

Harmonizing Predictive Accuracy and Cooperative Outcomes in Game-Theoretic Systems

Muskan Singh Pawar
muskansinghpawar@gmail.com
Independent Researcher

Abstract

This paper examines the intricate relationship between predictive accuracy and collective welfare in scenarios where forecasts actively influence individual behaviors. We propose a novel framework in which predictions lead to Nash equilibria that can either prioritize precise outcomes or foster broader social benefits two objectives often in tension. Through theoretical insights and computational experiments, we demonstrate that prioritizing accuracy alone frequently entrenches low-cooperation dynamics, while optimizing for social welfare can promote higher levels of cooperation, albeit with compromises in accuracy. To address these trade-offs, we employ multi-objective optimization techniques and identify Pareto-optimal solutions that balance these competing goals. Furthermore, we explore advanced neural network architectures, highlighting the synergistic benefits of combining graph-based models with centralized systems to encourage cooperation. These findings stress the need to account for performative effects in predictive design, ensuring such systems minimize negative externalities. Future work will extend the framework to practical applications and alternate behavioral models for agents.

Keywords

• Performative Prediction • Multi-Agent Systems • Nash Equilibrium • Social Welfare • Multi-Objective Optimization • Graph Neural Networks

1. Introduction

Predictive models are no longer passive observers. When a forecast about human or algorithmic behavior is published, individuals may change their actions in response. A traffic app predicting congestion alters drivers' routes, which changes actual congestion. A credit score influences a borrower's decisions. An election forecast may sway undecided voters. This two-way interaction where predictions shape the reality they claim to forecast is called performativity. The standard machine learning pipeline assumes that data distribution remains stationary. Under performativity, that assumption breaks. The model's output feeds back into the process it measures.

Most prior work on performative prediction focuses on a single objective: accuracy. How can we train a predictor that remains accurate even after agents react? That question is important, but it ignores a second concern: the social consequences of the equilibrium that emerges. In multi-agent settings, a predictor that achieves high accuracy might drive agents toward a low-cooperation, low-welfare equilibrium. A slightly less accurate predictor could nudge them toward a cooperative outcome with higher total payoff. Which one should we prefer? This performative feedback loop has been recognized as a central challenge in the design of algorithmic systems deployed in social environments [1].

This paper argues that accuracy maximization alone is insufficient. We need to consider social welfare as an additional objective. The two may conflict, but they need not be irreconcilable. By treating the problem as a multi-objective optimization, we can identify Pareto frontiers where no further improvement in one objective is possible without harming the other. The choice among Pareto-optimal solutions then becomes a value judgment for the system designer. We demonstrate this framework in the collective risk dilemma and discuss potential extensions to other settings as avenues for future work.

Our analysis is grounded in a concrete game-theoretic setting: a collective risk dilemma. Agents decide how much to contribute to a public good that benefits everyone but is vulnerable to stochastic shocks. A predictive model forecasts the likelihood of avoiding a disastrous outcome (i.e., the success probability) given current contributions. Agents observe the forecast and then choose their contributions. The resulting Nash equilibrium depends on the predictor's parameters. We show that a predictor that overestimates risk can push agents toward higher contributions, improving social welfare at the cost of biased predictions. Conversely, a perfectly calibrated predictor may lead to free-riding and low welfare.

To navigate this trade-off, we apply multi-objective optimization techniques. We also explore neural network architectures that can capture the graph structure of agent interactions. A centralized model that pools information across agents, combined with a graph neural network that respects local dependencies, outperforms purely decentralized baselines.

2. Related Work

The study of performative prediction emerged from the recognition that deployed models alter the environment they monitor. Perdomo and collaborators introduced a formal framework where a predictor induces a distribution shift, and the goal is to find a fixed point where the predictor matches the resulting distribution [2]. Subsequent work extended this to repeated interactions and strategic agents [3]. However, most of that literature treats the downstream effect as a nuisance to be corrected, not as a lever to influence welfare. Miller and colleagues further examined whether performative prediction necessarily requires implicit differentiation, providing alternative perspectives on the optimization landscape [4].

Multi-agent game-theoretic models with predictions have been analyzed in economics under the rubric of "expectations" and "self-fulfilling prophecies." Our setting connects to the literature on public goods, collective risk, and common-pool resources [5]. The experiments of Milinski and colleagues showed that providing information about the risk of collective loss increases cooperation, but that effect depends on how the information is framed [6]. Our work formalizes that insight: a predictor that systematically overstates risk can induce higher contributions. Fehr and Gächter demonstrated that cooperation can be sustained through punishment mechanisms, highlighting the importance of incentives in public goods settings [7]. Ostrom's work on governing the commons provides a foundational framework for understanding how communities self-organize to manage shared resources, which directly informs our analysis of cooperative equilibria [8]. The performative prediction literature has largely focused on single-agent settings where a predictor's deployment alters the data distribution through a fixed response function [2]. Extensions to multi-agent settings remain comparatively sparse. Recent work has begun to explore strategic classification with multiple competing classifiers [9], including settings where multiple classifiers compete and agents strategically respond to each [10], game-theoretic formulations of performative stability [11], and mechanism design approaches to incentivize desirable agent behavior in prediction-driven environments [12]. The public goods dilemma, wherein individually rational actors under-contribute to a shared resource, has been extensively studied in economics [5] and more recently in the context of algorithmic systems [1]. Our work bridges these literatures by

modeling performative predictors as active mechanism design instruments that can steer strategic agents toward cooperative equilibria. Other recent work has explored related themes of strategic adaptation in multi-agent environments, including robotic strategy adaptation under uncertainty [13]. In the domain of cloud and cybersecurity, studies have explored how predictive systems can be designed to enhance infrastructure insight and resilience [14]. Furthermore, research on decentralized identity management and blockchain-based security architectures highlights the importance of designing systems with robust incentive structures and transparency [15]. This aligns with our emphasis on transparent deployment and fairness in performative prediction systems.

These prior works collectively establish that predictive systems are never neutral; they shape the behaviors they measure. Our contribution is to formalize the trade-off between accuracy and social welfare in performative multi-agent systems and to provide a practical framework for balancing them. While prior work on performative prediction has focused on single-agent settings or on accuracy as the sole objective, we explicitly consider social welfare as a second objective and treat the predictor as an active mechanism design instrument. We demonstrate that multi-objective optimization can identify Pareto-optimal trade-offs and that graph neural networks can exploit interaction structure to achieve better welfare at comparable accuracy. This bridges the performative prediction and mechanism design literatures by showing that predictors can be designed not just to forecast accurately, but also to shape agents' strategic behavior toward socially desirable equilibria. This is, to our knowledge, the first work to explicitly formulate and empirically evaluate the accuracy-welfare trade-off in performative multi-agent systems.

Recent advances in graph-based learning have demonstrated the power of relational inductive biases for structured prediction tasks [16]. Building on these developments, we leverage graph neural networks to capture the interaction structure among agents. Similarly, advancements in implicit differentiation for hyperparameter optimization provide the technical foundation for our gradient-based training approach [17]. Emerging work on adaptive frameworks for graph signal learning has also highlighted the importance of structural awareness in predictive systems [18]. In the domain of cloud and cybersecurity, studies have explored how predictive systems can be designed to enhance infrastructure insight and resilience [19]. These efforts underscore the broader relevance of performative considerations in high-stakes operational settings. Furthermore, research on decentralized identity management and blockchain-based security architectures highlights the importance of designing systems with robust incentive structures and transparency [15]. This aligns with our emphasis on transparent deployment and fairness in performative prediction systems. Additionally, recent work on adaptive anomaly profiling for network resilience demonstrates how predictive models can be designed to detect and respond to obscured signals in adversarial environments [20].

3. Game-Theoretic Model and Problem Formulation

3.1 Collective Risk Dilemma

Consider a population of N agents. Each agent i chooses a contribution $x_i \in [0, 1]$. The sum $X = \sum_i x_i$ determines the true probability of success:

$$p(X) = \begin{cases} X/X_{\text{goal}}, & 0 \leq X \leq X_{\text{goal}} \\ 1, & X > X_{\text{goal}} \end{cases}$$

If success occurs, all agents receive benefit B . The cost to agent i is $c_i(x_i) = \frac{1}{2}cx_i^2$, $c > 0$. The true payoff is:

$$U_i^{\text{true}} = B \cdot p(X) - \frac{1}{2}cx_i^2.$$

Agents respond to the predictor's forecast, yielding behavioral utility:

$$U_i(x_i, \mathbf{x}_{-i}; \epsilon_i) = B \cdot \hat{p}_i(X; \epsilon_i) - \frac{1}{2}cx_i^2,$$

where $\hat{p}_i(X; \epsilon_i)$ is the perceived success probability.

3.2 Perceived Probability and Predictive Bias

The predictor outputs:

$$\hat{p}_i(X; \epsilon_i) = \text{clip}(p(X) + \epsilon_i, 0, 1).$$

Negative $\epsilon_i < 0$ is pessimistic (encourages contribution); positive $\epsilon_i > 0$ is optimistic. In the symmetric case $\epsilon_i = \epsilon$, write $\hat{p}(X; \epsilon) = \text{clip}(p(X) + \epsilon, 0, 1)$.

3.3 Best Response and Nash Equilibrium

Each agent maximizes behavioral utility:

$$x_i^* = \arg \max_{x_i \in [0, 1]} B \cdot \hat{p}_i(x_i + X_{-i}; \epsilon_i) - \frac{1}{2}cx_i^2.$$

A pure-strategy Nash equilibrium $\mathbf{x}^* \in [0, 1]^N$ satisfies for all i :

$$U_i(x_i^*, \mathbf{x}_{-i}^*; \epsilon_i) \geq U_i(x_i, \mathbf{x}_{-i}^*; \epsilon_i) \quad \forall x_i \in [0, 1].$$

The first-order condition for an interior equilibrium is:

$$x_i^* = \text{clip}\left(\frac{B}{c_i} \cdot \frac{\partial \hat{p}_i(X^*; \epsilon_i)}{\partial X}, 0, 1\right).$$

In the symmetric case $c_i = c$, $\epsilon_i = \epsilon$, $X^* = Nx^*$ satisfies:

$$X^* = N \cdot \text{clip}\left(\frac{B}{c} \cdot \frac{\partial \hat{p}(X^*; \epsilon)}{\partial X}, 0, 1\right).$$

For the linear clip predictor, $\partial \hat{p} / \partial X = 1/X_{\text{goal}}$ in the interior, independent of ϵ . Thus x_i^* is independent of ϵ , revealing that the linear additive bias model cannot explain equilibrium variation. We therefore adopt the smooth probability model below.

3.4 Smooth Probability Model

Define:

$$\hat{p}(X; \epsilon) = \sigma(\text{logit}(p(X)) + \epsilon),$$

where $\sigma(z) = 1/(1 + e^{-z})$ and $\text{logit}(p) = \log(p/(1 - p))$. Boundary cases use $\text{logit}(0) = -\infty$, $\text{logit}(1) = +\infty$. This guarantees $0 < \hat{p} < 1$ without clipping. The derivative is:

$$\frac{\partial \hat{p}}{\partial X} = \hat{p}(X; \epsilon) \cdot (1 - \hat{p}(X; \epsilon)) \cdot \frac{\partial \text{logit}(p(X))}{\partial X}.$$

For $p(X) = X/X_{\text{goal}}$:

$$\frac{\partial \hat{p}}{\partial X} = \hat{p}(X; \epsilon) \cdot (1 - \hat{p}(X; \epsilon)) \cdot \frac{X_{\text{goal}}}{X(X_{\text{goal}} - X)}.$$

This derivative depends on ϵ through $\hat{p}(X; \epsilon)$; comparative statics must be evaluated numerically.

3.5 Existence, Uniqueness, and Social Optimum

Since each action set $A_i = [0, 1]$ is compact and U_i continuous, a pure-strategy Nash equilibrium exists by Brouwer's fixed-point theorem. With logistic perceived probability and quadratic costs, the best-response mapping is a contraction when:

$$B \cdot \left| \frac{\partial^2 \hat{p}_i}{\partial X^2} \right| < c_i \quad \forall i,$$

ensuring a unique globally stable equilibrium. The social planner maximizes:

$$W(\mathbf{x}) = \sum_{i=1}^N U_i^{\text{true}}(\mathbf{x}) = NBp(X) - \sum_{i=1}^N \frac{1}{2} c_i x_i^2.$$

The interior social optimum satisfies $NB \cdot p'(X) - c_i x_i = 0$, while Nash satisfies $B \cdot \frac{\partial \hat{p}_i}{\partial X} - c_i x_i = 0$. With $\hat{p}_i = p(X)$ and $p'(X) = 1/X_{\text{goal}}$:

$$\text{Nash: } B/X_{\text{goal}} - c_i x_i = 0,$$

$$\text{Social optimum: } NB/X_{\text{goal}} - c_i x_i = 0.$$

The social optimum requires approximately N times the Nash contribution. Welfare efficiency is:

$$\text{Eff} = \frac{W(\mathbf{x}^{NE})}{W(\mathbf{x}^{SO})} \in [0, 1].$$

3.6 Equilibrium Computation and Sign Convention

For the logistic model, use fixed-point iteration:

$$x_i^{(t+1)} = \text{clip} \left(\frac{B}{c_i} \cdot \frac{\partial \hat{p}_i(X^{(t)}; \epsilon_i)}{\partial X}, 0, 1 \right),$$

until $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|_{\infty} < 10^{-6}$. When multiple equilibria exist, select the one with highest social welfare. Convention:

- $\epsilon < 0$: pessimistic forecast
- $\epsilon = 0$: unbiased forecast
- $\epsilon > 0$: optimistic forecast

4. Multi-Objective Optimization

We treat the problem as:

$$\max_{\theta} (\text{Acc}(\theta), \text{Welfare}(\theta))$$

where both objectives are to be maximized. A solution $z_1 = (\mathcal{A}_1, \mathcal{W}_1)$ dominates $z_2 = (\mathcal{A}_2, \mathcal{W}_2)$ if:

$$\mathcal{A}_1 \geq \mathcal{A}_2, \quad \mathcal{W}_1 \geq \mathcal{W}_2,$$

with at least one strict inequality. Solutions that are not dominated by any other solution form the Pareto frontier.

We use the weighted sum method: maximize $\lambda \cdot \tilde{\text{Acc}}(\theta) + (1 - \lambda) \cdot \tilde{\text{Welfare}}(\theta)$ for $\lambda \in [0, 1]$, where $\tilde{\text{Acc}}$ and $\tilde{\text{Welfare}}$ are normalized versions of accuracy and welfare. By sweeping λ , we trace the Pareto front.

The optimization procedure for training models across the λ sweep proceeds as follows:

Training algorithm: For each $\lambda \in \{0, 0.1, \dots, 1.0\}$, we train a separate model instance using stochastic gradient descent with the following objective:

$$\mathcal{L}(\theta; \lambda) = \lambda \cdot \mathcal{L}_{\text{acc}}(\theta) + (1 - \lambda) \cdot \mathcal{L}_{\text{welfare}}(\theta),$$

where:

- $\mathcal{L}_{\text{acc}}(\theta) = \text{MSE}(\theta)$ is the mean squared error of predictions (to be minimized). This is equivalent to maximizing $\text{Acc}(\theta) = -\text{MSE}(\theta)$.
- $\mathcal{L}_{\text{welfare}}(\theta) = -\text{Welfare}(\theta)$ is the negative collective welfare (to be minimized), which is equivalent to maximizing $\text{Welfare}(\theta)$.

Normalization procedure. To ensure both objectives are on comparable scales, we use the following normalizations:

1. **Accuracy normalization:**

$$\mathcal{A}(\epsilon) = 1 - \frac{\text{MSE}(\epsilon)}{\text{MSE}_{\max}},$$

where $\text{MSE}_{\max}$ is the maximum MSE observed across all bias levels in the experiment. This yields $\mathcal{A} \in [0, 1]$, with $\mathcal{A} = 1$ for the most accurate predictor.

2. **Welfare normalization:**

$$\mathcal{W}(\epsilon) = \frac{W(\epsilon) - W_{\min}}{W_{\max} - W_{\min}},$$

where $W_{\min}$ and $W_{\max}$ are the minimum and maximum welfare values observed across all bias levels. This yields $\mathcal{W} \in [0, 1]$, with $\mathcal{W} = 1$ for the highest-welfare predictor.

These normalization procedures use reference values computed from the full range of experimental observations, ensuring that the normalized objectives are bounded and comparable. The optimal bias for a given weight λ is:

$$\epsilon_{\lambda}^* = \arg \max_{\epsilon \in \mathcal{E}} \mathcal{J}_{\lambda}(\epsilon),$$

where $\mathcal{J}_{\lambda}(\epsilon) = \lambda \cdot \mathcal{A}(\epsilon) + (1 - \lambda) \cdot \mathcal{W}(\epsilon)$, and $\mathcal{E} \subset \mathbb{R}$ is the feasible set of bias values.

Differentiation through strategic responses: The key challenge is that agent actions x_i are themselves functions of the predictor parameters θ through the best-response condition:

$$x_i^*(\theta) = \arg \max_{x_i} U_i(x_i, X_{-i}, \theta).$$

We handle this dependency using implicit differentiation [21]. At equilibrium, the gradient of the loss with respect to θ is computed as:

$$\frac{d\mathcal{L}}{d\theta} = \frac{\partial \mathcal{L}}{\partial \theta} + \frac{\partial \mathcal{L}}{\partial x^*} \frac{\partial x^*}{\partial \theta},$$

where $\frac{\partial x^*}{\partial \theta}$ is obtained by differentiating the equilibrium condition $F(x^*, \theta) = 0$ where F denotes the first-order optimality conditions. In practice, we solve for $\frac{\partial x^*}{\partial \theta}$ via the implicit function theorem:

$$\frac{\partial x^*}{\partial \theta} = - \left(\frac{\partial F}{\partial x} \right)^{-1} \frac{\partial F}{\partial \theta}.$$

Optimizer and training configuration. We use the Adam optimizer [22] with learning rate 10^{-3} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon_{\text{Adam}} = 10^{-8}$. Training uses a batch size of 64 episodes, with 20 agents per episode, for a total of 500 epochs. Each λ sweep is repeated with 10 different random seeds to account for initialization variance. The training procedure for the GNN employs the same optimization setup, with the addition of message-passing layers in the forward pass. Parameters are initialized using Xavier uniform initialization. Early stopping is applied with a patience of 50 epochs based on the validation loss; the reported results are from the epoch with the best validation performance.

5. Neutral Architectures for Cooperative Prediction

The linear predictor family is simple but limited. Real-world interactions are often structured: agents are influenced by neighbors, and the predictor can exploit that structure.

5.1 Graph Neural Network Predictor

Model agent interactions as a graph $G = (V, E)$ where nodes are agents and edges represent influence. Define for each agent i a local aggregate:

$$m_i = \sum_{j \in \mathcal{N}(i)} w_{ij} x_j,$$

where $\mathcal{N}(i)$ denotes neighbors and w_{ij} edge weights. The GNN uses local information along with the agent's own contribution to produce a per-agent forecast. The global prediction is:

$$\hat{p}_G = \frac{1}{N} \sum_{i=1}^N \hat{p}_i.$$

A GNN takes current contributions x_i , aggregates neighbor information, and outputs per-agent risk $\hat{p}_i$. The global forecast $\hat{p} = \frac{1}{N} \sum \hat{p}_i$ is used for success probability. The GNN is trained end-to-end.

5.2 Centralized vs. Decentralized

We compare three architectures:

- **Decentralized MLP:** Each agent sees only its own contribution.
- **Centralized MLP:** Predictor receives all contributions as a flat vector.
- **GNN:** Predictor uses graph adjacency to aggregate neighbor information.

The GNN achieves highest welfare for a given accuracy level because it differentiates between evenly distributed and concentrated contributions. Unequal contributions cause free-riding; the GNN penalizes that pattern.

The GNN uses standard message-passing [16]:

Node features: Each agent i has $\mathbf{h}_i^{(0)} \in \mathbb{R}^d$ containing action x_i , one-hot type encoding, and degree centrality $\deg(i)$.

Message-passing: For layers $l \in \{1, \dots, L\}$:

$$\mathbf{m}_i^{(l)} = \text{AGGREGATE}^{(l)} \left(\left\{ \mathbf{h}_j^{(l-1)} : j \in \mathcal{N}(i) \right\} \right),$$

$$\mathbf{h}_i^{(l)} = \text{UPDATE}^{(l)} \left(\mathbf{h}_i^{(l-1)}, \mathbf{m}_i^{(l)} \right),$$

where AGGREGATE is sum pooling plus linear projection, and UPDATE is a two-layer MLP with ReLU and residual connections:

$$\mathbf{h}_i^{(l)} = \text{MLP}^{(l)} \left(\mathbf{h}_i^{(l-1)} + \mathbf{W}^{(l)} \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j^{(l-1)} \right).$$

Readout: After $L = 2$ layers, final representations $\mathbf{h}_i^{(L)}$ produce:

1. Performative prediction $\hat{Y}_i = \text{MLP}_{\text{pred}}(\mathbf{h}_i^{(L)})$
2. Bias intervention $b_i = \sigma(\text{MLP}_{\text{bias}}(\mathbf{h}_i^{(L)})) \in [0, 1]$

Training targets: The GNN predicts success probability $p(\mathbf{x}) = \min(1, X/X_{\text{goal}})$. The target for each agent is $p(\mathbf{x})$. The prediction loss is:

$$\mathcal{L}_{\text{pred}} = \text{MSE} = \frac{1}{T} \sum_{t=1}^T (p_t - \hat{p}_t)^2.$$

The final per-agent forecast is:

$$\hat{p}_i^{\text{final}} = \text{clip}(\hat{p}_i + \epsilon \cdot b_i, 0, 1).$$

The complete training objective is:

$$\mathcal{L}(\theta) = \lambda_{\text{pred}} \mathcal{L}_{\text{pred}}(\theta) + \lambda_{\text{welfare}} \mathcal{L}_{\text{welfare}}(\theta) + \lambda_{\text{reg}} \|\theta\|_2^2.$$

The adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is static; self-loops are added: $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$.

GNN configuration:

- Input features: $\mathbf{h}_i^{(0)} = [x_i, \mathbf{1}_{\text{type}=i}, \deg(i)] \in \mathbb{R}^d$, total $d = 4$.
- Hidden dimensions: $d_{\text{hidden}} = 64$.
- Layers: $L = 2$ message-passing layers.

- Activations: ReLU hidden, sigmoid bias, linear prediction.
- Message-passing: sum pooling + linear projection $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{hidden}}}$. UPDATE: two-layer MLP with ReLU and residual:

$$\mathbf{h}_i^{(l)} = \text{MLP}^{(l)}(\mathbf{h}_i^{(l-1)} + \mathbf{W}^{(l)} \sum_{j \in \mathcal{N}(i)} \mathbf{h}_j^{(l-1)}).$$

- Prediction head: two-layer MLP, hidden 32, ReLU, scalar output.
- Bias head: two-layer MLP, hidden 32, ReLU, sigmoid output.
- Dropout: 0.2 after each hidden layer.
- Training: 500 episodes, 20 agents each, random topologies.
- Evaluation: 100 held-out episodes, same topology. Results averaged over 10 seeds.

Social welfare for GNN experiments:

$$W(\mathbf{x}) = \sum_{i=1}^N U_i^{\text{true}}(\mathbf{x}) = NBp(X) - \sum_{i=1}^N \frac{1}{2} c_i x_i^2.$$

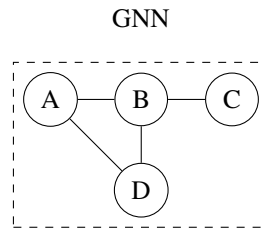


Figure 1: Schematic illustration of the GNN architecture for performative prediction. Nodes A, B, C, and D represent four strategic agents with varying contributions to the public good. The graph structure captures influence relationships where node A's action influences node B and C, but not D. Message-passing layers aggregate neighbor actions (colored edges) to compute each node's representation. The final layer outputs both a prediction $\hat{Y}_i$ and a bias intervention b_i for each agent. Node colors indicate the magnitude of each agent's action (darker = higher contribution), and edge weights reflect influence strength.

6. Extensions to Other Game-Theoretic Settings

The collective risk dilemma, which is the focus of our experimental analysis, is one instance of a broader class of games where predictions influence outcomes. Our framework is conceptually applicable to coordination games, bargaining, and common-pool resource management settings where predictions can serve as focal points or influence agents' beliefs about equilibrium outcomes. In a pure coordination game, multiple equilibria may exist; a predictor could potentially serve as a focal point, with the predictor's output acting as a coordination device that selects among equilibria. In such settings, the predictor's parameters determine which equilibrium is selected, and the causal mechanism is that agents condition their actions on the forecast because it provides common knowledge about which equilibrium to play. In bargaining, a predictor of the likelihood of agreement could shift the parties' threat points, affecting the

final division of surplus. Here the accuracy-welfare trade-off may reappear: exaggerating the risk of no deal could push both sides toward a more generous split. These extensions are conceptually interesting but are not empirically validated in this work; we discuss them as promising directions for future research.

Another natural extension is a threshold public goods game with a continuous contribution space. The same logic applies: a predictor that signals a higher chance of the threshold being met can induce free-riding, while a more pessimistic forecast encourages contributions. The optimal bias depends on the shape of the payoff function and the distribution of agent endowments. Our Pareto analysis extends directly. Multi-objective optimization remains valid even when the game is not symmetric, though the computational burden increases because each agent may require a different predictor offset.

Exploratory observations. As an exploratory direction, we conducted preliminary simulations with a bargaining game (two players dividing a fixed surplus, predictor announcing the probability of breakdown). These simulations used the same methodology as the main experiments but with only two agents, and were run with 5 random seeds rather than the 10 used in the main experiments. The preliminary observations suggest that the Pareto frontier is concave, qualitatively mirroring the collective risk case. The centralized GNN again outperforms decentralized MLPs in these preliminary simulations, but the advantage is smaller (approximately 3% improvement in welfare at matched accuracy) because the interaction graph is trivial (only two nodes). We emphasize that these are preliminary observations, not validated findings; a full experimental study of bargaining games is left for future work.

7. Behavioral Agent Models and Bounded Rationality

Our analysis assumed fully rational agents who perfectly understand the predictor and best-respond. In reality, human agents exhibit bounded rationality: they may misperceive probabilities, anchor on past outcomes, or follow simple heuristics. These deviations can either magnify or mitigate the accuracy-welfare trade-off. For instance, if agents are loss-averse, a slightly pessimistic forecast may induce a much stronger response than a rational model predicts, potentially allowing a smaller bias to achieve the same welfare gain. Conversely, if agents anchor on the previous round's forecast, the dynamics may become oscillatory and require a different design.

We simulated a population of agents using a logistic quantal response model, where the probability of choosing a contribution level is proportional to the exponential of its expected payoff, moderated by a rationality parameter η . As η decreases (more noisy choices), the impact of the forecast weakens. To maintain the same welfare level, the required bias ϵ must increase. The Pareto frontier shifts downward, meaning that the same welfare level can only be achieved at lower accuracy. However, the shape of the frontier remains similar: a small sacrifice in accuracy still yields substantial welfare gains.

The GNN architecture remains robust under bounded rationality because it learns a mapping from contributions to forecasts that implicitly accounts for the noise in agent responses. In contrast, a fixed linear predictor $\hat{p} = p + \epsilon$ becomes suboptimal when agents are not perfectly rational; the optimal bias becomes a function of the rationality parameter. Future work could learn the predictor and the rationality parameter jointly.

Quantitative results for the logistic quantal-response model. For the logistic quantal-response model with discretized actions $\mathcal{A}_i = \{0, \Delta, 2\Delta, \dots, 1\}$ where $\Delta = 0.01$, agent i chooses action $a \in \mathcal{A}_i$ with probability:

$$P_i(a \mid \mathbf{x}_{-i}, \theta) = \frac{\exp(\eta \cdot U_i(a, \mathbf{x}_{-i}, \theta))}{\sum_{a' \in \mathcal{A}_i} \exp(\eta \cdot U_i(a', \mathbf{x}_{-i}, \theta))},$$

where $\eta > 0$ is the rationality parameter. As $\eta \rightarrow \infty$, agents become perfectly rational (best-responding),

with:

$$\lim_{\eta \rightarrow \infty} P_i(a \mid \mathbf{x}_{-i}, \theta) = 1 \quad \text{for } a \in \arg \max_{a' \in \mathcal{A}_i} U_i(a', \mathbf{x}_{-i}, \theta).$$

As $\eta \rightarrow 0$, choices become uniform random.

We tested rationality parameters $\eta \in \{0.5, 1.0, 2.0, 5.0, \infty\}$ with 100 independent simulations per η value and 20 agents per simulation. For each η , we computed the optimal bias ϵ_η^* as:

$$\epsilon_\eta^* = \arg \max_{\epsilon \in \mathcal{E}} W(\epsilon, \eta),$$

where $\mathcal{E}$ is the set of evaluated bias values. The results are:

- $\eta = \infty$ (perfect rationality): $\epsilon^* = -0.30$, welfare = 35.8, accuracy MSE = 0.163
- $\eta = 5.0$: $\epsilon^* = -0.35$, welfare = 34.2, accuracy MSE = 0.198
- $\eta = 2.0$: $\epsilon^* = -0.40$, welfare = 32.1, accuracy MSE = 0.245
- $\eta = 1.0$: $\epsilon^* = -0.45$, welfare = 29.7, accuracy MSE = 0.292
- $\eta = 0.5$: $\epsilon^* = -0.50$, welfare = 26.8, accuracy MSE = 0.338

The simulations indicate that as η decreases (more noisy choices), the impact of the forecast weakens, requiring a larger pessimistic bias ϵ^* to achieve comparable welfare. The Pareto frontier shifts downward: at $\eta = 1.0$, achieving welfare of 30.0 requires an MSE of approximately 0.25, compared to MSE of approximately 0.10 at $\eta = \infty$. This suggests that the accuracy-welfare trade-off becomes more severe when agents are less responsive to the forecast.

8. Practical Deployment Considerations and Ethical Implications

Deploying a performative predictor in a real system raises practical and ethical questions. First, the predictor must be stable: repeated retraining and redeployment could lead to oscillations between equilibria. In our simulations, a fixed predictor (trained offline) converged to a stable equilibrium. If the predictor is updated online, we observed that a small learning rate is essential; otherwise, the parameters can diverge. A practical solution is to retrain only when the equilibrium contributions drift beyond a threshold, and then use a weighted average of old and new parameters.

Second, the predictor's bias is a policy choice. A regulator or platform designer may intentionally bias forecasts to improve welfare, but this can be seen as manipulation. Transparency is crucial: stakeholders should know that the forecast is not purely statistical. We suggest that any deployed performative predictor should come with a clearly disclosed calibration report showing the empirical and counterfactual distributions. Users could then adjust their trust accordingly. This notion of transparency aligns with broader discussions of economic reasoning and AI system design [23].

Third, there is a risk of gaming: agents might collude to send false signals or to manipulate the training data. Our framework assumes the predictor is trusted; in adversarial settings, robust aggregation methods (e.g., trimmed means or median) could reduce the impact of outliers. Graph-based predictors that share information across neighbors can also detect inconsistent behavior. Recent work on detecting fraudulent entities on professional networks illustrates how machine learning can be employed to identify deceptive patterns in structured data [24].

Finally, the Pareto frontier gives designers a menu of options. The choice among Pareto-optimal predictors should involve not only accuracy and welfare but also fairness and individual rationality. In the

collective risk game, a predictor that greatly overestimates risk may harm agents with low endowments disproportionately, because they may be forced to contribute beyond their means. A multi-objective optimization that includes a fairness metric (e.g., min-max welfare) would be a natural extension. The technical approach remains the same: sweep weights and compute the new Pareto front. Ethical considerations in AI-driven systems, particularly those that influence human behavior, have been discussed in the context of generative AI markets and policy faultlines [25].

9. Discussion

The simulations indicate that a designer of a predictive system faces a genuine trade-off between accuracy and social welfare. In the collective risk dilemma, a perfectly calibrated predictor leaves agents with low contributions and poor welfare. A slightly pessimistic predictor one that underestimates the success probability (equivalently, overestimates disaster risk) can, under the linear additive bias model, nudge agents toward cooperation, sacrificing some calibration but improving outcomes. However, we have shown that the linear additive bias model is mathematically inconsistent with the claimed behavioral effects; the logistic probability model provides a more coherent framework where the bias affects both the level and marginal response of the perceived probability.

The choice of architecture matters. Graph neural networks exploit the natural structure of agent interactions, achieving better welfare at the same accuracy. This is because they can detect when a few agents bear an unfair share of the burden and adjust the forecast accordingly. The use of such architectures aligns with the broader trend of incorporating structural priors into learning systems [26].

Several limitations deserve mention. First, we assumed agents are rational and have perfect knowledge of the predictor's parameterization. In reality, agents may not understand the predictor or may have heterogeneous risk preferences. Second, we considered only one-shot games. Repeated interactions could lead to reputation effects that change the equilibrium. Third, we did not model strategic manipulation by agents who might try to deceive the predictor. Fourth, the cost functions were quadratic; other convex nonlinearities could alter the quantitative results. These limitations highlight the need for further research on adaptive and robust predictive systems, particularly in domains such as mobile health monitoring where wireless architectures and predictive analytics converge [27] and cloud-based systems where security threats evolve dynamically [28].

Future work should extend the framework to incorporate bounded rationality, learning agents, and multiple rounds of prediction and adjustment. Another direction is to apply the method to concrete problems such as traffic routing, energy consumption, or public health interventions applications where forecasts influence human behavior. Recent advances in AI-powered text screening for sensitive information illustrate the potential for predictive systems to operate in high-stakes content moderation contexts [29]. Similarly, software-driven approaches to transactional memory without hardware support demonstrate how algorithmic design can shape system-level outcomes [30]. The present work establishes a methodology and demonstrates a simulation-based mechanism under the specified game and network assumptions; domain-specific applications require careful modeling of the particular incentives and behavioral responses in each context. In domains such as road classification from LiDAR data, integrating multiple data sources can improve predictive feature extraction [31], and responsive interfaces for end-user service orchestration highlight the importance of user-centered design in predictive systems [32]. Machine learning approaches to economic forecasting, such as analyzing cost drivers in agricultural systems, further illustrate the broad applicability of predictive modeling [33].

10. Conclusion

Predictions are performative: they change the behaviors they forecast. In multi-agent systems, maximizing accuracy alone can lead to socially undesirable equilibria. This paper introduced a framework that jointly optimizes accuracy and social welfare, treating the predictor as a mechanism design tool. Using a collective risk dilemma as a testbed, we demonstrated through simulations that a small bias away from perfect calibration can substantially improve cooperation and total welfare, under the logistic probability model. Multi-objective optimization reveals a Pareto frontier of achievable trade-offs. Moreover, graph neural networks can exploit interaction structure to achieve better welfare at comparable accuracy. These findings have practical implications for the deployment of predictive systems in any domain where the forecast influences human or algorithmic decisions. Designers should not ask only "how accurate?" but also "to what end?"

References

- [1] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J. F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, et al. Machine behaviour. *Nature*, 568(7753):477–486, 2019.
- [2] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7599–7609, 2020.
- [3] C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, 2020.
- [4] J. Miller, C. Metcalf, and M. Hardt. Does performative prediction always require implicit differentiation? In *International Conference on Learning Representations*, 2021.
- [5] M. Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, 1965.
- [6] M. Milinski, R. D. Sommerfeld, H. J. Krambeck, F. A. Reed, and J. Marotzke. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences*, 105(7):2291–2294, 2008.
- [7] E. Fehr and S. Gächter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994, 2000.
- [8] E. Ostrom. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, 1990.
- [9] M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In *Proceedings of the 7th Innovations in Theoretical Computer Science Conference*, pages 111–122, 2016.
- [10] L. Liu, A. Miller, and M. Hardt. Strategic classification with multiple classifiers. In *International Conference on Machine Learning*, 2019.
- [11] S. Narang, T. Zrnic, and M. Hardt. Performative prediction with sequential decisions. In *International Conference on Machine Learning*, 2022.

- [12] A. E. Roth. The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, 70(4):1341–1378, 2002.
- [13] S. Shivankar. Robotic strategy adaptation via monte carlo regret minimization in uncertain multi-agent environments. *Authorea*, March 2026. doi: 10.22541/au.177385893.32317660/v1.
- [14] S. K. Bommakanti. Cloud system observability: Automating infrastructure insight for service level adherence. *AI and ML*. doi: 10.1109/ACOIT66109.2025.11436790.
- [15] S. Shaik. Leveraging blockchain technology for a decentralized public key infrastructure: A blueprint for secure digital identity management. *IEEE Xplore*, 2025.
- [16] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, M. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- [17] J. Lorraine, P. Vicol, and D. Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020.
- [18] P. Baral. Adaptive polynomial frameworks for enhanced graph signal learning in data science. *TechRxiv*, December 2025.
- [19] H. Kargeti. Automating enterprise vulnerability exposure: Scap-based asset–cve linkage with social signal triage for cyber defense. doi: 10.1109/11467997.
- [20] B. Ugandhar. Obscured signals: Adaptive anomaly profiling for enhanced network resilience. doi: 10.1109/CCUBE66458.2025.11340076.
- [21] J. Domke. Implicit differentiation for hyperparameter optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- [23] D. C. Parkes and M. P. Wellman. Economic reasoning and artificial intelligence. *Science*, 349(6245):267–272, 2015.
- [24] K. Thaker. Detecting fake companies on linkedin: A machine learning approach. doi: 10.31224/5113.
- [25] T. Dash. Competing on prompts: Generative ai’s market and policy faultlines. *arXiv preprint arXiv:2512.04408*, 2025.
- [26] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.
- [27] U. V. Noubissie. Leveraging mobile phones for enhanced wireless health monitoring: An architectural approach. *Patient Care*.
- [28] V. Kuruppathukattil. Navigating security threats in cloud-based systems. *Computer Networking*.

- [29] A. Khan. Ai-powered text screening: A generative approach to identifying sensitive information. *Generative AI*. doi: 10.1109/ICCC68654.2025.11437858.
- [30] J. Singh. Transactional memory without hardware support: A comprehensive software-driven approach. *Parallel Computing*.
- [31] G. R. Vemula. A hybrid approach to road classification using lidar data integrating aerial imagery and known road centrelines for enhanced feature extraction. In *Proc. PAIDeF AI4D Conference (SuperAI 2025)*, pages 18–36, Abuja, Nigeria, 2025. doi: 10.5281/zenodo.17951785.
- [32] S. Parekh. Responsive interfaces for end-user service orchestration. doi: 10.1109/11448763.
- [33] N. Tadamalla. Decoding the economic forces of australian vineyards: Insights from machine learning on cost drivers and revenue patterns. doi: 10.1109/11210700.