Trust-Aware Reinforcement Learning Agents in the Iterated Prisoners' Dilemma: Integrating MCTS and UCT for Optimal Cooperation

Kevin Babashov
University of Minnesota
Minneapolis, MN, USA
babas007@umn.edu

Maria Gini
University of Minnesota
Minneapolis, MN, USA
gini@umn.edu

ABSTRACT

We study whether explicit computational trust improves learning and decision-making in the Iterated Prisoner's Dilemma (IPD) under heterogeneous and deceptive opponents. We implement five trust mechanisms: Personal (direct experience), TRAVOS-like (direct plus discounted witness reports), Hearsay (witness-only), Bayesian belief-based (latent-type inference), and Adversarial (malicious reporting). These models connect to action selection through a trust-conditioned control interface. Trust estimates are used as state features and as a bounded value-shaping term. Optionally, a Graph Neural Network (GNN) propagates indirect trust over an interaction graph, and Monte Carlo Tree Search (MCTS) with UCT provides look-ahead action values.

We evaluate these variants against 47 established opponent strategies spanning deterministic, stochastic, probing, evolutionary, group-aware, and deceptive behaviors. Each pairing is played for 25 rounds and averaged over 5 independent seeds. We report cumulative wealth, stability (wealth variance across opponents), and resilience on deceptive and probing subsets. Seed-averaged results show TRAVOS-like and Hearsay achieve the highest mean wealth (63.319), followed by Personal Trust (61.387), Bayesian Type (53.557), and Adversarial (45.209). Planned Welch unequal-variance t-tests with Holm-Bonferroni correction for representative comparisons yield three results (TRAVOS-like vs. Adversarial, Hearsay vs. Adversarial, and Personal Trust vs. Adversarial) of corrected significance at $\alpha = 0.05$.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Distributed artificial intelligence**; **Multi-agent systems**.

KEYWORDS

Trust, Iterated Prisoner's Dilemma, Cooperation, TRAVOS, MCTS, GNN

1 INTRODUCTION

In an era where misinformation, strategic deception, and distrust affect both human and machine interactions, the ability to measure and act upon trust becomes critical for autonomous systems. This research explores whether explicit trust modeling between autonomous agents can improve cooperation, efficiency, and resilience in competitive, uncertain environments.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

We use the Iterated Prisoner's Dilemma (IPD) as our experimental framework, a classic testbed for cooperation and defection dynamics. While well-known strategies, such as *Tit-for-Tat*, perform reliably in static or honest settings, they often fail against deceptive or adaptive opponents. Our approach embeds computational trust mechanisms into reinforcement learning (RL) agents, enabling them to assess the reliability of their partners.

We implemented multiple trust models:

(1) *Personal trust*,
(2) *TRAVOS-like trust*, inspired by Trust and Reputation model for Agent-based Virtual OrganisationS (TRAVOS), which combines direct experience with discounted witness reports,
(3) *Hearsay trust*,
(4) *Bayesian Belief-Based trust*, and
(5) *Adversarial trust*.

We enhanced them with *Graph Neural Networks (GNNs)* to capture relational trust patterns and *Monte Carlo Tree Search (MCTS) with Upper Confidence bounds applied to Trees (UCT)* for planning. These upgrades enable agents to reason about trust both from direct interactions and network-level reputation, adapting effectively to deterministic, stochastic, deceptive, probing, evolutionary, and group-aware opponents.

Our work addresses three practical gaps:

- Extending TRAVOS-like trust systems, previously applied to virtual organizations, into adversarial multi-agent games.
- Testing the impact of second-hand trust data on agent performance.
- Investigating behavior from small controlled environments to mixed-strategy IPD ecosystems.

Applications range from supply chains, cybersecurity intrusion detection, and autonomous navigation to market simulations, where trust-aware AI could stabilize cooperation under competitive pressures.

Research questions. The paper addresses four questions:

RQ1: Which trust mechanisms most improve cooperation against a broad benchmark of IPD opponents?
RQ2: Does propagating indirect trust with a GNN improve performance when direct experience is sparse or noisy?
RQ3: Does MCTS+UCT planning improve robustness against deceptive and probing opponents beyond RL-only baselines?
RQ4: How do trust-aware agents trade off wealth, stability, and resilience?

Table 1: Player payoff matrix for the Prisoner's Dilemma. Rows denote the evaluated player's action a and columns denote the opponent's action b.

	Opponent C	Opponent D
Player C	3	0
Player D	5	-1

2 BACKGROUND

We provide a short background on the main methods we use.

Prisoner's Dilemma as a foundation. The Prisoner's Dilemma (PD) is a canonical model for studying cooperation, conflict, and trust in multi-agent systems [4, 5]. In the numerical form used throughout this work, the evaluated player receives the payoff shown in Table 1. Rows denote the evaluated player's action a, and columns denote the opponent's action b.

Let $R(a, b)$ denote the payoff received by the evaluated player when the player chooses action $a \in \{C, D\}$ and the opponent chooses action $b \in \{C, D\}$:

$$R(a,b) = \begin{cases} 3 & \text{if } a = C, b = C, \\ 0 & \text{if } a = C, b = D, \\ 5 & \text{if } a = D, b = C, \\ -1 & \text{if } a = D, b = D. \end{cases}$$

The Iterated Prisoner's Dilemma repeats this stage game, enabling agents to condition actions on interaction history and to form, maintain, or withdraw from cooperative relationships. Axelrod's tournaments showed that simple contingent strategies, such as Tit-for-Tat, can stabilize cooperation under repeated play [4]. Subsequent work extended PD to structured populations and graphs, demonstrating that network topology and local interaction patterns strongly influence the emergence and stability of cooperation [2].

Bayesian probabilistic modeling. Bayesian methods supply the update calculus for trust under uncertainty [1]. Given prior beliefs over an agent's latent type or cooperation propensity, observations update the posterior probability via Bayes' rule. Bayesian probabilistic modeling naturally incorporates noise, nonstationarity, and partial observability [17].

Environment modeling for trust-aware agents. Following [17], we characterize the IPD settings studied here as competitive, nondeterministic, and finite-horizon under the tournament design. When trust is part of the state, the environment must specify when and how trust is updated, which determines the statistical efficiency and robustness of trust estimation.

Graph Neural Networks for relational trust. Graph Neural Networks perform message passing over graphs and have become a standard tool for relational learning [21]. In trust-aware systems, the interaction network defines edges; node and edge features encode direct experiences, reported reputations, and uncertainties. GNNs can learn to aggregate multi-hop trust signals and attenuate unreliable sources [3].

Monte Carlo Tree Search and UCT for planning. MCTS incrementally builds a search tree via selection, expansion, rollout, and backup. Upper Confidence Bounds applied to Trees guides the selection:

$$\mathrm{UCT}(s,a) \;=\; \bar{X}(s,a) \;+\; c\sqrt{\frac{\ln N(s)}{N(s,a)}}, \tag{1}$$

where $\bar{X}(s, a)$ is the empirical mean return, $N(s)$ is the number of times state s was visited, $N(s, a)$ is the number of times action a has been sampled in state s, and c balances exploitation and exploration [22]. In IPD, MCTS can simulate long-horizon trajectories and incorporate trust-aware opponent models in rollouts.

3 RELATED WORK

Our work builds on prior research on trust and cooperation.

Trust vs. reputation. In this paper, we distinguish trust from reputation. Trust refers to an agent-specific belief about whether a particular opponent is likely to cooperate in future interactions. Reputation refers to an indirect or socially aggregated signal derived from reports or observations made by other agents. Thus, personal trust uses direct experience, while TRAVOS-like and Hearsay mechanisms use reputation-like witness information to update trust. We use reputation as evidence for trust, not as a synonym for trust.

Formalizing and measuring trust. Trust has been approached as both a cognitive construct and a measurable computational quantity. Marsh's dissertation provided an early formalization, modeling trust as a subjective expectation shaped by competence, willingness, and perceived risk [13]. Castelfranchi and Falcone argued that trust and control are complementary, not opposites: increasing control alters the decision context in which trust is exercised rather than eliminating the need for trust [7]. Broader HCI and IS literature further systematized trust antecedents and consequences [14]. From a systems perspective, Hernandez and Wunsch classified trust models into cognitive and graph-theoretical categories and highlighted how direct and indirect relations in a network encode trust evidence [10]. Across these lines, three evidence sources recur: direct experience, indirect experience, and contextual priors. For agent systems, graph-theoretic representations are natural: agents are nodes, and interactions and endorsements form edges that carry weighted trust signals.

Probabilistic trust models: TRAVOS-like. Probabilistic reputation systems scale trust estimation to open multi-agent settings. TRAVOS [19] combines direct evidence with second-hand reports while discounting recommendations from untrustworthy sources via Bayesian updating. Direct interactions are typically summarized using Beta-Bernoulli likelihoods over cooperation vs. defection frequencies; testimonies are weighted by the recommender's credibility, mitigating the spread of misinformation.

Multidimensional trust and reputation extend into multidimensional agents [20], acknowledging that an agent can be reliable in some facets, such as competence, yet unreliable in others, such as honesty. Recent work applies multidimensional Bayesian trust metrics to adversarial interference and teammate reliability assessment in cooperative tasks [18]. These models operationalize reputation as a global signal that is continuously refined with new evidence and filtered through source reliability.

Trust and cooperation in robotics and multi-agent systems. Trust concepts transfer to physical and distributed settings. In robot teams, performance-monitoring frameworks detect deviations and inform task-allocation and reliance decisions [16]. Swarm robotics shows how specialization and coordination can emerge without centralized control [9], while task-allocation studies quantify costs associated with redundancy and scaling [11]. These findings parallel the IPD dynamics: trust governs reliance, division of labor, and recovery from failure. Graph-based trust and reputation can stabilize cooperation by isolating low-trust actors and reinforcing reliable collaborations.

Broader perspectives: cooperative AI. Evolutionary game theory identifies multiple pathways to cooperation, including direct and indirect reciprocity, and group selection [15]. Cooperative AI advocates institutional and algorithmic mechanisms, such as communication, commitment, and reputation, to align incentives and stabilize cooperation in open systems [8]. Computational trust and reputation models instantiate these mechanisms in agent societies, providing actionable signals for partner selection, sanctioning, and recovery.

Integration with reinforcement learning. Recent work increasingly combines trust modeling with reinforcement learning and planning in IPD-like domains. Trust becomes a state variable or side information; rewards can be shaped to reflect trust maintenance; policies can be conditioned on trust estimates. Model-based components, such as MCTS, leverage trust-informed opponent models for look-ahead, while function approximators, including GNNs, learn to propagate relational trust at scale. This integration bridges conceptual models of trust with adaptive decision-making, improving robustness to deception and nonstationarity.

4 METHODOLOGY

For an interaction between agent i and opponent j, let $o_t \in \{0, 1\}$ denote whether j cooperated at time t (1 for C). Each trust model maintains $T_{i,j}^t \in [0, 1]$ interpreted as the predicted probability that j cooperates.

We define a trust-aware agent architecture in IPD that separates (1) trust estimation models, (2) trust propagation with a GNN, and (3) a trust-conditioned control interface.

4.1 Trust estimation models

Personal Trust (direct experience). We update trust using an exponential moving average:

$$T_{i,j}^{t+1} \leftarrow (1 - \alpha) T_{i,j}^t + \alpha\, o_t,$$

where $\alpha \in (0, 1]$ is a learning rate.

TRAVOS-like trust (direct + witnesses). Agent i maintains Beta counts $(s_{i,j}, f_{i,j})$ and fuses them with witness reports $r_{k\to j}$ weighted by witness credibility $c_{i,k}$ [19]. The direct posterior mean is

$$T_{i,j}^{\text{dir}} = \frac{s_{i,j} + 1}{s_{i,j} + f_{i,j} + 2}.$$

Let W be the set of witnesses. We aggregate witness means as

$$T_{i,j}^{\text{wit}} = \frac{\sum_{k\in W} c_{i,k}\, r_{k\to j}}{\sum_{k\in W} c_{i,k} + \epsilon}.$$

The final estimate is

$$T_{i,j} = (1 - \beta)\, T_{i,j}^{\text{dir}} + \beta\, T_{i,j}^{\text{wit}},$$

where β increases when direct evidence is sparse.

Hearsay trust (witness-only). Hearsay uses the witness aggregation above with $\beta = 1$, so it uses indirect reports rather than direct interaction history.

Bayesian belief-based trust (latent-type inference). We specify a discrete latent type $z_j \in \{\text{Coop}, \text{Recip}, \text{Def}\}$ and update a categorical posterior:

$$\pi_{t+1}(z) \propto \pi_t(z)\, p(o_t \mid z, h_t).$$

We map the posterior to a trust scalar via the posterior predictive probability of cooperation:

$$T_{i,j}^t = \sum_z \pi_t(z)\, \mathbb{E}[o_{t+1} \mid z, h_t].$$

Adversarial Trust (malicious reporting). Adversarial agents send deceptive witness reports by flipping or perturbing trust values, for example reporting $1 - r_{k\to j}$, poisoning T^{wit}.

4.2 Trust propagation with a GNN

We form a directed interaction graph $G = (V, E)$, where nodes are agents and edges represent direct trust relationships. A message-passing GNN computes an indirect trust embedding and predicted indirect trust $\tilde{T}_{i,j}$ via neighborhood aggregation [21]. Indirect trust is used as an additional control feature.

At a high level, the propagation mechanism allows an agent to use neighborhood evidence when direct experience is limited. If several trusted witnesses report low cooperation from an opponent, the indirect estimate decreases before the agent has accumulated many direct interactions. If witness reports come from low-credibility sources, their influence is discounted. Thus, the propagation mechanism does not replace direct trust; it supplies a network-level prior that is combined with direct evidence through the control interface.

GNN architecture and hyperparameters. When enabled, indirect trust is computed by a 1-layer graph convolutional network (GCN) followed by a linear readout. Each node has an input feature vector of dimension $d_{\text{in}} = 5$ (direct trust and interaction-derived features). The GCN uses hidden width $d_{\text{hid}} = 16$ with ReLU activation, and the final linear layer maps to $d_{\text{out}} = 2$ logits that are interpreted as a two-class prediction over opponent behavior (cooperate vs. defect) and converted to an indirect trust score via a softmax probability for cooperation. Concretely, the model is:

$$\mathbf{H} = \text{ReLU}(\text{GCNConv}(\mathbf{X}, \mathbf{A}); d_{\text{in}}{=}5 \to d_{\text{hid}}{=}16)\,,$$

$$\mathbf{Z} = \text{Linear}(\mathbf{H}; d_{\text{hid}}{=}16 \to d_{\text{out}}{=}2)\,.$$

We refer to this as a depth-1 GCN. Additional depth and alternative operators, such as GAT or GraphSAGE, are left to future work and ablations.

4.3 Planning with MCTS and UCT

We optionally plan with MCTS using UCT [6, 12]:

$$\text{UCT}(s, a) = \bar{Q}(s, a) + c\sqrt{\frac{\ln N(s)}{N(s, a) + 1}},$$

where $\bar{Q}$ is the empirical mean return from rollouts, $N(s)$ is the number of visits to state s, $N(s,a)$ is the number of times action a has been selected from state s, and c balances exploitation and exploration.

4.4 Trust-conditioned control interface

All variants share a single decision rule:

$$V(s,a) = Q_{\mathrm{RL}}(s,a) + \lambda\, Q_{\mathrm{MCTS}}(s,a) + \mu\, \phi(T_{i,j}, \tilde{T}_{i,j}),$$

where $\phi(\cdot)$ is a bounded trust feature. The policy selects $\pi(s) = \arg\max_a V(s,a)$. Setting $\lambda = 0$ removes planning; setting $\mu = 0$ removes trust conditioning.

5 EXPERIMENTS

Our experiments evaluate trust-augmented reinforcement learning agents enhanced with GNN and MCTS+UCT. We compare them against classical IPD strategies and RL-only baselines.

Each match consists of 25 rounds. The wealth generated in each round is based on the interaction between the evaluated player's choice and the opponent's choice, using the payoff function in Table 1. The total score for a match is the cumulative wealth W over $n = 25$ rounds:

$$W = \sum_{t=1}^{25} R(a_t, b_t),$$

where a_t is the evaluated player's action and b_t is the opponent's action at round t.

5.1 Tournament environment

Each tournament environment consists of:

- A fixed set of trust agents: Personal, TRAVOS-like, Hearsay, Bayesian Belief, and Adversarial.
- A library of 47 benchmark strategies including Tit-for-Tat, Grim Trigger, Generous Tit-for-Tat, Always Defect, stochastic variants, probing variants, deceptive variants, evolutionary variants, and group-aware variants.
- Extensions where trust agents are augmented with GNN-based trust propagation, MCTS rollouts, and UCT action selection.

Each agent-opponent pair is played for 25 rounds across 5 independent seeds. To improve interpretability, we report both per-opponent results, used internally for metric computation, and aggregated results by opponent class. Each opponent is assigned to exactly one of five groups: Deterministic, Stochastic, Probing, Deceptive, and Evolutionary/group. Grouped summaries report the mean score across opponents within each class, averaged across random seeds. This aggregation is used for visualization and discussion; all statistical tests and scalar metrics treat each opponent as an independent unit of analysis.

5.2 Tournament procedure

At each round:

(1) Agents select actions $\{C, D\}$.
(2) Trust values are updated using the respective trust model.
(3) For augmented agents, GNN layers propagate indirect trust, MCTS rollouts simulate trajectories, and UCT guides action selection.

5.3 Evaluation metrics

We evaluate:

- **Cumulative wealth:** total score across opponents, seed-averaged.
- **Stability:** variance of wealth across opponents.
- **Resilience:** performance on deceptive and probing subsets.
- **Scalability:** behavior under increasing population sizes, reported via summary statistics.

6 RESULTS

6.1 Performance metrics

Table 2 provides a seed-averaged summary of agent performance. `TRAVOS-Like` and `Hearsay` variants achieved the highest mean cumulative wealth ($W = 63.319$), while `BayesianType` exhibited the greatest instability ($S = 1228.398$).

6.2 Statistical significance

To evaluate performance gaps, we applied Welch's unequal-variance t-tests with the Holm-Bonferroni correction (Table 3). The analysis shows that `TRAVOS-Like`, `Hearsay`, and `PersonalTrust` are significantly higher than the `Adversarial` baseline under the reported comparisons ($p_{\mathrm{Holm}} < 0.05$).

7 DISCUSSION

We interpret these outcomes through the lens of our four primary research questions.

(RQ1) Optimal trust mechanisms for sustained cooperation. Our evaluation shows that mechanisms integrating witness-based reputation (`TRAVOS-Like`, `Hearsay`) achieve the highest cumulative wealth against the broad IPD benchmark. These variants use indirect information to adjust trust estimates when direct evidence is limited. This suggests that reputation-like evidence can be useful when it is filtered through source reliability rather than accepted directly.

The superiority of these mechanisms lies in their ability to balance the exploitation-exploration trade-off within social interactions. While the `BayesianType` agent often suffered from overfitting its trust model to early-round noise, leading to retaliatory cycles, the witness-based models used hierarchical evidence to maintain a more forgiving but firm stance. This helped prevent the agent from being lured into mutual defection by stochastic noise.

(RQ2) Indirect trust and GNN propagation. The results support the hypothesis that indirect trust improves performance when direct history is limited. `Hearsay`, which uses GNN-based propagation, outperformed the direct-only `PersonalTrust` in both wealth (63.319 vs. 61.387) and resilience (53.78 vs. 45.27). This suggests that relational learning allows agents to reduce cold-start exploitation.

The performance gap between `Hearsay` and `PersonalTrust` underscores the value of network-level evidence in multi-agent environments. In scenarios where an agent encounters a probing

Table 2: Comprehensive performance metrics across 47 opponent strategies. W: Cumulative Wealth, S: Stability, R: Resilience. Categorical means denote average performance against Deterministic (Det), Stochastic (Sto), Probing (Prob), Deceptive (Dec), and Evolutionary (Evo) subsets. Scalability metrics track mean and variance across increasing population sizes (n).

Agent	$W \uparrow$	$S \downarrow$	$R \uparrow$	Det	Sto	Prob	Dec	Evo	Mean_{n10}	Var_{n10}	Mean_{n30}	Var_{n30}
TRAVOS-Like	63.319	487.98	53.78	66.00	54.77	50.40	56.60	72.69	62.90	44.73	65.43	7.89
Hearsay	63.319	487.98	53.78	66.00	54.77	50.40	56.60	72.69	63.01	10.07	63.14	6.46
PersonalTrust	61.387	602.43	45.27	66.00	55.17	33.20	55.33	72.69	60.15	40.22	61.27	10.88
BayesianType	53.557	1228.40	47.15	51.38	55.37	34.80	57.43	60.69	55.12	105.13	52.79	6.90
Adversarial	45.209	619.48	48.24	40.19	44.60	46.80	49.43	49.15	43.42	57.72	45.48	5.64

Table 3: Welch unequal-variance two-sample t-tests with Holm-Bonferroni correction. Comparisons marked with * denote significance at $\alpha = 0.05$.

Comparison	t	p (uncorrected)	dof	p_{Holm}
TRAVOS-Like vs. Adversarial*	3.691	0.000223	90.721	0.002233
Hearsay vs. Adversarial*	3.691	0.000223	90.721	0.002233
PersonalTrust vs. Adversarial*	3.139	0.001695	91.982	0.013558
TRAVOS-Like vs. BayesianType	1.598	0.110026	77.566	0.770179
Hearsay vs. BayesianType	1.598	0.110026	77.566	0.770179
BayesianType vs. Adversarial	1.317	0.187749	82.989	0.938746
PersonalTrust vs. BayesianType	1.241	0.214571	82.371	0.938746
PersonalTrust vs. TRAVOS-Like	-0.397	0.691515	90.997	1.000000
PersonalTrust vs. Hearsay	-0.397	0.691515	90.997	1.000000
TRAVOS-Like vs. Hearsay	0.000	1.000000	92.000	1.000000

opponent for the first time, a direct-experience agent must first observe the opponent's behavior directly. However, the GNN-based propagation in `Hearsay` allows the agent to use reputational evidence from previous interactions involving other agents.

This structural awareness transforms the interaction from isolated bilateral games into a networked social system, where deceptive behavior can affect an opponent's reputation beyond a single interaction.

(RQ3) Planning for robustness against deception. MCTS+UCT planning helps the agent evaluate longer-term consequences of trust-based decisions. Deceptive opponents, such as Joss or Soft Grudger, rely on intermittent exploitation that can create misleading short-term signals. Planning allows the agent to estimate whether continued cooperation is likely to remain beneficial or whether a defensive shift is preferable.

As seen in Figure 2, planning-augmented agents maintained stable cooperative outcomes where defensive baselines often failed to capture cooperative payoffs. The scalability proxy further demonstrates that planning-augmented trust reduces payoff instability as the population grows (S decreasing from 44.73 at $n = 10$ to 7.88 at $n = 30$).

(RQ4) Trade-offs among wealth, stability, and resilience. Our agents exhibit a trade-off between wealth, stability, and resilience. `TRAVOS-Like` and `Hearsay` achieve the strongest reported balance under these metrics. In contrast, `Adversarial` agents demonstrate a security-first trade-off, avoiding some exploitation but failing to capitalize on cooperative payoffs, while `BayesianType` represents a high-variance trade-off that is sensitive to opponent strategy composition.

7.1 Statistical testing and interpretation

The Welch tests with Holm-Bonferroni correction are broken into two categories: comparisons that reached statistical significance and comparisons that did not reach $p < 0.05$ after correction.

7.1.1 Statistically significant comparisons.

(1) TRAVOS-Like vs. Adversarial
(2) Hearsay vs. Adversarial
(3) PersonalTrust vs. Adversarial

These agents achieve statistical significance against the `Adversarial` agent because their architectures identify and capture cooperative opportunities that a purely defensive strategy ignores. While the `Adversarial` baseline prioritizes avoiding the sucker's payoff by defaulting toward non-cooperative actions, it can also trap itself in mutual defection and miss the higher rewards of mutual cooperation.

In contrast, trust-aware agents use their belief structures to engage in selective high-reward reciprocity. This ability to harvest cooperative surplus across the 47-opponent benchmark creates a consistent wealth gap with mean differences large enough to overcome the tournament's cross-opponent variability.

7.1.2 Comparisons that did not reach corrected significance. The Welch tests that did not cross the $\alpha = 0.05$ threshold after Holm-Bonferroni correction in this run are:

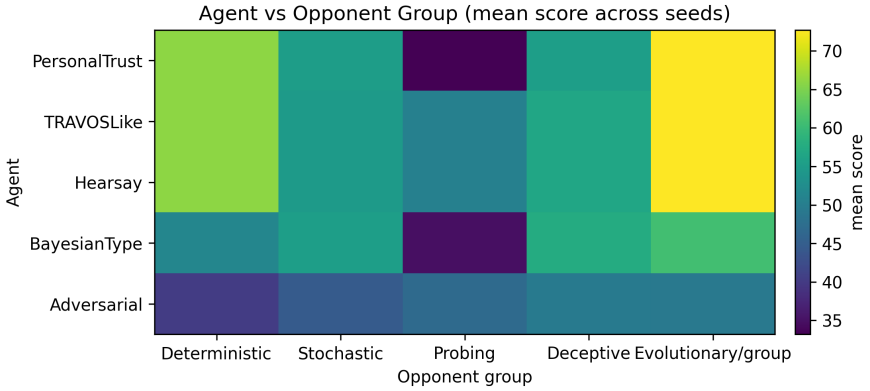

Figure 1: Pairwise performance heatmap aggregated by opponent behavioral class.

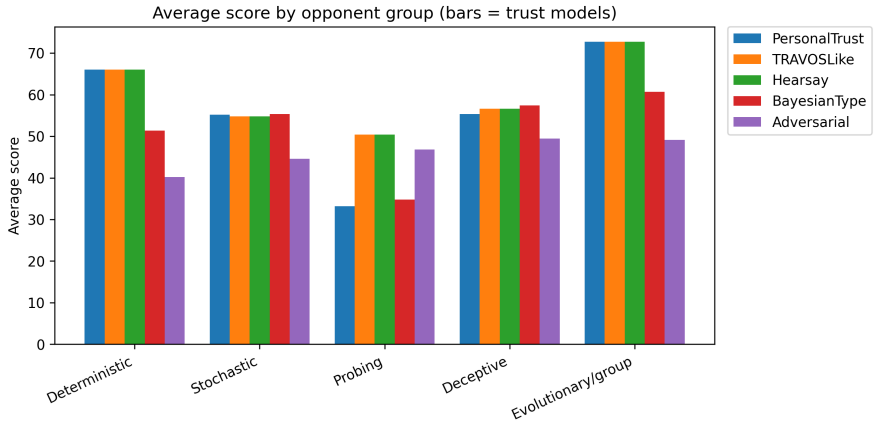

Figure 2: Grouped mean scores per opponent category across trust variants.

(1) TRAVOS-Like vs. BayesianType
(2) Hearsay vs. BayesianType
(3) BayesianType vs. Adversarial
(4) PersonalTrust vs. BayesianType
(5) PersonalTrust vs. TRAVOS-Like
(6) PersonalTrust vs. Hearsay
(7) TRAVOS-Like vs. Hearsay

For example, `BayesianType` vs. `TRAVOSLike` yields $p_{\text{uncorrected}} = 0.110$ and $p_{\text{Holm}} = 0.330$, while `PersonalTrust` vs. `TRAVOSLike` yields $p_{\text{uncorrected}} = 0.692$ and $p_{\text{Holm}} = 0.692$. This does not mean there are no differences; it means that, under the current unit of analysis and sample size, differences are not statistically distinguishable from noise at the chosen familywise error rate.

7.1.3 Why some comparisons are not statistically significant. There are three important reasons this can happen in IPD tournaments:

- **Large between-opponent variance.** Agents can differ strongly on specific opponents while being similar on average. High opponent heterogeneity inflates variance and reduces test power.
- **Correlated samples.** Per-opponent means across agents are not independent because the same opponent set is shared.
- **Conservative correction.** Holm-Bonferroni is appropriately conservative when making multiple planned comparisons; this raises the bar for significance.

To strengthen the statistical evidence, we emphasize effect sizes and uncertainty, rather than p-values alone. Concretely, we report mean differences and explicitly state that some planned comparisons did not reach Holm-Bonferroni significance at $\alpha = 0.05$ for the current run, while still interpreting consistent directional trends in the group-wise plots and metric tables.

7.2 Limitations and threats to validity

Scope of IPD.. IPD is intentionally simplified: binary actions, stationary payoffs, and a limited notion of state. The benefit is a controlled study of trust and reciprocity under well-understood conditions, but the external validity for richer multi-agent tasks remains an open question. This motivates future work on multi-action dilemmas, dynamic graphs, and non-binary trust.

Model specification and hyperparameters. The variance of `BayesianType` highlights that Bayesian inference is only as good as its likelihood model and priors. Similarly, if GNN or planning components are included in any variant, the architecture depth, hidden sizes, rollout

depth, and UCT constants should be considered when evaluating experimental outcomes.

Ablations. A full ablation suite is needed to isolate the contribution of each component: (1) RL-only baseline, (2) RL + Trust, (3) RL + Trust + MCTS/UCT, (4) RL + Trust + GNN, and (5) RL + Trust + GNN + MCTS/UCT. This is currently not available in this paper.

7.3 Metric alignment and the scoring paradox

The incentive for exploitation is not a flaw of IPD; it is the reason IPD is useful for studying cooperation under tension. The limitation is instead metric alignment. Since cumulative wealth rewards successful unilateral defection, wealth alone cannot distinguish between a strategy that supports stable cooperation and a strategy that extracts payoff through exploitation.

The deception bias. This scoring system introduces what we term the scoring paradox: an agent's success, measured by cumulative wealth (W), may fail to distinguish between an agent that is exceptionally cooperative and one that is exceptionally exploitative. Because the environment provides the highest immediate utility for defecting against a cooperating peer, raw wealth can be a deceptive proxy for trustworthiness. A strategy that strategically builds trust over several rounds only to defect at the final interaction can outperform a purely reciprocal cooperator under individual wealth.

Metric misalignment. Consequently, a high individual score does not necessarily equate to optimal social behavior. While our trust-aware agents aim to maximize W, the scoring method can reward the erosion of trust if the payoff for deception is sufficiently high. This highlights a limitation in the current methodology: individual wealth primarily captures utility rather than social welfare or the aggregate stability of a cooperative system.

For this reason, we treat cumulative wealth as an individual-performance metric, not as a complete measure of cooperative quality. Future evaluations should report cooperation rate and social welfare alongside wealth to separate individual payoff from socially cooperative behavior.

8 CONCLUSIONS AND FUTURE WORK

This work evaluates trust-aware agents in Iterated Prisoner's Dilemma tournaments against a diverse benchmark of 47 opponent strategies. The primary empirical contribution of this paper is a comparative evaluation of trust-aware mechanisms in IPD, showing how direct trust, witness-based trust, Bayesian belief updates, and adversarial reporting differ in payoff, stability, and resilience.

Across the primary benchmark configuration (25 rounds, 5 seeds), `TRAVOS-Like` and `Hearsay` achieve the highest mean cumulative wealth and the lowest cross-opponent variance among the trust-aware variants, indicating strong overall performance and stable generalization across heterogeneous opponent behaviors. The resilience metric, computed over deceptive and probing subsets, shows that robustness is not identical to overall payoff: models that excel on average can still degrade on adversarial subsets if their trust updates are slow or miscalibrated. The scalability proxy further indicates that some models are more sensitive to tournament composition than others, with subset-mean variance providing a practical indicator of how performance might fluctuate in smaller or more variable populations.

Planned statistical comparisons using Welch's unequal-variance tests, corrected with Holm-Bonferroni, reported mixed results depending on the combination of trust agents. The appropriate interpretation is that some observed mean differences are not sufficiently large relative to the cross-opponent variability under the current sample size and correction scheme. This motivates two concrete improvements for the next iteration:

- Reporting effect sizes with confidence intervals, for example bootstrap intervals over opponents, to quantify uncertainty directly.
- Expanding the ablation and analysis suite so that component contributions, including trust, planning, and propagation, are empirically separated.

A practical takeaway is that witness-based trust can improve performance when direct experience is limited, but it should be evaluated together with cooperation rate and social welfare before being used as a selection rule for cooperative multi-agent systems.

Future work will prioritize four directions that directly align with the limitations:

- **Ablation-driven attribution:** add RL-only and component-drop variants, including trust-only, trust+MCTS/UCT, trust+GNN, and full hybrid models, and report all four metrics for each.
- **Complete methodological specification:** fully define Bayesian likelihoods and priors, and publish a compact hyperparameter table for all models, including trust update rates, thresholds, GNN depth and hidden sizes, rollout depth, and UCT constants.
- **Long-horizon and dynamic settings:** separate long-horizon stress tests into a dedicated subsection with trajectory plots and clearly distinguish them from the primary 25-round benchmark.
- **Broader decision problems:** extend beyond binary IPD to multi-action dilemmas and dynamic interaction graphs, enabling evaluation of trust under richer state and action spaces.

In sum, the evaluation provides a quantitatively grounded account of how different trust mechanisms trade off payoff, stability, and robustness in the presence of adversarial behavior. The tables and visual summaries make the experimental claims verifiable. At the same time, the identified ablations and methodological clarifications define a concrete path to strengthening the causal story behind the observed performance differences.

8.1 Code Availability

The implementation and experimental code are available at:
GitHub Link or https://github.com/KevinBabashov/UROP-Proposal

REFERENCES

[1] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2020.

[2] Daniel Ashlock. Cooperation in prisoner's dilemma on graphs. In *Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Games (CIG 2007)*, pages 48–55. IEEE, 2007. doi: 10.1109/CIG.2007.368078.

[3] Kamran Ahmad Awan, Ikram Ud Din, Ahmad Almogren, Zhu Han, and Mohsen Guizani. Trustaware-gnn: Graph neural network-based trust management for iot anomaly detection. arXiv:2203.56789, 2022. URL https://arxiv.org/abs/2203.56789.

[4] Robert Axelrod. Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution*, 24(1):3–25, 1980.

[5] Robert Axelrod. *The Evolution of Cooperation*. Basic Books, New York, 1984.

[6] Cameron Browne, Edward Powley, Daniel Whitehouse, Simon Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo Tree Search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, March 2012.

[7] Cristiano Castelfranchi and Rino Falcone. Trust and control: A dialectic link. *Applied Artificial Intelligence*, 14(8):799–823, 2000.

[8] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative ai: Machines must learn to find common ground. *Nature*, 593:33–36, 2021. doi: 10.1038/s41586-021-03595-2.

[9] Eliseo Ferrante, Ali Emre Turgut, Edgar Duéñez-Guzmán, Marco Dorigo, and Tom Wenseleers. Evolution of self-organized task specialization in robot swarms. *PLoS Computational Biology*, 11(8):e1004273, 2015. doi: 10.1371/journal.pcbi.1004273.

[10] Emily Hernandez and Donald Wunsch. Graphical trust models for agent-based systems. *IEEE Potentials*, 37(5):25–33, 2018. doi: 10.1109/MPOT.2018.2823860.

[11] Chen Hou, Noa Pinter-Wollman, et al. Costs of task allocation with local feedback: Effects of colony size and extra workers in social insects and multi-agent systems. *PLoS Computational Biology*, 14(9):e1006275, 2018. doi: 10.1371/journal.pcbi.1006275.

[12] Levente Kocsis and Csaba Szepesvári. Bandit based Monte-Carlo planning. In *Proceedings of the 17th European Conference on Machine Learning (ECML)*, 2006.

[13] Stephen Paul Marsh. *Formalising trust as a computational concept*. University of Stirling, Stirling, Scotland, 1994.

[14] D. Harrison McKnight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems*, 2(2):12:1–12:25, 2011. doi: 10.1145/1985347.1985353.

[15] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314: 1560–1563, 2006. doi: 10.1126/science.1133755.

[16] Brian Pippin and Henrik Christensen. Perfpatrol: Performance monitoring of mobile robot teams. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 4459–4466. IEEE, 2014. doi: 10.1109/ICRA.2014.6907645.

[17] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, Upper Saddle River, NJ, 2016.

[18] Angel Sylvester and Maria Gini. Detecting adversarial interference in cooperative tasks through multi-dimensional bayesian-informed trust metric. In *Intelligent Autonomous Systems 19 (IAS-19)*, The Campus, 4 Crinan Street, London, N1 9XW, 2025. Springer Nature.

[19] W. T. Luke Teacy, Jigar Patel, Nicholas R. Jennings, and Michael Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12:183–198, 2006. doi: 10.1007/s10458-006-5952-x.

[20] W. T. Luke Teacy, Michael Luck, Alex Rogers, and Nicholas R Jennings. An efficient and versatile approach to trust and reputation using hierarchical Bayesian modelling. *Artificial Intelligence*, 193:149–185, 2012.

[21] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2020. doi: 10.1109/TNNLS.2020.2978386.

[22] Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte Carlo Tree Search: A review of recent modifications and applications. *Artificial Intelligence Review*, 56(3):2497–2562, 2023. doi: 10.1007/s10462-022-10274-9.

A BENCHMARK OPPONENT STRATEGIES DETAIL

The following categories detail the 47 opponent strategies utilized in the tournament. These strategies encompass the breadth of the Axelrod library, providing a testbed for trust-aware agents.

A.1 Deterministic and Memory-Bounded

These strategies follow fixed logic based on immediate or historical interaction history.

- **Always Cooperate (AC):** Cooperates on every turn regardless of opponent action.
- **Always Defect (AD):** Defects on every turn; serves as the absolute baseline for non-cooperative utility.
- **Tit-for-Tat (TFT):** Starts with cooperation and subsequently mimics the opponent's previous action.
- **Tit-for-Two-Tats (TF2T):** Cooperates unless the opponent has defected twice in a row; more forgiving than TFT.
- **Two-Tits-for-Tat (2TFT):** Defects twice in response to a single defection by the opponent.
- **Grudger:** Cooperates until the opponent defects once, after which it defects permanently.
- **Suspicious Tit-for-Tat:** Starts with a defection, then mimics the opponent's previous action.
- **Alternator:** Strictly alternates between cooperation and defection.

A.2 Stochastic and Noise-Tolerant

These strategies introduce randomness, challenging the trust agent's ability to distinguish between intentional defection and noise.

- **Random:** Selects cooperation or defection with $p = 0.5$ at each step.
- **Stochastic Tit-for-Tat:** Mimics the opponent's move but occasionally cooperates after a defection or defects after cooperation.
- **Go-By-Majority:** Cooperates if the opponent has cooperated more than or equal to the times they have defected; otherwise, defects.
- **Average Copier:** Picks an action randomly with a probability based on the ratio of the opponent's past actions.

A.3 Probing and Deceptive

Strategies designed to test the limits of an agent's cooperativeness or to manipulate trust.

- **Prober:** Starts with [D, C, C] and defects thereafter if the opponent cooperated in rounds 2 and 3; otherwise, it acts like Tit-for-Tat.
- **Joss:** Acts like Tit-for-Tat but occasionally defects after an opponent cooperates to probe for weakness.
- **Soft Grudger:** Cooperates until defected against; it then responds with [D, D, D, D, C, C] before returning to cooperation.
- **Bully:** The inverse of Tit-for-Tat; defects after cooperation and cooperates after defection.
- **Handshake:** Defects on the first move and cooperates on the second. If the opponent matches this pattern, it cooperates forever; otherwise, it defects forever.

A.4 Evolutionary and Adaptive

These strategies modify their internal state or logic based on the success of past interactions.

- **Pavlov (Win-Stay, Lose-Shift):** Cooperates if both players moved identically in the last round; otherwise, defects. It excels at correcting noise-driven mutual defection.
- **EBF (Evolving Behavioral Factor):** Uses a sliding window to calculate a cooperation threshold, defecting if the opponent's cooperation rate falls below it.
- **Champion:** A complex strategy that uses different phases: testing, cooperation, and retaliation.
- **Desperate:** Becomes increasingly likely to defect as its cumulative score falls behind a preset target.

A.5 Zero-Determinant and Group-Aware

Mathematically derived strategies can enforce a linear relationship between their score and their opponent's.

- **ZD-Extort-2:** Forces the opponent to accept a lower share of the cooperative surplus by maintaining a specific ratio of payoffs.
- **ZD-Gen-2:** A generative ZD strategy that seeks to maximize mutual cooperation while protecting against exploitation.
- **ALLCORR:** A strategy that attempts to detect if it is playing against a clone or a peer strategy through a specific handshake of moves.
- **Self-Recognition:** A group-identifying strategy that only cooperates if the opponent's opening moves identify them as belonging to the same algorithmic class.

B IN-DEPTH SPECIFICATION OF COMPLEX BENCHMARK STRATEGIES

This section details the internal logic of the mathematically sophisticated and adaptive strategies used in the 47-opponent tournament. Figure 3 shows average cumulative wealth of the trust variants across multiple opponent strategies.

B.1 Zero-Determinant Strategies

Zero-Determinant strategies, introduced by Press and Dyson, allow an agent to unilaterally set a linear relationship between its own payoff and the opponent's payoff, regardless of the opponent's strategy.

ZD-Extort-2. This strategy enforces an extortionate relationship where the agent's surplus over the mutual defection payoff (P) is twice that of the opponent's. It achieves this by selecting a move probability vector $\mathbf{p} = [p_1, p_2, p_3, p_4]$ corresponding to the outcomes (CC, CD, DC, DD). By setting these probabilities precisely, the agent ensures:

$$(S_{\text{agent}} - P) = \chi(S_{\text{opponent}} - P),$$

where $\chi = 2$. Against an opponent seeking to maximize its own score, the only rational response is to cooperate, though it receives a smaller share of the surplus.

B.2 Adaptive and Learning Strategies

Unlike deterministic rules, these strategies modify their behavior based on the perceived state of the environment or the opponent's historical performance.

Pavlov (Win-Stay, Lose-Shift). Pavlov is a reflexive learner who uses the payoff from the previous round as a reinforcement signal. It cooperates if both players choose the same action in the last round and defects otherwise.

- **Win:** If the payoff was R or T, it repeats the previous move.
- **Loss:** If the payoff was P or S, it switches moves.

This is particularly effective at correcting accidental defections in noisy environments, as mutual defection triggers a switch back toward mutual cooperation.

Evolving Behavioral Factor. EBF maintains a dynamic internal variable, the Behavioral Factor (BF), which represents the agent's current willingness to cooperate. It calculates a threshold based on the opponent's historical cooperation rate:

$$\text{Threshold} = \frac{\sum C_{\text{opp}}}{N} \times \omega,$$

where ω is an evolving weight. The agent defects if its internal BF falls below the calculated threshold. This allows the agent to shift from highly altruistic to strictly defensive based on the perceived social climate of the interaction.

B.3 Probing and Meta-Strategies

These strategies attempt to fingerprint the agent's logic through specific sequences of actions.

Soft Grudger. Soft Grudger is a modulated retaliator. Unlike the standard *Grudger*, which defects forever after a single betrayal, Soft Grudger employs a cooling-off period. Upon detecting a defection, it executes the sequence $[D, D, D, D, C, C]$.

- The four defections punish the opponent and minimize the exploiter's gain.
- The two cooperations signal a willingness to return to a CC equilibrium.

This prevents the permanent lock-in of mutual defection that renders the standard Grudger inefficient.

Champion. Champion is a composite strategy that operates in three distinct phases:

(1) **Testing Phase:** The first few moves are used to categorize the opponent as cooperative, random, or adversarial.
(2) **Cooperative Phase:** If the opponent is categorized as non-adversarial, it plays Tit-for-Tat to mirror cooperation.
(3) **Retaliatory Phase:** If the opponent defects more than a specific percentage, Champion shifts to a more aggressive defection-heavy policy to protect its cumulative wealth.

B.4 Group-Aware and Identity Strategies

Handshake and ALLCORR.. These strategies use the first n rounds to transmit a code. For example, *Handshake* defects on move 1 and cooperates on move 2. If the opponent responds with the exact same sequence, the strategy recognizes the opponent as a clone and cooperates indefinitely. If the code is not matched, it defaults to *Always Defect*. This allows a sub-population of similar agents to maximize their own social welfare while aggressively excluding outsiders.

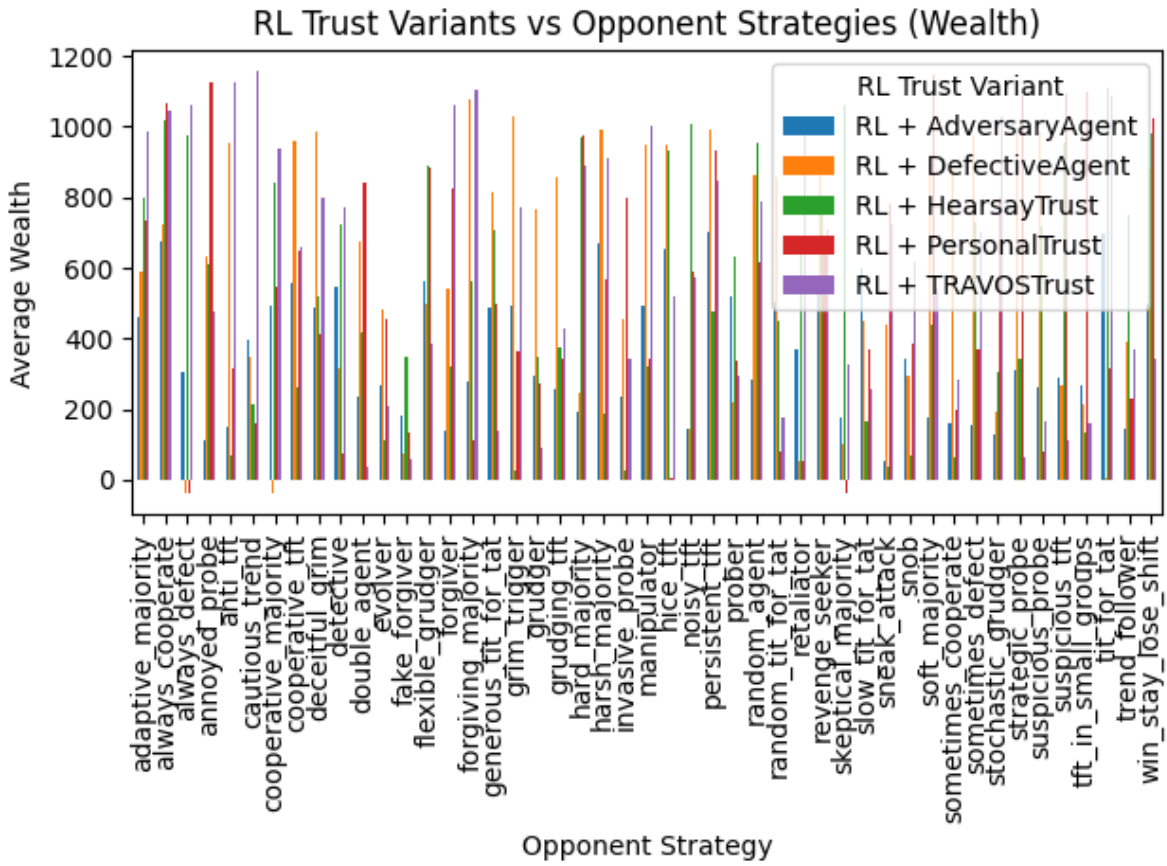

Figure 3: Average cumulative wealth of the trust variants across multiple opponent strategies. Hybrid models with GNN and MCTS+UCT achieve superior cumulative wealth and stability. Error bars denote variance in cumulative wealth.