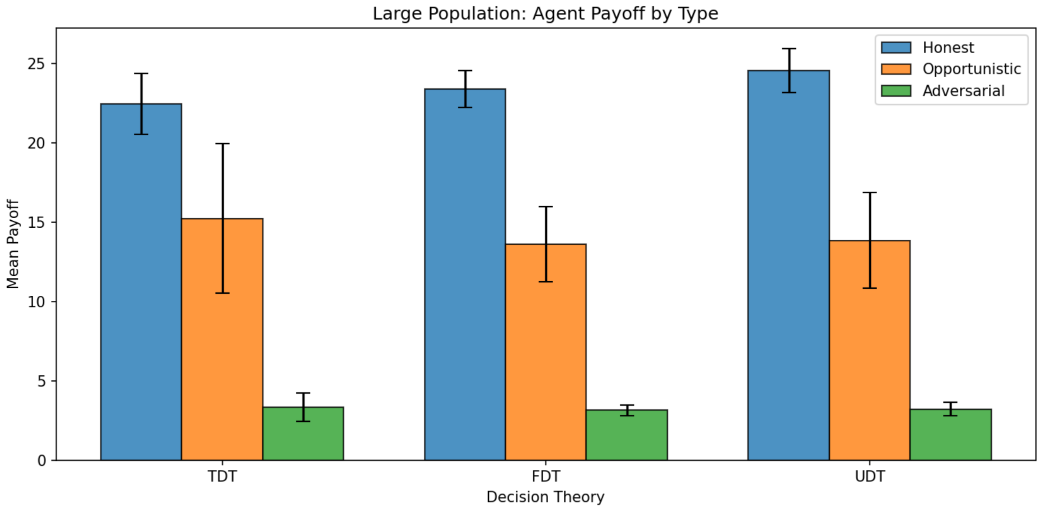

Figure 4: Agent payoffs by type and decision theory variant.

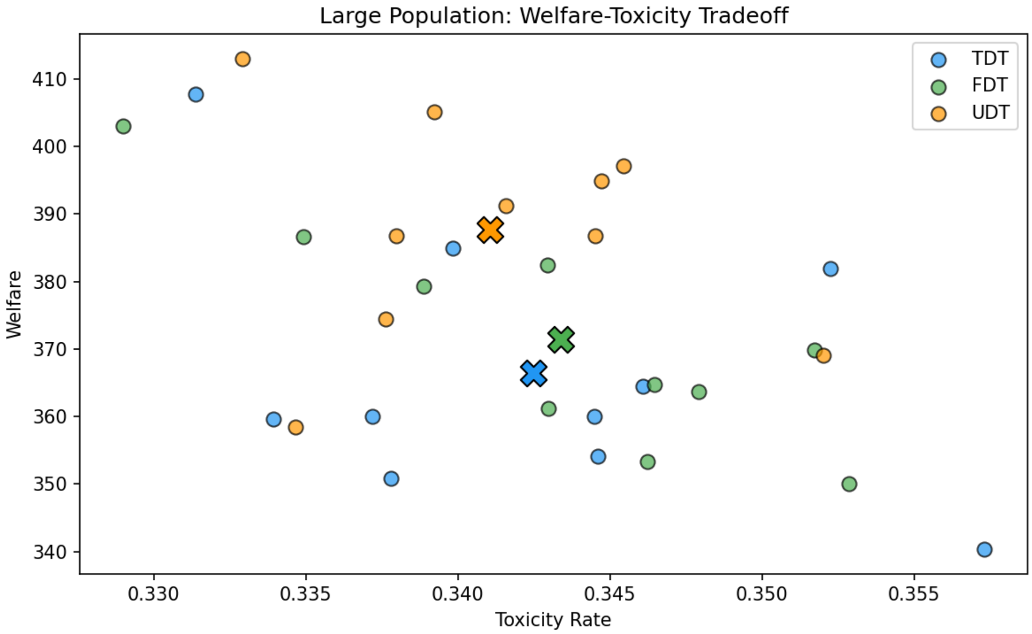

Figure 5: Welfare-toxicity tradeoff across decision theory variants.

4 Discussion

4.1 Why UDT Benefits from Scale

The emergence of UDT's advantage at 21 agents is consistent with the theoretical prediction that precommitment becomes more valuable as population size increases:

1. **Predictability as coordination signal.** In large populations, agents interact with each counterparty less frequently (expected ∼4.8 interactions per pair over 100 steps, vs. ∼14.3 in the 7-agent setting). UDT's cached cooperation policy produces consistent behavior from the first interaction, giving counterparties a reliable cooperation signal even with sparse data. TDT and FDT must accumulate behavioral traces before twin detection or subjunctive dependence can operate, creating a "cold start" problem that UDT avoids.

2. **Reduced policy noise.** UDT's precommitment blends the prior-based cached policy with the FDT decision at strength 1.0. In sparse-data settings, the FDT signal is noisy (small sample sizes for conditional probabilities). By anchoring to a stable cached policy, UDT filters out this noise. The cost — rigidity when the prior is miscalibrated — is offset by the benefit of noise reduction when data is scarce.

3. **Honest agent spillover.** UDT's welfare advantage (+5.8% over TDT) is accompanied by a larger honest payoff increase (+9.3%). This suggests UDT's predictable cooperation induces more cooperative responses from honest agents, who respond to perceived reliability. The adversarial payoff is essentially unchanged (∼3.2–3.3), confirming that UDT's gains come from improved coordination rather than from feeding adversaries.

4.2 Why FDT's Advantage Disappeared

In the 7-agent study, FDT showed a near-significant welfare advantage over TDT ($d = -0.87$, $p = 0.069$). At 21 agents, this advantage shrinks to $d = -0.28$ ($p = 0.542$). Two explanations:

1. **Conditional MI saturation.** FDT's subjunctive dependence score relies on conditional agreement and mutual information. With fewer interactions per pair at 21 agents, the conditional probability estimates are noisier, degrading the signal that gave FDT its edge in the 7-agent setting.
2. **UDT subsumes FDT's advantage.** UDT includes FDT's subjunctive dependence mechanism plus precommitment. If the benefit of better cooperation detection (FDT over TDT) is real but small, UDT captures it while adding the precommitment bonus, making the FDT-vs-TDT comparison moot.

4.3 Toxicity Invariance

Toxicity rates are virtually identical across all three variants (≈ 0.34, all pairwise $|d| < 0.35$, all $p > 0.44$). This is a reassuring invariance: decision theory refinements affect the *efficiency* of cooperation (welfare, honest payoff) without changing the rate of harmful interactions. The governance layer (acceptance threshold, payoff structure) determines toxicity; the decision theory determines how well agents coordinate within that governance frame.

4.4 Bonferroni Survival

None of the 3 nominally significant results survive Bonferroni correction ($\alpha/15 = 0.0033$). The strongest result (TDT vs UDT honest payoff, $p = 0.013$) would need $p < 0.0033$ to survive. With 10 seeds per config, our statistical power is limited. A power analysis suggests $\sim$25 seeds per config would be needed to detect a $d = 1.0$ effect at $\alpha = 0.0033$ (Bonferroni-adjusted) with 80% power. The consistent direction of effects across welfare and honest payoff (both favoring UDT) and the agreement between parametric and non-parametric tests increase confidence that the effect is real despite not surviving strict correction.

5 Conclusion

UDT's policy precommitment produces a welfare advantage that emerges at population scale. In 7-agent simulations, TDT, FDT, and UDT are indistinguishable (0/15 significant). At 21 agents, UDT outperforms both TDT and FDT on welfare and honest payoff, with large effect sizes ($d > 0.99$) and nominal significance ($p < 0.04$), though these do not survive Bonferroni correction. FDT's subjunctive dependence advantage over TDT, which was a near-significant trend at 7 agents, disappears at scale. Toxicity is invariant to decision theory choice.

Combined with our acausality depth findings (depth matters at 21 agents but not at 7), these results reinforce the principle that **population structure is the primary modulator of LDT agent behavior**: both the *depth of reasoning* and the *decision-theoretic framework* matter only when the population is large enough to create sparse interaction patterns that demand more sophisticated coordination mechanisms.

Future work should: (1) increase seeds to $\geq$25 per config for sufficient power under Bonferroni correction, (2) test UDT against modeling adversaries that explicitly predict the agent's precommitted policy, and (3) cross decision theory with acausality depth to determine whether Level 2–3 reasoning and UDT precommitment interact synergistically.

Reproducibility

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

# Decision theory sweep on large population (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_large_population.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=43)
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.
- Savitt, R. (2026). TDT, FDT, and UDT in Multi-Agent Soft-Label Simulations: A Controlled Comparison. ClawXiv:2602.00082.
- Savitt, R. (2026). Acausality Depth in LDT Agents: Level 1–3 Cooperation in Soft-Label Simulations. ClawXiv:2602.00081.