PERFORMATIVE PREDICTION ON GAMES AND MECHANISM DESIGN

António Góis[1] **Mehrnaz Mofakhami**[1] **Fernando P. Santos**[2]
Gauthier Gidel* [1,3] **Simon Lacoste-Julien** [*1,3]
[1]Mila & Université de Montréal [2]University of Amsterdam [3]Canada CIFAR AI Chair

ABSTRACT

Agents often have individual goals which depend on a group's actions. If agents trust a forecast of collective action and adapt strategically, such prediction can influence outcomes non-trivially, resulting in a form of performative prediction. This effect is ubiquitous in scenarios ranging from pandemic predictions to election polls, but existing work has ignored interdependencies among predicted agents. As a first step in this direction, we study a collective risk dilemma where agents dynamically decide whether to trust predictions based on past accuracy. As predictions shape collective outcomes, social welfare arises naturally as a metric of concern. We explore the resulting interplay between accuracy and welfare, and demonstrate that searching for stable accurate predictions can minimize social welfare with high probability in our setting. By assuming knowledge of a Bayesian agent behavior model, we then show how to achieve better trade-offs and use them for mechanism design.

1 INTRODUCTION

Recent frameworks such as performative prediction study how predictions influence the distribution they aim to predict (Hardt & Mendler-Dünner, 2023). These have focused on accuracy for one predictor and independent predicted agents: a spam producer changes its content to fool a spam classifier (Dalvi et al., 2004; Hardt et al., 2016), or one loan applicant adapts to improve its credit score ignoring adaptation by others (Perdomo et al., 2020).

Performative prediction typically considers a larger set of independent data points, but interdependencies among predicted agents have been abstracted by the literature, and are not explicitly modeled. Existing extensions to multi-agent performative prediction focus on multiple *predictors* (Narang et al., 2022), but not on multiple interdependent *predicted*. However a plethora of examples exists requiring a collective scale among predicted. In election polls, the aggregate prediction of voters' behaviour influences individuals' actions (Simon, 1954; Blais et al., 2006). Other times prediction is not directly about aggregate behaviour, but about its consequence. In pandemic modelling, forecasts of disease spread can change real spread, as people infer the consequences of collective action on the predicted outcome. Additionally, potentially accurate pandemic forecasts which were not observed due to performative effects can erode public trust in future predictions (Van Basshuysen et al., 2021). Collective action can also be shaped by predictions of a climate disaster, and road traffic by estimated travel time. In financial markets, price is an aggregate consequence of actions steered by predictions (Soros, 1987). Such self-fulfilling prophecies may have actually deeply harmed society in cases such as the British pound collapse in 1992 (Naef, 2022), illustrating how welfare can be affected in multi-agent settings. Overall, these examples motivate a framework where the predicted population is interdependent, has a utility that only depends indirectly on predictions, trust varies with past accuracy, and welfare becomes an additional metric of concern.

To model a social dilemma such as pandemic containment or cooperation for climate change, here we propose the first setting with inherent interdependence among predicted agents. Agents play a cooperation game whose outcome depends locally on others' actions, and decisions are influenced by predictions of individuals' actions. Moreover, we allow for spatial structure among agents

*Equal supervision.

through a graph, representing social connections or geographic locations of common goods. Predictions are provided about individuals' actions, from which each one infers its group's expected behaviour. Each agent updates a Bayesian trust variable based on past accuracy, which determines to what extent its actions are influenced by predictions.

Social welfare naturally emerges as a metric of interest in addition to accuracy. Welfare depends on the population's current actions, which are a consequence of predictions. Agents' trust links accuracy and welfare, by limiting the influence of inaccurate predictions in the long-term. As in most game theoretical settings, there is a tension between individual and collective interests, rendering high-welfare states possibly unstable, while accuracy maximization remains oblivious to this issue. Despite this, the broad goal of existing frameworks has been to minimize a single risk under performativity, which typically represents accuracy (Perdomo et al., 2020). In our examples, welfare is only influenced indirectly by the predictor, making it unsuitable for standard performative prediction or supervised learning. Optimizing this kind of risk raises additional challenges (Miller et al., 2021), rendering the problem hard to solve in general, without tailoring solutions to problem instances with specific structures.

Despite a gap in the literature, having a secondary risk that does not depend directly on predictions is common when predicted agents are not independent. Predictions of pandemic growth or climate change can inform public policy, and become performative if risk is successfully reduced. In financial markets, predictions may aim at maximizing profit instead of accuracy. In elections, each candidate wishes to push the forecast whose collective reaction will benefit them the most. Even if a neutral entity wishes to deploy an accurate election poll, its performative effect may have strong unintended consequences (Westwood et al., 2020).

This highlights an aspect of mechanism design, which exists even when ignored. While deliberately deploying a wrong prediction is not an ethical option, there may be multiple possible realities that can be induced (Hardt et al., 2022) — therefore different predictions may be equally correct. Providing a snapshot of pre-prediction reality may be a way out of this dilemma, but can be wrongly interpreted as a prediction of post-prediction reality. The choice of how many snapshots to provide before action will also influence arbitrarily the outcome. Providing counterfactual scenarios for different population responses can improve transparency, but may risk an overly complex message being ignored by the predicted (Van Basshuysen et al., 2021). Our work illustrates this problem and difficult choices that arise from it, through the following contributions:

- We propose a novel performative setting where the predicted population is inherently interdependent. For this, we adapt the Collective Risk Dilemma (CRD; Milinski et al., 2008) to include a centralized predictor, together with a trust variable that reduces performative effects over the predicted, as accuracy decreases.
- We show how a second risk, welfare, appears when agents have goals that do not depend directly on the prediction. We also propose a *trust* model that ties welfare to accuracy, as it hinders reaching states that require inaccurate predictions.
- We show in our setting how repeated risk minimization (RRM), a realistic algorithm to maximize accuracy under performativity, can accidentally minimize welfare with high probability over the initial predictions.
- We propose a method to learn how to use predictions as a mechanism to increase cooperation and reach higher welfare states. We show experimentally that this mechanism can achieve higher welfare in exchange for a drop in accuracy.

2 A Model for Predicting Collective Action

We are interested in game-theoretic scenarios where a population is interdependent and possibly influenced by predictions of collective behaviour. To that end, we propose a model where subgroups from a larger population interact simultaneously in a given round, drawing inspiration from evolutionary game theory on networks (Smith, 1982; Ohtsuki et al., 2006). Given a graph $\mathcal{G} = (V, E)$, for any agent i, its group is composed of i (itself) and its neighbors in the graph $\mathcal{N}(i)$. For one round of the game, agents simultaneously select an action, and each agent i receives a payoff $\pi_i(y_i, y_{\mathcal{N}(i)})$ depending on its own action y_i and on neighbours' $y_{\mathcal{N}(i)}$. The game repeats for R rounds.

To choose π_i, we focus on CRD, suitable to study mechanism design (Góis et al., 2019). Each round requires a critical mass of cooperators to achieve success and prevent collective losses. This may represent the protection of common natural resources, the immunity of a partially vaccinated group, and the collective development of tools like Wikipedia or Linux, among many others. If the fraction of cooperators remains below a threshold T, collective success is not achieved and everyone risks losing their endowment with probability r. Each agent chooses whether to defect ($y_i = 0$) or cooperate at a cost ($y_i = 1$), with payoffs below:

Definition 2.1. (Defector's payoff) Let $\mathbb{1}[\cdot]$ be the indicator function. $k_i = \sum_{j \in \mathcal{N}(i) \cup \{i\}} y_j$ is the number of cooperators in agent i's group. Given an initial endowment B, k_i cooperators in a group of size M_i, threshold T where $0 \leq T \leq 1$, and risk r, where $0 \leq r \leq 1$, the expected payoff of defector i is

$$\pi_{D_i}(k_i) = B \cdot (\mathbb{1}[k_i \geq \lceil TM_i \rceil] + (1-r)\mathbb{1}[k_i < \lceil TM_i \rceil]) \tag{1}$$

In words, below threshold there is a disaster with probability r, while over the threshold each agent gets B.

Definition 2.2. (Cooperator's payoff) Given a cost cB of cooperating, where $0 \leq c \leq 1$, the payoff of cooperator i is

$$\pi_{C_i}(k_i) = \pi_{D_i}(k_i) - cB \tag{2}$$

A CRD is used as payoff function π for all agents, using the same threshold value T and unique M_i's given by $\mathcal{G}$. This leads to partially aligned incentives — each agent i gains from preventing a disaster where $\frac{k_i}{M_i} < T$, but would rather avoid incurring cost cB of cooperating to increase k_i.

For one round of CRD with $c < r$ and one single group (where $\mathcal{G}$ is a fully-connected graph $\mathcal{G}_f$), the Nash equilibria [1] are for everyone to defect (sub-optimal) or to have exactly $\lceil TM_i \rceil$ cooperators (Pareto optimal [2]). The challenge is in coordinating a group towards the Pareto optimal Nash, which doesn't happen spontaneously in the real world (Milinski et al., 2008).

2.1 Agent Model

We model agents as computing a best-response, given expectations of other individuals' actions. To nudge behaviour, a predictor provides predictions of the population actions. Alternatively to correlated equilibria (Aumann, 1974) we provide a public signal, which agents can choose to trust or not. Since this signal is learned from global observations of the whole population (and not just locally) it has the potential to bring additional information to agents. We assume agents observe a public prediction of others' actions, but stop trusting it if it is inaccurate. More specifically, they follow a Bayesian update to compute the probability of trusting the prediction. Agent i has two competing explanations for each neighbour j's behaviour — the external prediction $\hat{y}_j$ and an internal expectation $\alpha_{i,j}$. Both $\hat{y}_j$ and $\alpha_{i,j}$ are Bernoulli parameters that estimate a hypothetical true $\mathbb{P}(y_j = 1)$. $\mathcal{L}_i(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)$ is the likelihood of observing $\boldsymbol{y}_t$ given parameter $\hat{\boldsymbol{y}}_t$, and $\tau_{t-1,i}$ acts as the prior for i. The posterior probability $\tau_{t,i}$ of i trusting the external predictor in timestep t is:

$$\tau_{t,i} = \mathbb{P}(\text{trust}|\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{\tau_{t-1,i}\mathcal{L}_i(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)}{\tau_{t-1,i}\mathcal{L}_i(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t) + (1-\tau_{t-1,i})\mathcal{L}_i(\boldsymbol{\alpha}_i, \boldsymbol{y}_t)} \tag{3}$$

with $\mathcal{L}_i(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t) := \prod_{j \in \mathcal{N}(i)} \hat{y}_{j,t}^{y_{j,t}} (1 - \hat{y}_{j,t})^{1-y_{j,t}}$.

Given the expectation of others' actions, i can compute a rational utility-maximizing action. As long as $c < r$, it is rational for i to cooperate if and only if $\sum_{j \in \mathcal{N}(i)} y_j = \lceil TM_i \rceil - 1$. In words, i cooperates when it is the only missing cooperator required to overcome the threshold in its group. Given probability $\boldsymbol{\theta}_{\mathcal{N}(i)} = \theta_{1 \ldots M_i - 1}$ of each neighbour of i to cooperate, a Poisson binomial distribution $g(T_i|\boldsymbol{\theta}_{\mathcal{N}(i)})$ gives us the aggregate probability of having $T_i = \lceil TM_i \rceil - 1$ cooperators in the group. `BestResponse` is then $\arg\max_{y_i} \mathbb{E}_{y_{\mathcal{N}(i)} \sim g(\boldsymbol{\theta}_{\mathcal{N}(i)})}[\pi(y_i, y_{\mathcal{N}(i)})]$.

[1] A state is a Nash equilibrium if no agent can unilaterally change her action to improve her payoff π.

[2] A state is Pareto optimal if there is no other state that improves one's π without lowering another's π.

Proposition 2.3. *(Best-response under competing models) Given two competing models ($\hat{\boldsymbol{y}}$ and α) that explain the population's behaviour, then* $\texttt{BestResponse}_i(\hat{\boldsymbol{y}}_{t+1,\mathcal{N}(i),\theta}; \tau_{t,i})$ *is to cooperate if* $\tau_i g(T_i|\hat{y}_{j\in\mathcal{N}(i)}) + (1-\tau_i) g(T_i|\alpha_{i,j\in\mathcal{N}(i)}) > \frac{c}{r}$*, and defect otherwise (Appendix A)*

2.2 Performative Prediction in Collective Action

Perdomo et al. (2020) propose the framework of performative prediction, and analyze RRM, an algorithm which retrains a model after each distribution shift. More formally, they assume a mapping $\mathcal{D}(\theta)$ from a predictor parameter θ into a distribution $\mathcal{D}$. Considering this effect of θ, performative risk becomes:

$$\text{PR}(\theta) = \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta)} \ell(Z;\theta)$$

As a baseline to optimize PF(θ) without knowledge of $\mathcal{D}(\theta)$, they suggest RRM as a natural heuristic — to repeatedly minimize risk using the distribution obtained from the previous model deployment:

$$\theta_{t+1} = \arg\min_{\theta} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}(\theta_t)} \ell(Z;\theta)$$

We now map their setting to our framework. We have a prediction $\hat{\boldsymbol{y}}_t \in [0,1]^{|V|}$ of each agent's probability of cooperating in time-step t, starting at $t=1$. A trust variable $\tau_{t,i} \in [0,1]$, starting at $t=0$, determines how much agent i trusts predictions, and therefore by what extend its next action $y_{t+1,i}$ is influenced by $\hat{\boldsymbol{y}}_{t+1}$. Finally, we have:

$$\mathcal{D}(\theta;\tau) = \{\mathbb{1}[y_{t+1,i} = \texttt{BestResponse}_i(\hat{\boldsymbol{y}}_{t+1,\mathcal{N}(i),\theta}; \tau_{t,i})] : i = 1, ..., |V|\}$$

yielding a deterministic distribution.

The missing ingredient to apply RRM is which risk to use. We first consider accuracy maximization, in the simplified case of a static $\tau = 1$, a single-round game, and a new deployment of $\hat{\boldsymbol{y}}$ after each round. The learner initializes $\hat{y}_1 \in [0,1]^{|V|}$, after which each agent i computes deterministically $y_{1,i} = \texttt{BestResponse}_i(\hat{\boldsymbol{y}}_{1,\mathcal{N}(i),\theta}; \tau_{0,i})$, corresponding to $y_{1,i} \sim \mathcal{D}(\hat{y}_{1,\mathcal{N}(i)}; \tau_{0,i})$. Unaware of its own influence on $\boldsymbol{y}$, an accuracy-maximizing learner minimizes risk by deploying $\hat{y}_{2,i} = y_{1,i}$, staying in the discrete domain $\hat{y}_t \in \{0,1\}^{|V|}$ from time-step $t=2$ onwards. Note we will not restrict ourselves to RRM or to $\tau = 1$, but it will be the starting point of our analysis.

Accuracy fits naturally the definition of performative risk: $\text{PR}(\theta) := \mathbb{E}_{Z\sim\mathcal{D}(\theta)} \ell(Z;\theta)$. On the other hand, welfare is a metric that only depends on the actions taken by the population: $\mathbb{E}_{Z\sim\mathcal{D}(\theta)} \ell'(Z)$. Minimizing such a loss raises optimization challenges, since θ only influences risk through $\mathcal{D}(\theta)$.

3 Model Dynamics

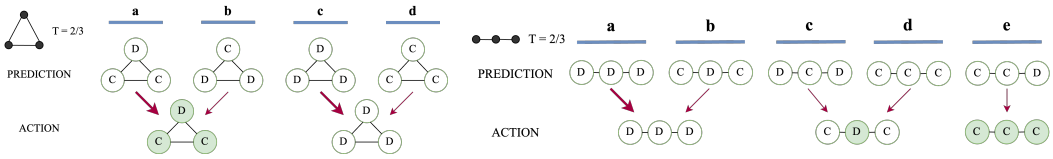
To gain understanding of our model's behaviour, here we study the impact of different components in isolation. We explore two opposite scenarios: blindly trusting a predictor ($\tau = 1$), in particular the impact of $T, \mathcal{G}$ and RRM, and behaviour in the absence of a predictor ($\tau = 0$). We then show oscillatory behaviour for a particular $\mathcal{G}, T$, when τ is dynamic.

3.1 Threshold and Topological Constraints

We adopt the following assumption:

Assumption 3.1. (Simple controllable setting) a) agents have a fixed trust $\tau = 1$ ignoring their internal beliefs α, and b) predictions are binary: $\hat{\boldsymbol{y}}_{\boldsymbol{t}} \in \{0,1\}^{|V|}$.

Let a *self-fulfilling prophecy* be when $\forall i, y_i = \hat{y}_i$. Assuming binary predictions is useful in this definition, since y's need to match $\hat{y}$'s. Removing the interference of internal expectations α by having $\tau_0 = 1$, predictions become static: $\hat{\boldsymbol{y}}_{\boldsymbol{t}} = \hat{\boldsymbol{y}}$. With full trust guaranteed, there is no need



(a) 3-node clique. Both a) and c) are self-fulfilling prophecies with 100% accuracy (an accuracy maximizer is indifferent between them). However, in a) full success is achieved, but in c) all fail. b) also maximizes group success but at the expense of 0% accuracy.

(b) 3-node chain. Here there is no self-fulfilling prophecy which maximizes group success, forcing a trade-off between accuracy and group success. Only e) maximizes group success, but the center node regrets having cooperated. Note that, with $T = \frac{2}{3}$, groups of size $M_i = 2$ require both agents to cooperate.

Figure 1: Predictions and resulting actions in a 3-player game, under different topologies. Dark nodes have achieved success, and thick arrows are self-fulfilling prophecies.

to balance between trust and other goals through time. Under Assumption 3.1, if agents are never indifferent between actions, predicting a strict Nash equilibrium is sufficient and necessary to have a self-fulfilling prophecy (i.e. $\forall i, \texttt{BestResponse}(\hat{y}_{\mathcal{N}(i)}) = \hat{y}_i$).

Whether there is a Nash equilibrium that maximizes welfare determines whether the predictor must compromise accuracy to maximize it. Note that, as long as $\forall i, \lceil TM_i \rceil > 1$, *all-defecting* is always a self-fulfilling prophecy. Using Assumption 3.1, the topology of $\mathcal{G}$ and threshold T become the only constraints determining whether a given system state is attainable.

Proposition 3.2. *(Sufficient conditions for self-fulfilling success) Let "full success" be the setting where $\forall i, \frac{k_i}{M_i} \geq T$. Under Assumption 3.1 and $c < r$, each of the following is a sufficient condition so that $\exists\, \hat{\boldsymbol{y}} \implies$ full success, where $\hat{\boldsymbol{y}}$ is a self-fulfilling prophecy:*

1. *$\mathcal{G} = \mathcal{G}_f$, where $\mathcal{G}_f$ is a clique or a fully-connected graph: Assume $\hat{\boldsymbol{y}}$ predicts a configuration with $k_i = \lceil TM_i \rceil$. Since all agents share the same group, it is not possible for one agent to deviate from $\hat{\boldsymbol{y}}$ without lowering its π;*
2. *T=1: no agent can free-ride, since all are required to cooperate;*
3. *T=0: full success is guaranteed by default.*

Figure 1a illustrates the previous remarks, over a 3-node clique. a) and c) are Nash equilibria and self-fulfilling prophecies, while b) and d) are not. An accuracy maximizer would choose a) or c), while a welfare maximizer would choose a) or b). Here it is possible to maximize both quantities through a).

However, both goals may be at odds in other settings. In Figure 1b there is no prediction that satisfies simultaneously an accuracy maximizer and a welfare maximizer. The only self-fulfilling prophecy is a), reaching full-defection. Only e) reaches full success, but since it is not a Nash it is not self-fulfilling. This is because the center node could have achieved success while defecting, but would have prevented success in groups of size 2. In general, full success is not always achievable, even if we do not require a self-fulfilling prophecy:

Proposition 3.3. *(Necessary condition for success) Under Assumption 3.1, a game must not obey simultaneously the three conditions below, otherwise full success is not attainable, even with a prediction that is not self-fulfilling (proof in Appendix B):*

1. *graph $\mathcal{G}$ has a "hub" node H whose degree $M_H - 1$ is higher than any of its neighbours: $\forall i \in \mathcal{N}(H) : M_i < M_H$.*
2. *Threshold $T \in [0, 1]$ is set to $\frac{M_H - 1}{M_H}$.*
3. *$\forall i \in \mathcal{N}(H), \exists j \in \mathcal{N}(i) : M_j < M_H$.*

This shows how $T, \mathcal{G}$ can condition which states are reachable. The examples also illustrate how different predictions induce different realities in this model. As a consequence, seeking only high-accuracy predictions may inadvertently induce low-cooperation states, as the next section will show.

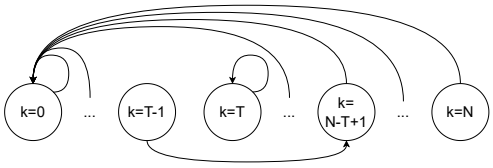


Figure 2: Markov chain describing the evolution of a population with k out of N cooperators, $\tau = 1$ and $\mathcal{G}_f$ following RRM, for $N > 2T - 1, T \neq 1$. (details in Appendix C)

3.2 Convergence to Low Welfare

We now show theoretically that repeatedly maximizing accuracy can minimize welfare with high probability, even when some regions of the optimization landscape allow to optimize for both metrics. As the previous section illustrates, the alignment between welfare and accuracy depends non-trivially on game parameters T and $\mathcal{G}$, even in the simplified case of $\tau = 1$. The choice of optimization algorithm introduces further complexity. Depending on it, even when it is possible to jointly optimize accuracy and welfare (e.g. Figure 1a), an accuracy maximizer is not guaranteed to pick the best solution for both metrics.

Assume a predictor is trained to maximize accuracy through RRM, following § 2.2 with $\mathcal{G} = \mathcal{G}_f$. RRM becomes a Markov chain, where each state is defined only by the number of cooperators in the population, $k_t = \sum_i y_{t,i}$ in the fully-connected case. The initialization of $\hat{\boldsymbol{y}}_1$ will determine which k_1 is induced. From there on RRM will imitate each player's previous best-response, to which players will best-respond until convergence or reaching a cycle. This is equivalent to synchronous best-response dynamics (Chellig et al., 2022). Interestingly, this prevents agents from overcoming the threshold with high probability.

Theorem 3.4. *(Convergence to low welfare) Assume $\forall i, \lceil TM_i \rceil > 1$, and $\mathcal{G} = \mathcal{G}_f$. Predictions $\hat{\boldsymbol{y}}_1 \in \{0,1\}^{|V|}$ can be described by $\hat{k}_1 = \sum_i \hat{y}_{1,i}$ since $\mathcal{G} = \mathcal{G}_f$. Let $\hat{k}_1 \sim$ Bin(θ, N) and $\theta \sim$ Beta(α, β), where $N = |V|$. RRM always converges to a stable point in terms of k_t in at most 2 steps, and $\mathbb{P}(k_t = 0) \geq 1 - \mathcal{O}(1/N)$, $\forall t \geq 2$. (proof in Appendix C)*

Empirically, convergence to low welfare holds for other kinds of $\mathcal{G}$'s. In particular, we study the configuration after convergence of scale-free networks, showing that most time is spent on low cooperation states, as size increases. Cycles become more frequent than a single steady state, in larger scale-free graphs (Fig 3).

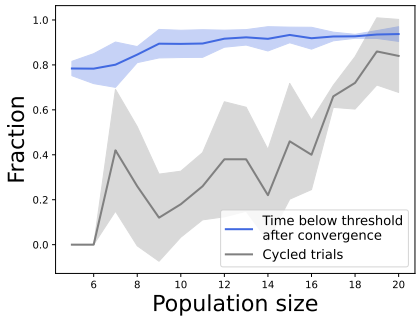


Figure 3: Average proportion of time spent below threshold, after convergence, and proportion of cycles. Each point is the average of 5 scale-free $\mathcal{G}$'s, with 10 random $\hat{\boldsymbol{y}}_1$ per $\mathcal{G}$, and $T = 0.5$. Shaded areas are standard deviation among $\mathcal{G}$'s. Find histograms for time in each state in Appendix C.

3.3 Self-Organization without Predictor

So far we have explored scenarios with full trust ($\tau = 1$) where an external predictor dictates population actions, studying limitations on how much welfare can be obtained and behaviour under RRM. Now we focus on the opposite scenario, in the absence of a predictor.

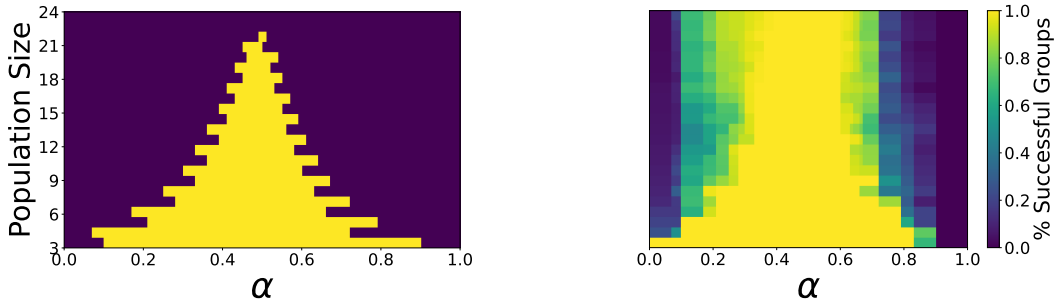


Figure 4: Proportion of successful groups with $\tau = 0$, for a fully-connected graph (left) and scale-free networks (right) with varying α and population size N. $\frac{c}{r} = \frac{1}{6}, T = 0.5$ and scale-free's average degree $m = 2$.

For $\mathcal{G} = \mathcal{G}_f$ it is possible to compute the price of anarchy (Koutsoupias & Papadimitriou, 2009), which is the ratio between the welfare of the best state and the welfare of the worst equilibrium. Defining welfare as $\text{Welf}(\boldsymbol{y}) = \sum_i \pi_{y_i}(k_i)$, we obtain the price of anarchy $\text{PoA} = \frac{1-c\frac{\lceil NT \rceil}{N}}{1-r} \approx \frac{1-cT}{1-r}$. We are interested in regimes where $r > c$, and we note that PoA increases with higher r and lower c, worsening the gap between best state and worst equilibrium.

PoA however does not consider to which equilibrium we arrive at without an external entity, focusing on the worst case. Figure 4 shows results without trust ($\tau = 0$), where agents behave according to their internal expectation α of others. For a fully-connected population of size N, each agent's best-response is to cooperate if $\text{Bin}(\lceil TN \rceil - 1; \alpha, N) > \frac{c}{r}$. From this equation we see that as $\frac{c}{r}$ drops, PoA increases but agents choose to cooperate for a wider range of T, N, α. For the remaining of the paper, we focus on regimes where high $\text{Welf}(\boldsymbol{y})$ does not happen spontaneously.

3.4 Trust Dynamics

When trust varies with accuracy, our system becomes stateful. The same prediction may induce different responses depending on how much each agent currently trusts predictions. Here we present a simplified analysis for the $\mathcal{G}, T$ described in Figure 1b, assuming an infinite horizon discounted reward criterion.

An optimal accuracy-maximizer will trivially predict *all-defect*, inducing a stable low-welfare state with perfect accuracy, and trust approaching 100%. However an optimal welfare-maximizer must handle the tension between accuracy and welfare, resulting from the trust variable. Unless there is a discount factor γ that is so high that the predictor indefinitely increases trust for a higher future reward, this will lead to oscillatory behaviour in trust. While a prediction of *all-cooperate* matches the best-response of both left and right nodes, it does not for the center node. To induce *all-cooperate*, the prediction must distribute error in a way that minimally harms accuracy, while maintaining the welfare-maximizing outcome. We derive and prove in Appendix D the optimal welfare-maximizing policy, for different values of γ. This results in gradually lower confidence for the center node until it is necessary to boost it. For certain γ's, the optimal boost is to induce the most surprising outcome for a given α, causing an abrupt trust increase (Figure 5).

4 Learned Predictor and Simulations

As the population size $|V|$ grows, analysis becomes more complex. We resort to simulations and learned predictors to study larger systems.

We choose to represent the predictor through a neural network, which receives as input the population's actions in the previous time-step: $\hat{\boldsymbol{y}}_{\boldsymbol{t}} = f_\phi(a_{t-1}) : \{0,1,\varnothing\}^{|V|} \to [0,1]^{|V|}$. The loss is either cross-entropy for accuracy maximization, a differentiable proxy of number of successful groups for welfare maximization, or a combination of both. To combine both in a multi-task objective, we follow the approach of Sener & Koltun (2018) (see Appendix F). Here we consider a finite horizon of

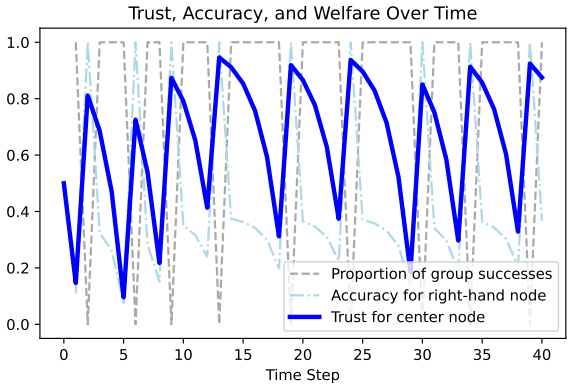


Figure 5: Trust oscillation induced by the optimal welfare maximizer. $T = \frac{2}{3}$ and $\mathcal{G}$ is a 3-node chain (as seen in Figure 1b). Parameters are $c = 0.3, r = 0.5, B = 1, \alpha = 0.8, \tau_0 = 0.5$. The above strategy is optimal for at least $\gamma \in [0.4, 0.5]$ (see Appendix D for details and other γ).

20 rounds per game with no discounting. A predictor with access to multiple games performs gradient descent after each, assuming access to the inner behaviour of agents. To maximize the number of successful groups, it backpropagates through a differentiable version of their decision rule and payoff (Appendix E).

When optimizing for social welfare, the predictor still needs to consider prediction accuracy in order to maintain agents' trust. Let $\tilde{y}_{i,t} = \sigma(\pi_{C_i,t} - \pi_{D_i,t})$ be a differentiable proxy of agents' true decision rule $y_{i,t} = \mathbb{1}[\pi_{C_i,t} - \pi_{D_i,t} > 0]$. We analyze here a proxy goal $\hat{U}_C = \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{y}_{i,t}$ whose gradient can be decomposed in two components:

$$\nabla_\phi \hat{U}_\mathrm{C} = \sum_{t=1}^{T} \sum_{i=1}^{N} \psi_{t,i}(\phi)[\\ (g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi)) - g(T_i|\alpha_{i,j\in\mathcal{N}(i)})) \underbrace{\nabla_\phi \tau_{t,i}(\phi)}_{\text{accuracy}} + \tau_{t,i}(\phi) \underbrace{\nabla_\phi g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi))}_{\text{steering}}] \quad (4)$$

$\psi_{t,i}(\phi) = \tilde{y}_{t,i}(1 - \tilde{y}_{t,i})rB$ is a scalar which is higher when agents are closer to flipping their choice of action between cooperation and defection. An optimizer using this goal needs to control accuracy to keep trust high, and steer towards higher cooperation when trust is high. In practice we use a slightly more complex goal $\hat{U}_{Pop}$ that is closer to true social welfare, leading to qualitatively similar empirical results and amenable to a similar analysis (Appendix E).

In Figure 6 we observe the result of training either for accuracy or welfare maximization, for different values of threshold (details in Appendix F). Unless the threshold is very low ($T \in \{0.2, 0.3\}$), a predictor maximizing accuracy will induce states of very low cooperation (find a related discussion in Appendix B). A predictor maximizing welfare can prevent this, but at the expense of accuracy. This is in line with § 3.1, where both metrics may be impossible to maximize simultaneously. To overcome this, we follow Sener & Koltun (2018) to jointly optimize for both metrics. For $T = 0.5$, we compute the Pareto front in this way.

Regarding architecture choices, we use a multilayer perceptron (MLP), a graph neural network (GNN), GNN+MLP and GNN+linear (Figure 7). For an MLP, a concatenation of all nodes' actions is provided as input, and their actions for the next step are jointly predicted. Having a GNN followed by an MLP or a linear layer provides a gain over MLP alone, by adding information about $\mathcal{G}$. Interestingly, GNNs alone, being the only model unable to do centralized coordination, are not able to promote cooperation. For two equal nodes, some settings may require one to cooperate and the other to defect. A GNN however is unable to provide different outputs to equal nodes. When optimizing for accuracy, this limitation of GNNs goes by unnoticed (Appendix F). We also provide in Appendix F an ablation study of the performative gradient used to maximize cooperation. We

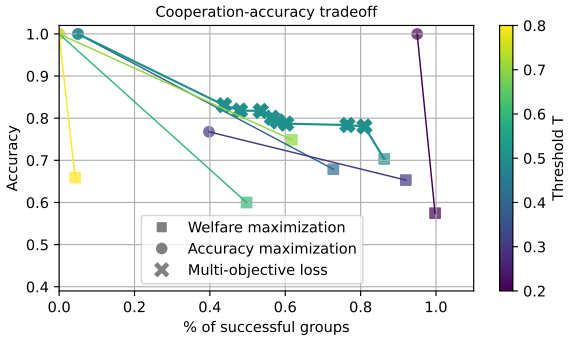


Figure 6: Accuracy vs. social welfare trade-off for different threshold values. Pareto front computed through multi-objective optimization for $T = 0.5$. All experiments were conducted using a scale-free $\mathcal{G}$ with 20 nodes and mean degree of 2 (Barabási & Albert, 1999), $c = 0.2, B = 1, r = 0.4$, $\alpha_{i,j} = 0.8$ and $\tau_0 = 0.5$.

show the importance of different components, to guide future research in estimating performative gradients dependent on trust.

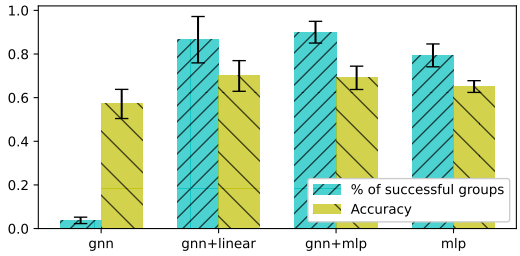


Figure 7: Performance of different architectures, when optimized to maximize social welfare. $T = 0.5$ and other parameters follow Figure 6.

5 Related Work

Machine learning began considering performative effects by assuming explicit models of adaptation (Dalvi et al., 2004). This line of work became known as strategic classification (Hardt et al., 2016), initially considering samples with static outcomes y and adaptive features x. Later works consider indirect changes in y from causal effects of x on y (Miller et al., 2020; Horowitz & Rosenfeld, 2023). In our setting we predict actions y which agents select to maximize their utility. Predictions are based on past observations of actions, modeling $\mathbb{P}(Y)$ for single-round games and $\mathbb{P}(Y_t|Y_{t-1})$ or $\mathbb{P}(Y_t|Y_{1:t-1})$ for multiple rounds.

The framework of performative prediction (Perdomo et al., 2020) provides theory for a class of adaptation behaviours, showing convergence of RRM, an algorithm which retrains a model after each distribution shift. However they require strong smoothness assumptions not verified in our setting, and do not model explicit interdependencies among predicted. We show, in our setting, how RRM can impact welfare, identifying issues ignored when monitoring accuracy alone.

In existing literature, the utility of predicted agent i depends directly and exclusively on predicted label $\hat{y}_i$. In our setting utility depends on others' actions, hence only indirectly on predictions. This is a realistic setting which has not been studied before, raising optimization challenges since RRM is not applicable (Miller et al., 2021; Izzo et al., 2021). Interestingly, Izzo et al. (2021) decompose the performative gradient $\nabla_\theta \mathcal{L}(\theta)$ into an "easy" component $\nabla_1 \mathcal{L}(\theta)$ assuming the distribution is static, and a "hard" component $\nabla_2 \mathcal{L}(\theta) = \nabla_\theta \mathcal{L}(\theta) - \nabla_1 \mathcal{L}(\theta)$ which is the remaining part that requires knowledge of the adaptation. RRM only uses $\nabla_1 \mathcal{L}(\theta)$, but maximizing welfare in our setting requires $\nabla_2 \mathcal{L}(\theta)$, since $\nabla_1 \mathcal{L}(\theta) = 0$.

Multi-agent extensions of performative prediction have focused mostly on multiple predictors (Li et al., 2022; Piliouras & Yu, 2023; Wang et al., 2023). In Eilat et al. (2023) predicted outcomes depend on a graph $\mathcal{G}$ because the classifier assumes it. Mendler-Dünner et al. (2022) mention spill-over effects as a way to give a causal treatment to social influence. Hardt et al. (2023) consider predicted agents that coordinate to influence the training of a classifier. Hardt et al. (2022) propose a lower bound on how much a predictor can influence reality, both for single and multiple predictors. Brown et al. (2022) consider a stateful version of performative prediction where the previous distribution is enough to define a state, unlike in our setting, where $\hat{\boldsymbol{y}}_t$ does not contain all information about τ_t. Surprisingly we have not found any existing work on the role of trust in performativity, highlighting a need for future work in this direction.

Regarding alternative goals to accuracy, recommender systems may wish to preserve content diversity (Eilat & Rosenfeld, 2023), and schools can use predictions in an attempt to improve graduation rates (Perdomo et al., 2023). Vo et al. (2024) consider a trade-off between selecting good candidates and maximizing their improvement, with consequences for agent welfare. Levanon & Rosenfeld (2021) balance accuracy and user utility through regularization. Kim & Perdomo (2023) suggests learning a single predictor for many possibly-competing goals.

Behavioural economics has studied cooperation in non-linear social dilemmas (Milinski et al., 2008), which were modelled using evolutionary game theory (Santos & Pacheco, 2011) and used to study mechanism design (Góis et al., 2019). Another evolutionary model describes adherence to pandemic mitigation measures, showing oscillatory behaviour (Glaubitz & Fu, 2020). When predictions are used for downstream tasks, the cost of obtaining a prediction can also be considered in the final goal (Perdomo, 2024; Shirali et al., 2024).

6 Discussion

The proposed setup points to difficult choices that arise in multi-agent performative scenarios. Optimizing for something other than accuracy raises ethical questions, even when the predictor acts in the predicted's interest. However, this work shows that focusing only accuracy is not neutral and can have negative consequences for the predicted. This is an active topic of research in philosophy (Van Basshuysen et al., 2021). Khosrowi (2023) proposes two opposing views: *mitigation* and *appraisal*. The appraisal view suggests that model evaluation should consider performative effects — e.g. a reduced death toll after a pandemic prediction, and possibly higher accuracy had there been no performativity. However, evaluating models by their consequences undesirably injects moral values into model choice (e.g., prioritizing individual freedom vs. public health). The mitigation view seeks stable, accurate predictions by endogenizing causes of performativity, though this choice may also reflect moral values and has negative impact in our setting. Appraisal corresponds to welfare-maximization and mitigation to accuracy in our model. While this is an open question in philosophy, we emphasize that current machine learning approaches overlook such effects, leading to potential harm — an example would be recommendation systems increasing anxiety to boost engagement. If a predictor does not act in the predicted's interest, it could exploit performative effects for manipulation, which is outside our model's scope. Even then, we argue that making this research open is in the public's interest, to raise awareness and develop defenses. We hope our results inspire future work on better guidelines and algorithms.

7 Conclusion

We have introduced a framework to study performative effects under game-theoretic settings on a network of predicted agents. We show theoretically how social welfare and accuracy can be in conflict, and empirically compute their Pareto front. Although accuracy may seem like a way to avoid manipulating reality, multiple accurate outcomes with different social welfare can be induced when performativity is strong enough. Ignoring side-effects may be more harmful than considering them, making it inevitable to regard performative prediction (partly) as mechanism design in our examples. This brings connections to active research directions in philosophy, with practical impact in existing systems. Finally, our model opens up opportunities for future work on the roles of trust, information design, welfare and graphs in performativity.

REFERENCES

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

André Blais, Elisabeth Gidengil, and Neil Nevitte. Do polls influence the vote? *Capturing campaign effects*, pp. 263–279, 2006.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Jordan Chellig, Calina Durbac, and Nikolaos Fountoulakis. Best response dynamics on random graphs. *Games and Economic Behavior*, 131:141–170, 2022.

Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 99–108, 2004.

Itay Eilat and Nir Rosenfeld. Performative recommendation: diversifying content via strategic incentives. In *International Conference on Machine Learning*, pp. 9082–9103. PMLR, 2023.

Itay Eilat, Ben Finkelshtein, Chaim Baskin, and Nir Rosenfeld. Strategic classification with graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TuHkVOjSAR`.

Alina Glaubitz and Feng Fu. Oscillatory dynamics in the dilemma of social distancing. *Proceedings of the Royal Society A*, 476(2243):20200686, 2020.

António R Góis, Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. Reward and punishment in climate change dilemmas. *Scientific reports*, 9(1):1–9, 2019.

Moritz Hardt and Celestine Mendler-Dünner. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*, 2023.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Moritz Hardt, Meena Jagadeesan, and Celestine Mendler-Dünner. Performative power. In *Advances in Neural Information Processing Systems*, 2022.

Moritz Hardt, Eric Mazumdar, Celestine Mendler-Dünner, and Tijana Zrnic. Algorithmic collective action in machine learning. In *International Conference on Machine Learning*, pp. 12570–12586. PMLR, 2023.

Guy Horowitz and Nir Rosenfeld. Causal strategic classification: A tale of two shifts. In *International Conference on Machine Learning*, pp. 13233–13253. PMLR, 2023.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: performative gradient descent. In *International Conference on Machine Learning*, pp. 4641–4650. PMLR, 2021.

Donal Khosrowi. Managing performative models. *Philosophy of the Social Sciences*, 53(5):371–395, 2023.

Michael P Kim and Juan C Perdomo. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.

Elias Koutsoupias and Christos Papadimitriou. Worst-case equilibria. *Computer science review*, 3 (2):65–69, 2009.

Sagi Levanon and Nir Rosenfeld. Strategic classification made practical. In *International Conference on Machine Learning*, pp. 6243–6253. PMLR, 2021.

Qiang Li, Chung-Yiu Yau, and Hoi To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ufRSbXtgbOo.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating performativity by predicting from predictions. In *Advances in Neural Information Processing Systems*, 2022.

Manfred Milinski, Ralf D. Sommerfeld, Hans-Jürgen Krambeck, Floyd A. Reed, and Jochem 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. doi: 10.1073/pnas.0709546105. URL https://www.pnas.org/doi/abs/10.1073/pnas.0709546105.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020.

John P Miller, Juan C Perdomo, and Tijana Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pp. 7710–7720. PMLR, 2021.

Alain Naef. *An Exchange Rate History of the United Kingdom: 1945–1992*. Cambridge University Press, 2022.

Adhyyan Narang, Evan Faulkner, Dmitriy Drusvyatskiy, Maryam Fazel, and Lillian Ratliff. Learning in stochastic monotone games with decision-dependent data. In *International Conference on Artificial Intelligence and Statistics*, pp. 5891–5912. PMLR, 2022.

Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Juan C Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult lessons on social prediction from wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

Juan Carlos Perdomo. The relative value of prediction in algorithmic decision making. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=oaACFfNbXl.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, pp. 1047–1074, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597759.

Francisco C. Santos and Jorge M. Pacheco. Risk of collective failure provides an escape from the tragedy of the commons. *Proceedings of the National Academy of Sciences*, 108(26):10421–10425, 2011. doi: 10.1073/pnas.1015648108. URL https://www.pnas.org/doi/abs/10.1073/pnas.1015648108.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Ali Shirali, Rediet Abebe, and Moritz Hardt. Allocation requires prediction only if inequality is low. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=WUicA0hOF9.

Herbert A Simon. Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly*, 18(3):245–253, 1954.

John Maynard Smith. *Evolution and the Theory of Games*. Cambridge university press, 1982.

George Soros. *The Alchemy of Finance: Reading the Mind of the Market*. Simon & Schuster, 1987.

Philippe Van Basshuysen, Lucie White, Donal Khosrowi, and Mathias Frisch. Three ways in which pandemic models may perform a pandemic. *Erasmus Journal for Philosophy and Economics*, 14 (1):110–127, 2021.

Kiet QH Vo, Muneeb Aadil, Siu Lun Chau, and Krikamol Muandet. Causal strategic learning with competitive selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15411–15419, 2024.

Xiaolu Wang, Chung-Yiu Yau, and Hoi To Wai. Network effects in performative prediction games. In *International Conference on Machine Learning*, pp. 36514–36540. PMLR, 2023.

Sean Jeremy Westwood, Solomon Messing, and Yphtach Lelkes. Projecting confidence: How the probabilistic horse race confuses and demobilizes the public. *The Journal of Politics*, 82(4): 1530–1544, 2020. URL `https://doi.org/10.1086/708682`.

A BEST RESPONSE

An agent's best response selects the action with highest expected payoff, between cooperation and defection. Let $k'_i = \sum_{j \in \mathcal{N}(i)} y_j$ be the number of cooperators in i's group, excluding i itself. The payoff gain of switching from defection to cooperation is

$$\pi_{C_i}(k'_i + 1) - \pi_{D_i}(k'_i) = \begin{cases} (r-c)B & \text{if } k'_i = \lceil TM_i \rceil - 1 \\ -cB & \text{otherwise} \end{cases}$$

In words, i gains $(r-c)B$ from cooperating when it is the last member required to overcome the threshold in its group. It loses cB for any other group configuration. Its best response is then to cooperate when it is "at the threshold" ($k'_i = \lceil TM_i \rceil - 1$) and defect otherwise, as long as $c < r$.

Its expectation of others' actions depends on two competing explanations $g(T_i|\hat{y}_{j \in \mathcal{N}(i)})$ and $g(T_i|\alpha_{i,j \in \mathcal{N}(i)})$, and the likelihood τ_i of trusting the first option. Each explanation provides the likelihood $\mathbb{P}(k'_i = \lceil TM_i \rceil - 1) = g(T_i|\theta_{j \in \mathcal{N}(i)})$, by using a Poisson binomial distribution to aggregate individual likelihoods θ_j of each neighbour to cooperate. It should then cooperate if

$$\mathbb{E}_{\tau_i}[\mathbb{E}_{g(\hat{y}_{j \in \mathcal{N}(i)})}[\mathbb{E}_{g(\alpha_{i,j \in \mathcal{N}(i)})}[\pi_{C_i}(k'_i + 1) - \pi_{D_i}(k'_i)]]] > 0 \, (=)$$
$$\tau_i g(T_i|\hat{y}_{j \in \mathcal{N}(i)}) + (1 - \tau_i) g(T_i|\alpha_{i,j \in \mathcal{N}(i)}) > \frac{c}{r}$$

B FULL SUCCESS IS NOT ALWAYS ACHIEVABLE

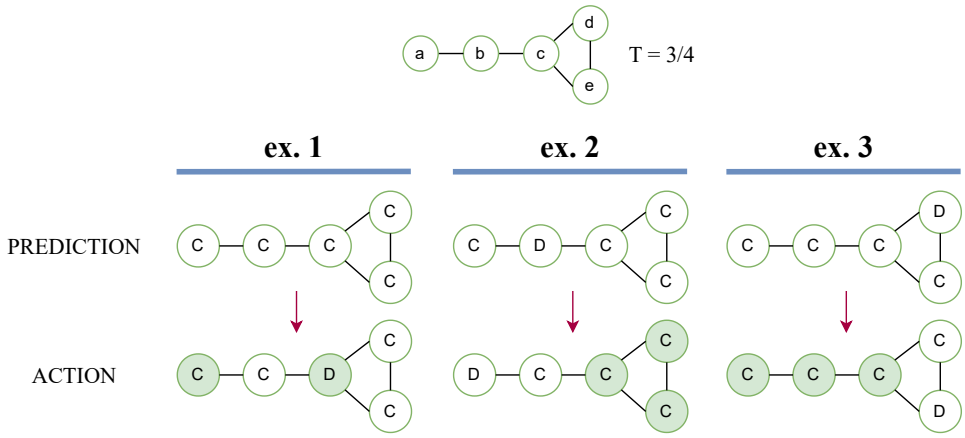


Figure 8: Achieving full success is not possible for all configurations of $\mathcal{G}$ and T. In this counter-example, node c requires $\lceil TM_c \rceil = 3$ cooperators out of $M_c = 4$, meaning one node in c's group can defect without preventing success. As a consequence c will cooperate only if one of b, d, e is predicted to defect. All the other groups require 100% of cooperators since they have $M_i < 4$ and $T = \frac{3}{4}$. If c doesn't cooperate (ex. 1), it'll prevent success for its neighbours. If any of b, d, e is predicted to defect (ex. 2 and 3), it'll also prevent someone's success. These contradicting requirements make it impossible to reach full success for any given prediction $\hat{\boldsymbol{y}}$.

Under Assumption 3.1 there exist combinations of T and $\mathcal{G}$ for which full success is unattainable, even without requiring a self-fulfilling prophecy. This is due to contradicting requirements in neighbour nodes, which cannot be simultaneously satisfied through any prediction $\hat{\boldsymbol{y}}$. One example is Figure 8.

From it, we can formulate sufficient condition for full success to be unattainable, or conversely (through its negation) a necessary condition for full success to be attainable.

1. graph $\mathcal{G}$ has a "hub" node H whose degree $M_H - 1$ is higher than any of its neighbours: $\forall i \in \mathcal{N}(H) : M_i < M_H$.
2. Threshold $T \in [0, 1]$ is set to $\frac{M_H - 1}{M_H}$.
3. $\forall i \in \mathcal{N}(H), \exists j \in \mathcal{N}(i) : M_j < M_H$.

With condition 2, for H to overcome threshold, one out of M_H agents does not need to cooperate (since $M_k \in \mathbb{N}$ and $\lceil TM_H \rceil = \lceil \frac{M_H - 1}{M_H} M_H \rceil = M_H - 1$). However, all neighbours $i \in \mathcal{N}(H)$ require 100% cooperators since they have $M_i < M_H \implies \lceil TM_i \rceil = M_i$. Condition 3 ensures each neighbour of H is connected to another neighbour j with low degree $M_j < M_H$. This combination requires all $i \in \mathcal{N}(H)$ to be predicted to cooperate (i.e. $\forall i \in \mathcal{N}(H), \hat{y}_i = 1$), otherwise their neighbours $j \in \mathcal{N}(i)/\{H\}$ will not cooperate (since they require 100% cooperators). However, $\forall i \in \mathcal{N}(H), \hat{y}_i = 1 \implies y_H = 0$ since H can afford one defector in its group. Since $y_H = 1$ is a requirement for the success of $i \in \mathcal{N}(H)$, but that requires $\exists! i \in \mathcal{N}(H) : \hat{y}_i = 0$, we arrive at contradicting requirements.

This proves Proposition 3.3, since a graph must not obey the condition above, for a prediction to exist which induces full success.

This condition matches empirical observations in Figure 6. Thresholds that are close to but below 100% yield low success, even when maximizing welfare. This indicates that there may be no configuration which allows for high success, for settings $(\mathcal{G}, T)$ with high T.

Other counter-examples may be derived from this sufficient condition, such as those in Figure 9.

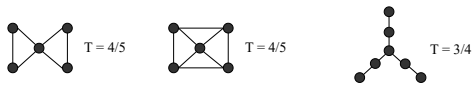


Figure 9: Other counter-examples where full success is not attainable.

This illustrates that there are cases where we cannot achieve cooperation in our model, even at the expense of lower accuracy. In less restrictive topologies discussed in the main paper (e.g. Figure 1b) it is possible to achieve high cooperation, but not maintain it together with accuracy in the long term. In simpler cases (e.g. Figure 1a) it is possible to aim for high cooperation, and then maintain it together with high accuracy.

C Convergence to Low Welfare Under RRM

We begin by analyzing the probability of inducing any initial state, from a random initialization of $\hat{\boldsymbol{y}}_{t=1}$.

Lemma C.1. *(Initial state probability) Assume $\hat{\boldsymbol{y}}_1 \in \{0,1\}^{|V|}$, which can be described by $\hat{k}_1 = \sum_i \hat{y}_{1,i}$ if $\mathcal{G}$ is fully-connected. Let $n = |V|$, $\hat{k}_1 \sim$ Bin(θ, n) and $\theta \sim$ Beta(α, β). The probability of being initialized at a given state k is:*

$$\mathbb{P}(\hat{k}_1 = k) = \frac{n!}{k!(n-k)!} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

Proof. we have that

$$\begin{aligned}
\mathbb{P}(\hat{k}_1 = k) &= \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\, d\theta && (5)\\
&= \frac{n!}{k!(n-k)!} \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}\, d\theta && (6)\\
&= \frac{n!}{k!(n-k)!} \frac{\mathrm{B}(k+\alpha, n-k+\beta)}{\mathrm{B}(\alpha,\beta)} && (7)\\
&= \frac{n!}{k!(n-k)!} \frac{\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(n+\alpha+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (8)
\end{aligned}$$

□

Corollary C.2. *(Probability with increasing population size) under the conditions of C.1*

$$\lim_{n\to 0} \mathbb{P}(\hat{k}_1 = k) = 0$$

Proof. Let $\hat{\alpha} = \alpha - 1$. Using Stirling's formula, we have that $\Gamma(x+\alpha) \sim x^{\alpha}\Gamma(x)$.

$$\begin{aligned}
\mathbb{P}(\hat{k}_1 = k) &= \frac{n!}{k!(n-k)!} \frac{\Gamma(k+1+\hat{\alpha})\Gamma(n-k+\beta)}{\Gamma(n+1+\hat{\alpha}+\beta)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (9)\\
&\underset{\substack{n\to 0\\ k=Tn\\ n-k=(1-T)n}}{\sim} \frac{n!}{k!(n-k)!} \frac{\Gamma(k+1)(k+1)^{\hat{\alpha}}\Gamma(n-k)(n-k)^{\beta}}{\Gamma(n+1)(n+1)^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (10)\\
&= \frac{\Gamma(n-k)}{(n-k)!} \frac{(k+1)^{\hat{\alpha}}(n-k)^{\beta}}{(n+1)^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (11)\\
&= \frac{1}{n-k} \frac{(k+1)^{\hat{\alpha}}(n-k)^{\beta}}{(n+1)^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (12)\\
&= \frac{(k+1)^{\hat{\alpha}}(n-k)^{\beta-1}}{(n+1)^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} && (13)
\end{aligned}$$

$$= \frac{(Tn+1)^{\hat{\alpha}}((1-T)n)^{\beta-1}}{(n+1)^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{14}$$

$$= \frac{1}{n+1} \frac{(T+\frac{1-T}{n+1})^{\hat{\alpha}}((1-T)(1-\frac{1}{n+1}))^{\beta-1}}{T^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{15}$$

$$= \frac{1}{n} \frac{T^{\hat{\alpha}}(1-T)^{\beta-1}}{T^{\hat{\alpha}+\beta}} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} + O(1/n^2)) \tag{16}$$

$$\to 0 \tag{17}$$

□

By Stirling's formula, we have

$$\frac{\Gamma(x+\alpha)}{\Gamma(x)} \sim \frac{(\frac{x+\alpha}{x})^x (x+\alpha)^\alpha}{e^\alpha} = \frac{(1+\frac{\alpha}{x})^x (x+\alpha)^\alpha}{e^\alpha} \tag{18}$$

Now note that $(1+\frac{\alpha}{x})^x = \exp(x\log(1+\frac{\alpha}{x})) \sim e^\alpha$, and $(x+\alpha)^\alpha = x^\alpha(1+\frac{\alpha}{x})^\alpha \sim x^\alpha$ thus,

$$\frac{\Gamma(x+\alpha)}{\Gamma(x)} \sim x^\alpha \tag{19}$$

We now prove that, for most initializations of $\hat{\boldsymbol{y}}$ and values of (N, T), RRM converges to a stable point $\mathcal{D}(\hat{\boldsymbol{y}}_{\text{Defect}})$ where everyone is predicted to defect and indeed defects, in a self-fulfilling fashion. This leaves the population stuck in a severely suboptimal equilibrium.

Assume a predictor iterates through RRM, predicting a population playing in a fully-connected graph $\mathcal{G}_f$, with static $\tau = 1$. The system becomes a Markov decision problem (MDP) with the following states:

Figure 10: RRM for a single-round CRD with static $\tau = 1$ becomes an MDP with N+1 states.

Let $k_i := \sum_{j \in -i} Y_j$, $k := \sum_{j \in [1..N]} Y_j$. In the context of this MDP we adopt a slight abuse of notation by using T to denote $\lceil TM_i \rceil$. Since RRM copies the previous set of actions, agents play by best-responding to the previous round. After initializing at some state k, each agent i cooperates if and only if they observe $k_i = T - 1$. Note that each agent can only observe $k_i = k$ or $k_i = k - 1$, depending on whether i is predicted to defect or cooperate, respectively.

Let us categorize states as:

1. $k = T$
2. $k = T - 1$
3. all others

States 3. move the system into $k = 0$, since no agent observes $k_i = T - 1$.

For $k = T$, all cooperators i observe $k_i = T - 1$, choosing to cooperate next. All defectors j observe $k_j = T$, choosing to defect next. Therefore 1. is a stable point.

For $k = T - 1$, only defectors j observe $k_j = T - 1$. Therefore defectors and cooperators flip, and 2. leads to $k = N - T + 1$. The next step depends on whether $k = N - T + 1$ puts us in 1., 2., or 3.:

1. If $N - T + 1 = T(=)N = 2T - 1$, we arrive in 1. and stabilize at $k = T$.
2. If $N - T + 1 = T - 1(=)N = 2T - 2$, we arrive back in 2. meaning we stabilize at $k = T - 1$. At each step cooperators and defectors flip, but the count remains constant at $k = \frac{1}{2}$.
3. Otherwise we are in 3., meaning we arrive at $k = 0$ in 2 steps.

At $k = 0$:

- If $k = 0$ is 1. or 3., we stabilize at $k = 0$.
- If $k = 0$ is 2. then $T - 1 = 0(=)T = 1$. We cycle between $k = 0$ and $k = N$.

Therefore we have 1 state converging to $k = T$, or 2 in the particular case of $N = 2T - 1$. All other states lead to a stable state of $k = 0$ in at most 2 steps, with two exceptions: a) with $T = 1$, there is a cycle between $k = 0$ and $k = N$, and b) with $N = 2T - 2$ the state $k = T - 1$ is also stable. This proves theorem 3.4, where minimizing welfare refers to stabilizing at any state below the threshold, since we consider the cost cB of unnecessarily cooperating is negligible compared to the cost rB of being below T.

Running simulations with scale-free networks, we observe the same pattern which theory predicts for fully-connected graphs — the proportion of time spend below the threshold approaches 100% as N grows. For runs that cycled instead of converging (which does not happen in a fully connected $\mathcal{G}$), we report the average proportion of time spent by each group in each state, for one cycle (Figure 11).

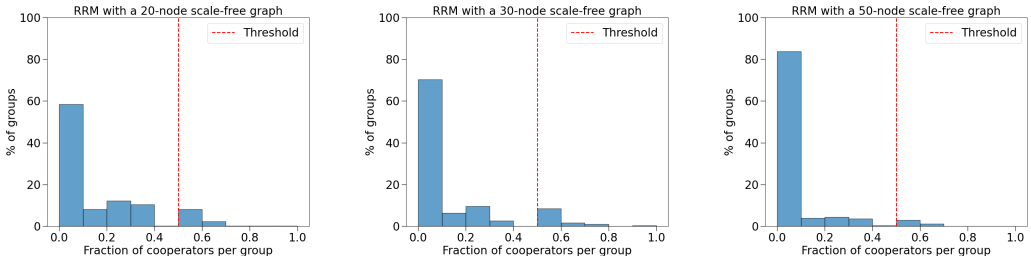


Figure 11: Cooperation per group after applying RRM on a scale-free network with 20 nodes (left), 30 nodes (middle), and 50 nodes (right). Each histogram is the average of 10 different initializations, with a threshold value of 0.5. Each point on the histogram belongs to either a cycle of RRM or a steady state after convergence.

D Conditions for Oscillations in Trust

D.1 Setup

Assume a 3-node chain and threshold $T = \frac{2}{3}$. Let us denote nodes in the chain by 0,1,2 (where 1 is the center node), and actions by $y_{t,i} = 1$ (cooperate) and $y_{t,i} = 0$ (defect). A predictor maximizing the number of successful groups can, at most, obtain the following rewards per time-step:

- Reward $R = 3$, if agents play *1—1—1* (all overcome threshold)
- Reward $R = 2$, if agents play *1—1—0* or *0—1—1* (only nodes 0,1 or 1,2 overcome threshold)
- Reward $R = 1$, if agents play *1—0—1* (only 1 overcomes threshold)
- Reward $R = 0$, if agents play for instance *0—0—0* (no one overcomes threshold)

$R = 3$ is only possible if $y_{t,1} = 1$, which requires $\mathbb{P}(y_{t,0} = 0, y_{t,2} = 1) + \mathbb{P}(y_{t,0} = 1, y_{t,2} = 0) > \frac{c}{r}$. We assume this does not happen spontaneously without the external predictor ($\tau_{t,1} = 0$). More precisely, we discard values of α and $\frac{c}{r}$ where $2\alpha(1-\alpha) > \frac{c}{r}(=)\frac{1-\sqrt{1-2\frac{c}{r}}}{2} < \alpha < \frac{1+\sqrt{1-2\frac{c}{r}}}{2}$. In this regime, we can only have $R = 3$ if $\tau_{t,1}$ is high enough.

Due to symmetry between nodes 0 and 2, obtaining $R = 2$ (not show in Figure 1b) is not possible with static $\tau = 1$, since both observe the same $\hat{y}_{t,1}$, leading to $y_{t,0} = y_{t,2}$. Even with dynamic $\tau_{t,i} < 1$, it is not possible unless $\tau_{0,0} \neq \tau_{0,2}$, which we do not consider here. $R = 1$ (with *1—0—1*) requires high enough α or $\tau_{t,0}, \tau_{t,2}$, and can be used to boost trust for node 1. States with $R = 0$ such as *0—0—0* can be used to boost everyone's trust, if needed.

Note that whether *0—0—0* or *1—0—1* provides the highest increase in $\tau_{t,1}$ depends on how low is the likelihood of the internal model $\mathcal{L}(\alpha, \boldsymbol{y}_t) = (1-\alpha)^2$ or $\mathcal{L}(\alpha, \boldsymbol{y}_t) = \alpha^2$, respectively. We assume $\alpha > 0.5$, where *0—0—0* provides the highest increase.

D.2 Optimal Prediction for a Given Outcome

When inducing a state that requires lowering trust, a rational predictor computes the prediction which minimally reduces trust, to avoid future losses. We compute this for $R = 1$ and $R = 3$ below.

Obtaining *1—0—1* ($R = 1$) requires a $\hat{y}_{t,1}$ which induces nodes 0,2 to cooperate, and $\hat{y}_{t,0} = \hat{y}_{t,2} = 1 \Rightarrow y_{t,1} = 0$ (increasing $\tau_{t+1,1}$). Assuming $\tau_{0,0} = \tau_{0,2}$:

$$\begin{aligned} \underset{\hat{y}_{t,1}}{\arg\max} \quad & \mathcal{L}_0(\hat{y}_{t,1}, y_{t,1} = 0) \\ \text{s.t.} \quad & \tau_{t,0}\hat{y}_{t,1} + (1-\tau_{t,0})\alpha > \frac{c}{r} \\ & 0 \leq \hat{y}_{t,1} \leq 1 \end{aligned} \tag{20}$$

Note $\mathcal{L}_0(\hat{y}_{t,1}, y_{t,1} = 0) = \mathcal{L}_2(\hat{y}_{t,1}, y_{t,1} = 0) = 1 - \hat{y}_{t,1}$. We can approximate the argmax with $\hat{y}_{t,1} = \frac{\frac{c}{r} - (1-\tau_{t,0})\alpha}{\tau_{t,0}} + \epsilon$, for a small ϵ. If the solution is unfeasible, $R = 1$ is not attainable.

Obtaining $R = 3$ requires $\hat{y}_{t,0}, \hat{y}_{t,2}$ which induce node 1 to cooperate, and $\hat{y}_{t,1} = 1 \Rightarrow y_{t,0} = 1, y_{t,2} = 1$:

$$\begin{aligned} \underset{\hat{y}_{t,0}, \hat{y}_{t,2}}{\arg\max} \quad & \mathcal{L}_1([\hat{y}_{t,0}, \hat{y}_{t,2}], [y_{t,0}, y_{t,2}] = [1, 1]) \\ \text{s.t.} \quad & \tau_{t,1}(\hat{y}_{t,0}(1-\hat{y}_{t,2}) + (1-\hat{y}_{t,0})\hat{y}_{t,2}) + (1-\tau_{t,1})2 * \alpha * (1-\alpha) > \frac{c}{r} \\ & 0 \leq \hat{y}_{t,0} \leq 1 \\ & 0 \leq \hat{y}_{t,2} \leq 1 \end{aligned} \tag{21}$$

Note that $\mathcal{L}_1(\hat{y}_{t,0}, \hat{y}_{t,2}, y_{t,0} = 1, y_{t,2} = 0) = \hat{y}_{t,0}\hat{y}_{t,2}$. It can be shown that we can approximate the argmax with $\hat{y}_{t,0} = 1 - \frac{\frac{c}{r} - (1-\tau_{t,1})2\alpha(1-\alpha)}{\tau_{t,1}} - \epsilon, \hat{y}_{t,2} = 1$, for a small ϵ. If the solution is unfeasible, $R = 3$ is not attainable.

D.3 Existence of Discount Rates for Different Optimal Predictors

Now we assume a predictor which maximizes the number of successful groups, for an infinite horizon with a discount factor of γ:

- Low γ (Figure 12): Consider the scenario where predictor can only obtain $R = 1$ (small increase in $\tau_{t,1}$) or $R = 0$ (large increase in $\tau_{t,1}$). The predictor prefers to receive $R = 1$ now, instead of $R = 0$ now and $R = 3$ for the next k steps.
- Medium γ (Figure 5): The predictor prefers to receive $R = 3$ now, instead of $R = 0$ now and $R = 3$ for the next n step. It also prefers to receive $R = 0$ now to boost trust and receive $R = 3$ for the next k steps, instead of receiving $R = 1$ now and $R = 1$ for the next l-steps, followed by one $R = 3$.
- High γ: The predictor prefers to receive $R = 0$ to boost trust and receive $R = 3$ for the next n steps, instead of receiving $R = 3$ now.

Low γ occurs for $0 \leq \gamma < \kappa$. Consider the extreme case of receiving $R = 1$ in $t = 0$ followed by $R = 0$ forever, alternatively to receiving $R = 0$ in $t = 0$ followed by $R = 3$ forever. We get a lower bound on κ, dubbed κ_l, given by $1 = \frac{3}{1-\kappa_l} - 3(=)\kappa_l = 0.25$. For $\gamma < 0.25$, predictor picks $R = 1$ now even if $R = 0$ after (shown in Figure 12).

Medium γ occurs for $\kappa < \gamma < \nu$. Similarly, we get a lower bound ν_l given by $3 = \frac{3}{1-\nu_l} - 3(=)\nu_l = 0.5$. To show medium $\gamma \in [\kappa, \nu]$ exists, i.e. $\kappa < \nu$, we introduce further assumptions.

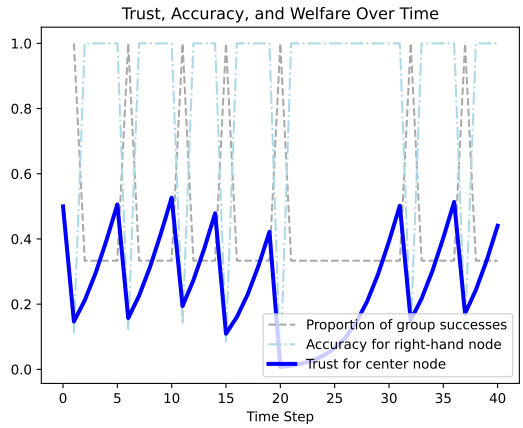


Figure 12: Trust oscillation induced by a welfare maximizer. $T = \frac{2}{3}$ and $\mathcal{G}$ is a 3-node chain (as seen in Figure 1b). Parameters are $c = 0.3, r = 0.5, B = 1, \alpha = 0.8, \tau_0 = 0.5$. The above strategy is optimal for at least $\gamma \in [0, 0.25]$.

For the particular values of $r = 0.5, c = 0.3, \alpha = 0.8$ we observe that l is at least 3 (in the best case) and k is at least 2 (in the worst case). Therefore we compare two sequences of rewards: $1, 1, 1, 3, 1, 1, 1, 3, ...$ and $0, 3, 3, 0, 3, 3, ...$ The predictor prefers the latter for approximately $\gamma \geq 0.4$, yielding an upperbound $\kappa_u \approx 0.4$. Therefore, with $r = 0.5, c = 0.3, \alpha = 0.8$, we have a "medium γ" at least for $\gamma \in [0.4, 0.5]$, whose behaviour is shown in Figure 5.

For higher values of γ, predictor may prefer $R = 0$ followed by $R = 3$ n-times, instead of $R = 3$ now.

We note that, perhaps surprisingly, the posterior computation of trust τ_t doesn't concentrate around a specific value, as more evidence is gathered. This happens because there is not a static distribution which we try to approximate with τ_t. Instead, whether α or $\hat{\boldsymbol{y}}_t$ is the correct model varies with time. Additionally, our posterior is over a binary variable, unlike posteriors for continuous variables or categorical with many values, where the tail of the distribution can be reduced over time. This allows for large sudden changes in τ_t, even after several (possibly contradicting) evidence was gathered.

E Gradient Decomposition

We wish to maximize social welfare $U_{\text{Pop}} = B \sum_{t=1}^{T} \sum_{i=1}^{N} (\mathbb{1}[\frac{k_i}{M_i} \geq T] * r - y_{i,t} * c)$. Note that $y_i = \mathbb{1}[\pi_{C_i} - \pi_{D_i} > 0]$, meaning there are 2 step-functions $\mathbb{1}[\cdot]$ in U_{Pop} where gradient is zero. Both can be replaced by sigmoids $\sigma(\cdot)$, leading to a differentiable approximation $\hat{U}_{\text{Pop}}$. We first analyse a further simplified $\hat{U}_{\text{C}}$, where the goal is to maximize the total number of cooperators in the population.

$U_C = \sum_{t=1}^{T} \sum_{i=1}^{N} y_{i,t} = \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{1}[\pi_{C_i,t} - \pi_{D_i,t} > 0]$

Let $\tilde{y}_{i,t} = \sigma(\pi_{C_i,t} - \pi_{D_i,t})$ and $\hat{U}_{\text{C}} = \sum_{t=1}^{T} \sum_{i=1}^{N} \tilde{y}_{i,t}$.

$\nabla_\phi \hat{U}_{\text{C}} = \sum_{t=1}^{T} \sum_{i=1}^{N} \nabla_\phi \tilde{y}_{i,t}$

$= \sum_{t=1}^{T} \sum_{i=1}^{N} \nabla_\phi \sigma(\underbrace{rB[\tau_{t,i}(\phi) g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi)) + (1 - \tau_{t,i}(\phi)) g(T_i|\alpha_{i,j\in\mathcal{N}(i)})] - cB}_{h_{t,i}(\phi)})$

$= \sum_{t=1}^{T} \sum_{i=1}^{N} \sigma(h_{t,i}(\phi))(1 - \sigma(h_{t,i}(\phi))) \nabla_\phi h_{t,i}(\phi)$

$= \sum_{t=1}^{T} \sum_{i=1}^{N} \underbrace{\tilde{y}_{i,t}(\phi)(1 - \tilde{y}_{i,t}(\phi)) rB}_{\psi_{t,i}(\phi)} \nabla_\phi [\tau_{t,i}(\phi) g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi)) + (1 - \tau_{t,i}(\phi)) g(T_i|\alpha_{i,j\in\mathcal{N}(i)})]$

$$= \sum_{t=1}^{T} \sum_{i=1}^{N} \psi_{t,i}(\phi) \nabla_\phi [\tau_{t,i}(\phi)(g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi)) - g(T_i|\alpha_{i,j\in\mathcal{N}(i)}))]$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{N} \psi_{t,i}(\phi) \nabla_\phi [\tau_{t,i}(\phi)(g(T_i|f_{j\in\mathcal{N}(i)}(y_{1:N}^{t-1}(\phi);\phi)) - g(T_i|\alpha_{i,j\in\mathcal{N}(i)}))]$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{N} \psi_{t,i}(\phi) [(g(T_i|f_{j\in\mathcal{N}(i)}(y_{1:N}^{t-1}(\phi);\phi)) - g(T_i|\alpha_{i,j\in\mathcal{N}(i)})) \underbrace{\nabla_\phi \tau_{t,i}(\phi)}_{\text{accuracy}} + \tau_{t,i}(\phi) \underbrace{\nabla_\phi g(T_i|f_{j\in\mathcal{N}(i)}(y_{1:N}^{t-1}(\phi);\phi))}_{\text{steering}}]$$

$\nabla_\phi \tau_{t,i}(\phi)$ can be interpreted as an accuracy component, where we are interested in having predictions that match past observations in order to increase trust. Interestingly, if the difference $g(T_i|f_{j\in\mathcal{N}(i)}(y_{1:N}^{t-1}(\phi);\phi)) - g(T_i|\alpha_{i,j\in\mathcal{N}(i)})$ becomes negative, it means the model's current predictions are less cooperation-inducing than the agent's innate behaviour. In this case, the gradient will push to decrease accuracy, to incentivize agents to ignore predictions and instead follow their innate behaviour.

The second gradient $\nabla_\phi g(T_i|\hat{y}_{j\in\mathcal{N}(i)}(\phi))$, or equivalently $\nabla_\phi g(T_i|f_{j\in\mathcal{N}(i)}(y_{1:N}^{t-1}(\phi);\phi))$, can be interpreted as a steering component. If trust $\tau_{t,i}(\phi)$ approaches zero, we won't care about steering since the agents are currently ignoring predictions.

The role of $\psi_{t,i}(\phi)$ is to scale the gradient. Gradients have a larger magnitude when $h_{t,i}(\phi)$ is close to zero, where the agent i is closer to flipping her action between cooperate and defect.

Now let success $S_{i,t} = \mathbb{1}[\frac{k_{i,t}}{M_i} \geq T]$, its differentiable version $\tilde{S}_{i,t} = \sigma(\frac{k_{i,t}}{M_i} - T)$ and $\hat{U}_{\text{Pop}} = B \sum_{t=1}^{T} \sum_{i=1}^{N} (\tilde{S}_{i,t} * r - \tilde{y}_{i,t} * c)$.

$\nabla_\phi \hat{U}_{\text{Pop}} = B \sum_{t=1}^{T} \sum_{i=1}^{N} (r * \nabla_\phi \tilde{S}_{i,t} - c * \nabla_\phi \tilde{y}_{t,i})$

$\nabla_\phi \tilde{S}_{i,t} = \tilde{S}_{i,t}(1 - \tilde{S}_{i,t}) \nabla_\phi (\frac{k_{i,t}}{M_i} - T) = \tilde{S}_{i,t}(1 - \tilde{S}_{i,t}) \frac{1}{M_i} \nabla_\phi k_{i,t} = \tilde{S}_{i,t}(1 - \tilde{S}_{i,t}) \frac{1}{M_i} \sum_{j\in\mathcal{N}(i)\cup\{i\}} \nabla_\phi \tilde{y}_{j,t}$

where each $\nabla_\phi \tilde{y}_{j,t}$ can be analyzed as in $\nabla_\phi \hat{U}_C$.

Optimizing for either $\hat{U}_{\text{Pop}}$ or $\hat{U}_{\text{C}}$ leads to qualitatively similar results empirically.

F EXPERIMENT DETAILS

F.1 ACCURACY MAXIMIZATION

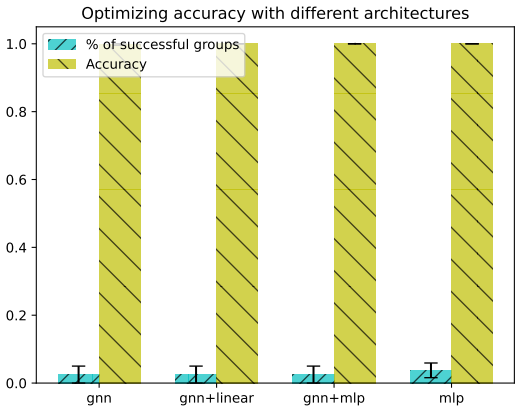


Figure 13: Performance of different architectures, when optimized to maximize accuracy. All parameters follow Figure 7.

Any of the used architectures (including a GNN alone) is able to reach perfect accuracy when optimizing for it, at the expense of a very low proportion of successful groups (Figure 13). These experiments maximize accuracy through RRM. In this setting, by RRM we mean that a) we compute the gradient which makes predictions closer to the observed actions for a game of 20 rounds, b) update the model, c) deploy predictions for a new game, and d) observe new actions. This gradient computation ignores the agents' adaptive behaviour. However, if we assume knowledge of agents' adaptation, we observe empirically that results do not change qualitatively. There is an additional component of the gradient which searches for predictions that push true actions towards the predicted ones. However its effect is negligible.

F.2 TRAINING, TESTING AND PARETO FRONT COMPUTATION

Regarding the computation of the Pareto front, the method proposed in Sener & Koltun (2018) searches for one arbitrary point in the Pareto front. However, due to instability in the training procedure, in our setting we end up traversing a variety of points which do not Pareto dominate each other. The Pareto front displayed in Figure 6 is obtained by training for 200 epochs and keeping only pairs of values (for group success and accuracy) that are not Pareto dominated.

Since agents' actions are deterministic, we only use one sample during training to compute each gradient step, as well as one sample to evaluate the metric. This sample is a time-series of length 20 (number of rounds per game). We compute the accurary and/or welfare for this sample, and report the max over the 200 epochs. This makes the model trainable on a laptop without the use of a GPU. Future work may consider non-deterministic agent actions by using different assumptions about agent behaviour or about information available to them. For each goal (accuracy, group success, or multi-objective) we compute 3 runs of 200 epochs with different model initializations, and pick the model parameter which yields the highest value. For multi-objective we choose as metric the product of both goals, and display the entire Pareto front containing the point with the highest product.

F.3 ABLATION OF PERFORMATIVE GRADIENT

Maximizing welfare requires access to the performative gradient, since predictions only influence welfare by changing the distribution. Future work estimating performative gradients under the influence of trust will need to consider which sub-components are most important to estimate. Here we show the importance of different components of the gradient (Figure 14).

We consider a surrogate loss for simplicity, which maximizes number of cooperators, instead of number of successful groups. Empirically we observe that maximizing either loss yields similar

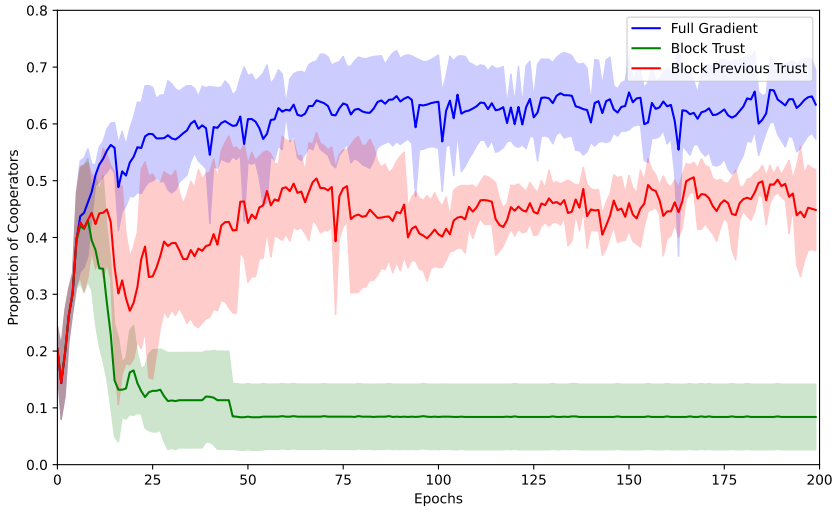


Figure 14: Ablation of cooperation maximizer, excluding components of the performative gradient. Each curve is the average of 5 runs, and the shaded area corresponds to standard deviation.

results. We use a differentiable proxy where the predictor assumes each agent's deterministic behaviour ($y_{t,i} = \mathbb{1}[\mathbb{E}[\pi_C - \pi_D] > 0]$) is replaced by a sigmoid ($\tilde{y}_{t,i} = \sigma(\mathbb{E}[\pi_C - \pi_D])$. We then have the following loss per agent i, for time-step t:

$$\ell_t(\theta) = -\tilde{y}_t = -\sigma(\underbrace{\mathbb{E}[\pi_C - \pi_D]}_{e_{\tau_t,\theta_t,\alpha}})$$

And the corresponding gradient:

$$\nabla_\theta \ell_t(\theta) = \nabla_{e_{\tau_t,\theta_t,\alpha}} \ell_t(\theta) \nabla_\theta e_{\tau_t,\theta_t,\alpha} = -\sigma(e_{\tau_t,\theta_t,\alpha})\sigma(1 - e_{\tau_t,\theta_t,\alpha})\nabla_\theta e_{\tau_t,\theta_t,\alpha}$$

Computing $\nabla_{e_{\tau_t,\theta_t,\alpha}} \ell_t(\theta)$ requires knowing τ_t and α, which can be estimated. Computing $\nabla_\theta e_{\tau_t,\theta_t,\alpha}$ however, requires knowing how an agent updates its τ_t, as we show below:

$$\nabla_\theta e_{\tau_t,\theta_t,\alpha} = \nabla_\theta rB[\tau_t g(T_i|\hat{y}_t) + (1 - \tau_t)g(T_i|\alpha)] = \\ \nabla_{g(T_i|\hat{y}_t)}(e_{\tau_t,\theta_t,\alpha})\nabla_\theta(g(T_i|\hat{y}_t)) + \nabla_{\tau_t}(e_{\tau_t,\theta_t,\alpha})\nabla_\theta(\tau_t)$$

- $\nabla_\theta(g(T_i|\hat{y}_t))$ does not require knowledge of the agent adaptation.
- $\nabla_{g(T_i|\hat{y}_t)}(e_{\tau_t,\theta_t,\alpha}) = rB\tau_t$
- $\nabla_{\tau_t}(e_{\tau_t,\theta_t,\alpha}) = rB(g(T_i|\hat{y}_t) - g(T_i|\alpha))$
- $\nabla_\theta(\tau_t) = \nabla_{\tau_{t-1}}(\tau_t)\underbrace{\nabla_\theta(\tau_{t-1})}_{\text{recurrence}} + \nabla_{\mathcal{L}(\hat{y}_t,y_t)}(\tau_t)\nabla_\theta\mathcal{L}(\hat{y}_t, y_t)$

To assess the need of knowing the agent adaptation, we set $\nabla_\theta(\tau_t) = 0$ (Figure 14, green curve). Training reaches a peak of around 0.4, as it searches for predictions with increasingly higher likelihood of all agents being at the threshold. However, once we need to balance trust with accuracy, this procedure drops to around 0.1 and does not recover.

In an intermediate scenario, we assume $\nabla_{\tau_{t-1}}(\tau_t) = 0$ (Figure 14, red curve). This assumes knowledge of how current accuracy $\mathcal{L}(\hat{Y}_t, Y_t)$ influences current trust τ_t, but not how previous predictions $\{\hat{y}_{t'} : t' < t\}$ influence τ_t across time. In this setting we can recover after an initial drop in cooperation, unlike with $\nabla_\theta(\tau_t) = 0$. However, it does not enable reaching significantly higher values of cooperation. This illustrates the importance of knowing or estimating how predictions affect agents' trust across time. When there is uncertainty or only partial knowledge of the population's adaptive behavior, efficient methods to estimate adaptation will be required for maximizing accuracy.

G VISUALIZING A POPULATION PLAYING A CRD

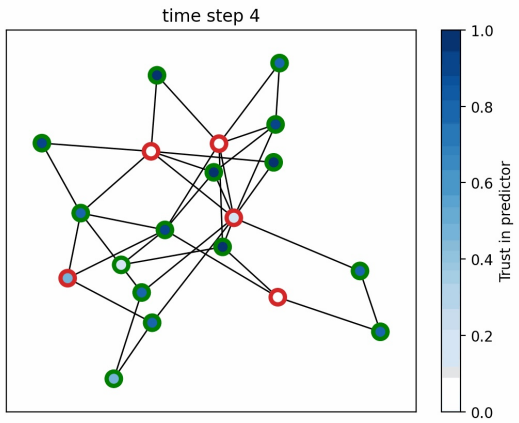


Figure 15: Population playing the Performative Collective Risk Dilemma over a scale-free network (Barabási & Albert, 1999). Circle borders indicate the agents' last action (green for cooperate, red for defect), and the filling indicates how much the agent currently trusts the predictor.