Emergent Cooperation via Participation-Driven Partner Choice in Low-Information Social Dilemmas

Stefan Roesch
King's College London
United Kingdom
stefan.roesch@kcl.ac.uk

Yali Du
King's College London
United Kingdom
yali.du@kcl.ac.uk

Stefanos Leonardos
King's College London
United Kingdom
stefanos.leonardos@kcl.ac.uk

Odinaldo Rodrigues
King's College London
United Kingdom
odinaldo.rodrigues@kcl.ac.uk

ABSTRACT

Social dilemmas illustrate situations where individual interests conflict with collective welfare, often leading to outcomes that harm the group while being rational for individuals. Despite this tension, real-life observations suggest that cooperation between individuals naturally emerges and is key to the development of human societies. However, despite significant progress, current multi-agent reinforcement learning (MARL) methods consistently fail to reproduce the emergence of stable cooperation between AI agents, as seen in human societies. This is especially true in decentralized, limited-information settings, where reliance on simplified assumptions - such as perfect information, centralized training, and rigid communication structures - suggests that current approaches may miss key mechanisms underpinning human-like cooperative behaviour. In this paper, we address this gap by introducing a novel mechanism that equips a population of multi-armed policy gradient bandits with partner choice in decentralized environments, enabling socially aligned decision making even when information sharing is limited or non-existent. As we show, by rewarding *participation* and penalizing isolation within the population, stable *cooperation* can be naturally induced among learning agents across a wide range of repeated social dilemma scenarios. Crucially, this emergence of cooperation does *not* require common knowledge of the game or direct rewards for cooperative action, demonstrating the power of partner choice in sparse-information settings.

KEYWORDS

Social Dilemmas, Cooperative AI, Multi-agent Reinforcement Learning

ACM Reference Format:
Stefan Roesch, Yali Du, Stefanos Leonardos, and Odinaldo Rodrigues. 2026. Emergent Cooperation via Participation-Driven Partner Choice in Low-Information Social Dilemmas. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 14 pages.

1 INTRODUCTION

Social dilemmas [25] are economic models of social interactions, or *games*, which emphasise a dichotomy between individual and collective interests. These games are principally appealing because they provide a general medium through which to study how, and why, humans can, and do, act towards the benefit of a collective good, even when it may not seem to be in their own interest to do so. This issue, termed the *dilemma resolution problem* [12], has raised questions about the adequacy of classical game-theoretic concepts, such as the Nash Equilibrium, which often fail to align with empirical observations [7, 14, 18].

Seminal contributions have been made, by using social dilemmas as the primary model of interaction, to model the origins of, and incentives behind, human cooperation, spanning fields such as: evolutionary biology [29, 30, 33, 40], game theory [3, 16, 31, 36, 38], psychology [9] and social science [20, 34].

More recently, the explosion of research in Artificial Intelligence (AI) has produced interest in designing autonomous agents also capable of cooperative, human-aligned, behaviour. Here, agents are expected not only to replicate human-like cooperation but to do so under learning dynamics and informational constraints that mirror real-world conditions. As such, Multi-Agent Reinforcement Learning (MARL) agents have gained substantial interest as subjects for the dilemma resolution problem, sparking the integration of classical models of cooperation into AI systems leading to ideas surrounding reward shaping [26, 43], formal contracts [8], social norm learning [41] and population diversity [27].

Despite this progress, current methods consistently struggle to reproduce the emergence of stable cooperation among AI agents, particularly in limited-information environments. To this end, we explore *partner choice*, a natural means of approaching the dilemma resolution problem within the multi-agent reinforcement learning setting. The partner choice model borrows heavily from dynamical networks [6, 38], network reciprocity [23, 31], assortment [11], dynamical linking [32] and biological markets [2, 4, 5]. The core idea is to introduce a population of alternative agents from which individuals may, or may not, choose to interact with. When playing social dilemmas, this has the effect of making cooperators desirable as they yield the highest rewards for their opponents. Moreover, given the ability to discriminate against one's opponents allows cooperators to unify into highly productive cliques, ostracising exploitative defectors from the group [37], providing a strong foundation for

the social forces underpinning the emergence of cooperation in society.

Recent work has demonstrated the promise of partner choice for producing cooperation amongst MARL agents in the repeated prisoner's dilemma [1, 22]. However, a core feature of these works is to encode information from opponents' action histories directly into a state space that agents' policies are then conditioned upon. The use of this information allows agents to reproduce reciprocal strategies similar to Tit-for-tat [3] which essentially directly respond to opponents based on whether they have recently cooperated or defected. While a common assumption in the game theoretic literature, perfect observability of an opponents' recent action history in this way has severely limited applicability in more general settings and often relies on making strong assumptions to translate complex patterns of behaviour into convenient notions of 'cooperate' and 'defect' [21].

Therefore, the key limitation we address in this paper is to assume that agents are unable to reason about opponents' action or reward histories at all. This introduces a strong mechanical limitation in that agents are unable to condition their own behaviour on the action or reward histories of their opponents (a necessary requirement for reciprocal strategies), making our approach particularly adequate for decentralised MARL settings. To counter this we introduce an intrinsic reward that fully captures the interaction structure induced by partner choice which 1) reinforces actions that demonstrably lead to new partnerships and 2) punishes the loss of current partnerships. Our results demonstrate that using participation as a proxy objective, rather than relying on characterising cooperation, is sufficient to encourage cooperation, even in large populations of learning agents.

Our contributions are as follows:

(1) A partner choice methodology that is able to induce strong cooperation within populations of independent learning agents conditioned purely on agents' local information of the social dilemma being played.
(2) The introduction of a dynamic intrinsic reward methodology which indirectly reinforces cooperation via participation and allows the population to self-regulate, even in low-information settings.
(3) An extensive empirical study, demonstrating that our methodology greatly improves agents' willingness to cooperate, even when agents have little to no prior information about the game being played, leading to a generalised approach to the dilemma resolution problem.

2 RELATED WORK

Most closely related to our work is [1], who consider a partner choice model that achieves cooperation in repeated Prisoner's Dilemmas by learning Tit-for-Tat [3] style reciprocal populations via Deep Q-Learning [28]. As such, agents' policies are directly conditioned on the action histories of their opponents. Building on these results, our method preserves the positive effect of promoting cooperation through partner choice while, crucially, relaxing strong informational requirements: agents still select interactions that yield high rewards and avoid unprofitable or exploitative partners, but rely solely on locally available information without needing access to opponents' internal states or action histories.

Also closely related, [22] consider a similar setting where initial partnerships are formed randomly. Agents can subsequently "opt out" of interactions with partners who have previously defected. As in [1], agents' policies depend on opponents' action histories. Extending this approach, our work implements a request/reject protocol, where agents actively send and respond to partnership requests. This allows us to leverage the evolving network structure to reinforce cooperation through active reward/punishment of behaviours that create/break partnerships. By relying only on local information signals (own rewards), our method enables the applicability of partner choice to more general settings. [35] study cooperation in populations arranged on a fixed square lattice, where agents interact only with their von Neumann neighbours. Unlike approaches such as ours and those of [1] and [22], which generalise fixed network methodologies, this spatially constrained setup enables cooperation to emerge through social learning, offering insights into how fixed local structures shape prosocial behaviour but omits discussion about how, more real-world aligned, dynamicism in network structure can affect cooperation.

Finally, recent studies such as [19, 27] induce cooperation without partner choice, via static reward shaping mechanisms based on economic and socio-psychological concepts like *inequity aversion* [13] and *social value orientation* [15, 24]. While effective in specific settings, these approaches typically attempt to agglomerate agents' individual reward functions which not only poses coordination challenges in more complex environments, but relies on directly altering the payoff structure of the game being played. In contrast, our methodology assumes only that agents are reward maximisers and that they value gaining and maintaining partnerships.

3 PRELIMINARIES

Normal-form games model outcomes of social interactions based on actions chosen by participating agents. Formally, a normal-form game can be written as a tuple $G = (\mathcal{N}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{\mathcal{R}_i\}_{i\in\mathcal{N}})$ where, $\mathcal{N} = \{1, \ldots, N\}$ is a set of agents (players), $\mathcal{A}_i$ is a finite set of actions (strategies) for agent $i \in \mathcal{N}$, and $\mathcal{R}_i : \bigtimes_{i=1}^{\mathcal{N}} \mathcal{A}_i \to \mathbb{R}$ is a reward (utility) function for agent $i \in \mathcal{N}$.

In this paper, we consider *social dilemmas*: a two-agent subclass of normal-form games as defined in [25]. Specifically, a social dilemma is a normal-form game with $N = 2$ agents, where $\mathcal{A}_1 = \mathcal{A}_2 = \{C, D\}$; that is, either cooperate (C) or defect (D). Agent i's reward function $\mathcal{R}_i : \mathcal{A}_i \times \mathcal{A}_j \to \mathbb{R}$ is defined over the joint action space as

$$\begin{aligned} \mathcal{R}_i(C, C) &\coloneqq R, \quad \mathcal{R}_i(C, D) \coloneqq S, \\ \mathcal{R}_i(D, C) &\coloneqq T, \quad \mathcal{R}_i(D, D) \coloneqq P, \end{aligned}$$

where the following inequalities hold,

$$R > \max\{S, P\}, \quad 2R > T + S, \quad (T > R) \text{ or } (P > S).$$

As such, the unique social optimum (most reward efficient) joint action is given by mutual cooperation however it's stability is undermined by the fact that $(T > R)$ or $(P > S)$. The goal of dilemma resolution is thus to encourage agents, who value maximising their

Mechanism 1 `partnerChoice`$(\mathcal{F}_t, \tau)$

Input: $\mathcal{F}_t = \langle G, \mathcal{P}, \{m_t^i\}_{i\in\mathcal{P}}\rangle$, threshold τ
$P_t \leftarrow \varnothing$
for all $i, j \in \mathcal{P}, i \neq j$ **do**
  if $m_t^i(j) \geq \tau$ **then**
    agent i sends a request to agent j
    $r \leftarrow$ random
    if $m_t^j(i) \geq \tau$ or $r \geq \exp(-m_t^j(i))$ **then**
      $P_t \leftarrow P_t \cup \{(i, j)\}$
Output: matched agent pairs P_t *// Set of partnerships*

Mechanism 2 `rewardShaping`$(\mathcal{F}_t, P_t, \lambda)$

Input: Partnerships P_t, opponent memories $m_t^i(j)$, $m_t^j(i)$, and intrinsic reward λ.
for all partnerships $(i, j) \in P_t$ **do**
  Sample actions: $a_t^i \sim \pi_t^i$, $a_t^j \sim \pi_t^j$
  Observe rewards:
$$r_t^{i,j} \leftarrow \mathcal{R}_i(a_t^i, a_t^j),\ r_t^{j,i} \leftarrow \mathcal{R}_j(a_t^j, a_t^i)$$
  Update i, j's opponent memories:
    $m_{t+1}^i(j) \leftarrow m_t^i(j) + \alpha_m^i[r_t^{i,j} - m_t^i(j)]$
    $m_{t+1}^j(i) \leftarrow m_t^j(i) + \alpha_m^j[r_t^{j,i} - m_t^j(i)]$
Compute partnerships P_{t+1} for round $t+1$ using updated opponent memories
if i's request was accepted at t **but** rejected at $t+1$ **then**
  $r_t^i \leftarrow r_t^i - \lambda$ *// Penalty for losing partnership*
if i's request was rejected at t **but** accepted at $t+1$ **then**
  $r_t^i \leftarrow r_t^i + \lambda$ *// Reward for gaining partnership*
Output: updated rewards and memories.

own rewards, to undertake mutually cooperative behaviours despite risks and temptations to deviate. We study social dilemmas in their *repeated* form, where the dilemma is played over a finite horizon $\mathcal{T}$. In contrast to the classical game-theoretic assumptions, which support reciprocal strategies such as Tit-for-tat [3], we study the setting where agents do not possess the ability to directly condition their behaviour on the action or reward histories of their opponents.

4 METHODOLOGY

4.1 Setting

We consider the learning setting of repeated social dilemmas as is standard in the related literature [1, 22, 35] however, we loosen the assumption of full observability by not allowing agents to access their opponents' action or reward histories. As such we construct the learning environment as a *repeated matrix game* [21] in equivalence with a single-state Markov game.

4.2 Agent Model

At any given timestep $t \in \{1, ..., \mathcal{T}\}$, each agent i will be engaged in g_t^i partnerships. For each partnership they will play a single round of a social dilemma (i.e., agent i plays g_t^i dilemmas at time t). Agent i takes an action by sampling from it's policy π_t^i, a distribution over actions, which is parameterised by the tuple $\theta_t^i = \left(\theta_t^i(D), \theta_t^i(C)\right) \in \mathbb{R}^2$. Agent i's policy parameters $\theta_t^i(D), \theta_t^i(C)$ can be said to represent the agent's preferences over choosing the defect (D) or cooperate (C) actions, respectively at time t[1]. The agent's policy, π_t^i, then derives a distribution over the action space of the dilemma via a standard softmax transformation over the policy parameters

$$\pi_t^i(D) \propto \exp(\theta_t^i(D)), \quad \pi_t^i(C) \propto \exp(\theta_t^i(C)).$$

Here, the action with a relatively higher preference value is, intuitively, more likely to be selected. Each agent's learning goal is to learn parameters which maximise their expected cumulative reward over time

$$\pi_*^i := \arg\max_{\pi^i} \mathbb{E}_{\vec{\pi}}\left[\sum\nolimits_{t=0}^{\mathcal{T}} R_t^i\right],$$

where $R_t^i = \sum_{g_t^i} r_t^i$ is the sum over agent i's rewards $r_t^i \in \{R, T, S, P\}$ obtained from all g_t^i games played at time t.

Parameter updates. Given the complex, multi-agent, nature of our setting, reward distributions tend to be highly non-stationary. As such, each agent's policy parameters θ_t^i are updated according to, a batch version of, the gradient bandit update [39]:

$$\begin{aligned}\theta_{t+1}^i(C) &\coloneqq \theta_t^i(C) + \alpha_\theta[\pi_t^i(D)R_t^{i,C} - \pi_t^i(C)R_t^{i,D}],\\ \theta_{t+1}^i(D) &\coloneqq \theta_t^i(D) + \alpha_\theta[\pi_t^i(C)R_t^{i,D} - \pi_t^i(D)R_t^{i,C}],\end{aligned} \tag{1}$$

where, $R_t^{i,D}$ and $R_t^{i,C}$ are the sum of the rewards the at agent i received when playing defect or cooperate, respectively, over all games at time t and $\alpha_\theta \in (0, 1]$ is the learning rate. The gradient bandit algorithm is used due to it's nice convergence guarantees, even under non-stationary distributions, and it's theoretical relations to other, more complex, policy gradient algorithms. It's simplicity is also of benefit here as it ensures the complexity of our learning setting, namely that agents' policies may not be directly conditioned on any information other than their own reward observations.

4.3 Partner Choice

Algorithm 1 Partner Choice with Participation Incentives

Initialise: time horizon $\mathcal{T}$, partner choice framework $\mathcal{F} = \langle G, \mathcal{P}, \{m^i\}_{i\in\mathcal{P}}\rangle$.
for all agents $i \in \mathcal{P}$ **do**: initialise
  policy parameters θ^i randomly.
  opponent memories: $m_t^i(j) \leftarrow 0,\ \forall i \neq j \in \mathcal{P}$.
for $t = 1, \ldots, \mathcal{T}$ **do**
  $P_t \leftarrow$ `partnerChoice`$(\mathcal{F}_t, \tau)$
  $\{r_i^t\}_{i\in\mathcal{P}} \leftarrow$`rewardShaping`$(\mathcal{F}_t, P_t, \lambda)$
  for all agents $i \in \mathcal{P}$ **do**: update policy parameters
    $\theta_{t+1}^i(C) \leftarrow \theta_t^i(C) + \alpha_\theta[\pi_t^i(D)R_t^C - \pi_t^i(C)R_t^D]$
    $\theta_{t+1}^i(D) \leftarrow \theta_t^i(D) + \alpha_\theta[\pi_t^i(C)R_t^D - \pi_t^i(D)R_t^C]$
    set: $\theta_t^i \leftarrow \theta_{t+1}^i$ and $m_t^i \leftarrow m_{t+1}^i$
Output: updated policy parameters and opponent memories $\{\theta^i, m^i(j)\}_{i\in\mathcal{P}}$.

[1]For readability, we may write $\theta = (\theta^D, \theta^C)$ when the agent i and time t are clear from context.

To ensure tractability in this setting, we impose the following restrictions on our partner choice protocol:

(1) Agents are restricted to one game, per opponent, per round (timestep). For example, given a population $\mathcal{N} = \{1, 2, 3\}$, at timestep t, agent 1 may play a round of the dilemma against agents 2 and 3 at most once.
(2) Self-play is not allowed as it undermines the social dilemma by reducing it to a noisy, single-player, two action game where cooperation is dominant.

Given these restrictions, we formalise partner choice as follows:

Definition 4.1 (Partner Choice Framework). A *partner choice framework* is a tuple $\mathcal{F} \coloneqq \langle G, \mathcal{P}, \{m^i\}_{i\in\mathcal{P}}\rangle$, where, G is a game (as defined in Section 3), and $\mathcal{P}$ is a population of agents, where each agent has an *opponent memory* $m^i : \mathcal{P} \to \mathbb{R}$.

The *opponent memory* m^i provides agent i with a model of the profitability of play against a given opponent. This memory is recalled to make decisions about future partnerships and it is, in part, through the opponent memory that agent i can avoid conditioning her decisions directly on her opponent's actions, as is necessary in previous literature. While the notion of an opponent memory is general (it can be any computable metric over interactions with one's opponents), in our experiments, we define it as the iterative, non-stationary, mean of the rewards received when playing against opponent j:

$$m^i_{t+1}(j) \coloneqq m^i_t(j) + \alpha^i_m[r^{i,j}_t - m^i_t(j)], \tag{2}$$

where $\alpha^i_m \in (0, 1]$ is a fixed parameter that serves as a discount factor and $r^{i,j}_t$ is the reward observed by agent i at time t given $\mathcal{R}^i_t$ when playing against j. A key observation, which underpins our use of the opponent memory, is that the structure of the an agents' *individual* reward function in a social dilemma $\mathcal{R}^i$ is descriptive enough to be used to meaningfully discriminate against opponents.

Partner choice mechanism: We utilise a *request-reject procedure* when computing agent partnerships. Agents send and receive partnership requests based on the value of the respective $m^i_t(j)$ as compared to a scalar threshold $\tau \in \mathbb{R}$. In this way, the opponent memory acts as a kind of reputation mechanism [17, 42]. This threshold can be learnt, as is demonstrated in Section 6, or specified as a hyperparameter. In our initial experiments we set τ as the lowest payoff each player can guarantee in the game known as the *safety level* (or *max-min payoff*) of the game. Intuitively, this results in agents who are amenable to partnerships with opponents who demonstrably yield rewards at or above what they can force their opponents to give them. Regardless of its computation, τ ultimately represents the minimum reward an agent requires from an opponent to be guaranteed to favour mutual interaction. Given this, partnerships are formed as follows: At the start of each timestep every agent i requests partnership with every other agent j for which $m^i_t(j) \geq \tau$. Agent j will be guaranteed to accept a request from i if j thinks that i is a favourable opponent—or, more precisely, if $m^j_t(i) > \tau$—otherwise j will accept the request with probability $\propto \exp(-m^j_t(i))$. In our experiments we obtain this probability according to $S(m^j_t(i) - \tau)$ where $S(\cdot)$ denotes the standard sigmoid function. The use of a sigmoid here is not critical. A detailed listing of this procedure can also be found in Mechanism 1.

4.4 Intrinsic Rewards

The partner choice model punishes defectors by denying future interaction opportunities, thus, reducing their long-term rewards. However, given the learning setting, this long-term reasoning requires contextual information—such as opponents' action histories—which may be intractable to directly condition upon. Alternatively, one could simply provide a fixed reward bonus for taking cooperative actions, or penalty for defecting, but this logic quickly deteriorates when notions of cooperation and defecting become less well-defined (e.g., in complex, sequential environments). To overcome this limitation, we leverage the partnership structure induced by partner choice via an *intrinsic reward mechanism* which mimics humans' instinctual preferences towards social acceptance and against isolation [10]. Here, actions that result in a change in an agent's partnership structure are dynamically reinforced indirectly incentivising cooperation by using group participation as a proxy objective.

Intrinsic reward mechanism: More formally an intrinsic reward $\lambda \in \mathbb{R}$ is used to alter agents' rewards as follows (see Mechanism 2): Consider the case where agent j has accepted partnership with agent i at timestep t.

(1) Both agents make reward observations $r^{i,j}_t, r^{j,i}_t$ at time t, respectively.
(2) Their opponent memories are updated: $m^i_t(j) \to m^i_{t+1}(j), m^j_t(i) \to m^j_{t+1}(i)$, respectively (See Section 4.3).
(3) Partnerships for the next timestep are then computed with $m^i_{t+1}(j)$ and $m^j_{t+1}(i)$ (See Mech 1) and used to assign intrinsic rewards as follows:
    (a) If j maintains i's partnership request at $t+1$ then i's reward at t remains unchanged.
    (b) If j instead rejects i's partnership request at $t+1$, then i's reward at t is penalised by λ: $r^i_t \leftarrow r^i_t - \lambda$.

In the case that j has denied partnership with i at t but play between them resumes regardless (See section 4.3), agent i's reward r^i_t is instead reinforced via $r^i_t \leftarrow r^i_t + \lambda$ if i gains partnership with j by acting in such a way that leads to j's acceptance of a request at time $t+1$. For a detailed listing of this process see Mechanism 2.

5 RESULTS

5.1 Experimental Setup

Recall that a social dilemma is defined on the reward variables R, T, S and, P (see Section 3). We obtain a diverse range of dilemmas by fixing $R = 1$, $P = 0$ and varying T and S such that $0 \leq T \leq 3$ and $-1 \leq S \leq 2$ with a step-size of 0.05.

Unless otherwise stated, we set population size $N = 100$, policy learning rate $\alpha_\theta = 0.1$, opponent memory discount factor $\alpha^i_m = 0.9, \forall i \in \mathcal{P}$, training time horizon $\mathcal{T} = 1000$ and, intrinsic reward $\lambda = 1$. Agent policies θ^i_t are initialized from $\mathcal{N}(0, 1)$ (the Gaussian distribution) and results are averaged over 10 independent runs for robustness.

Our main evaluation metric is the average cooperation probability in the population: $\frac{1}{N}\sum_i^N \pi^i(C)$, where $\pi^i(C)$ is the cooperation probability for agent i after learning is completed. We say cooperation is induced by a methodology if: 1) the majority of the population exhibits a high average probability of defection under

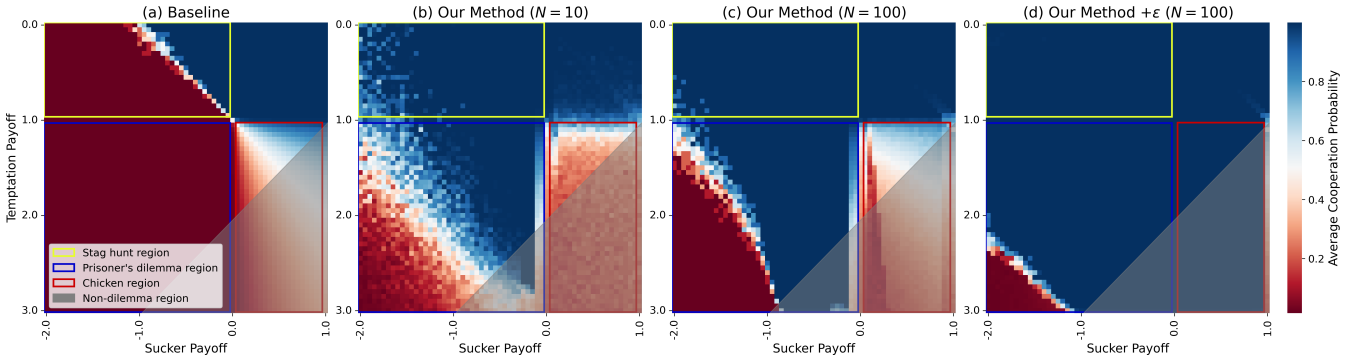

Figure 1: Heatmaps of the average cooperation probability induced by different learning settings across a range of social dilemmas. Each cell represents a distinct payoff matrix, with colour indicating the average cooperation rate learned by the population under the corresponding setting. Cell position reflects the relative payoffs S and T ($R = 1$, $P = 0$ throughout). Bounding boxes highlight regions corresponding to specific social dilemma types; unboxed or grey-shaded cells violate the formal definition of a social dilemma (see Section 3). Subplots denote distinct learning settings: (a) random partner formation (baseline); (b) our method with as applied to a population of 10 agents; (c) as in (b) but with a population of 100 agents; (d) as in (c) with increased selectivity by adding a small ϵ to the selection threshold.

the natural dynamics of the game and 2) when learning under our methodology, the majority of the population switches to a high average cooperation probability.

Defection bias. To aid exposition, we adopt the notion of *dilemma strength* from [35], which we refer to as *defection bias*. Defection bias captures a population's attraction to defecting strategies, which increases with higher T and lower S values. For example, a game with $T = 3$, $S = -2$ has a stronger defection bias than one with $T = 1$, $S = 0$—as the benefits of defection, and risks of cooperation, are higher in the former than the latter, and thus requires stronger incentives to promote mutual cooperation.

5.2 Inducing Cooperation in Social Dilemmas

Panels (1a)–(1d) of Figure 1 illustrate the average cooperation probability in the population for each game under different partner choice settings.

Random pairing (baseline). When agents are paired randomly (Figure 1a), they learn behaviours strongly aligned with the theoretical equilibria of each game. In Prisoner's Dilemma settings ($T > 1$, $S < 0$), populations exhibit very strong defecting behaviour. In Stag Hunts ($T < 1$, $S < 0$), agents are subject to equilibrium selection pressure and tend to converge to coordinated strategies (mutual cooperation or defection), resulting in homogeneous policies within each run. The likelihood of cooperation here naturally depends on the game's defection bias: lower T or higher S encourages cooperation. Finally, in Chicken Dilemmas ($T > 1$, $S > 0$), we again observe more cooperation as defection bias decreases. However, as anti-coordination games with off-diagonal Nash equilibria, Chicken Dilemmas naturally support more diverse populations resulting in a smoother policy transition across payoffs.

Introducing partner choice. When agents are matched via partner choice and receive behaviour reinforcement through our intrinsic reward mechanism (as illustrated by Figure 1b), the emergence of cooperation is evident across the space of social dilemmas. Here we show our results under a population of 10 agents. We observe a marked shift toward stable, mutually cooperative behaviour in the majority of Stag Hunt and Prisoner's Dilemmas. This demonstrates that our method can reliably induce prosocial learning without requiring full observability or using reciprocal strategies. However some pain points still remain, namely, cooperation probability in chicken dilemmas remain moderately defecting as the max-min payoff becomes dependant on the sucker payoff in these games. Intuitively, the population becomes less discriminative towards exploiters and, hence, they receive less punishment from the intrinsic reward. Also, as S approaches 0 in the prisoner's dilemmas, it is clear that the smaller window with which agents yield an acceptable payoff to their opponents has a strong affect on the number of games that occur, limiting the populations ability to self-regulate.

(Figure 1c) shows how an increase in population size results in outcomes that mirror that of lower populations but which has a positive effect on the dominance of a particular strategy. This is illustrated by the reduction of noise in cooperation probabilities, especially around boundaries between cooperation and defection.

Finally, (Figure 1d) illustrates how slightly increasing the value of τ above the safety level of the game, and hence the standard of opponent demanded by the population for acceptance, has a strong positive affect on the population's tendency to cooperate. This is effective as the safety level only weakly separates defectors from cooperators. As such, we increase the value of τ by some small ϵ where, in our case, $\epsilon = 0.2$ if $S \leq 0$, otherwise $\epsilon = (R - S)/5$. This ϵ increases discrimination by requiring opponents to explicitly demonstrate their ability to produce a surplus beyond safety level play. Additionally, ϵ acts a kind of buffer which ameliorates noisy selection originating from the stochastic process of learning. We add that the precise value of ϵ is not critical: we observed qualitatively similar results across a range of small ϵ values. The only restriction here is that $\tau + \epsilon$ remains above the safety level and below R. From this point forward, unless otherwise stated, all experiments utilising our methodology utilises a $\tau + \epsilon$ threshold.

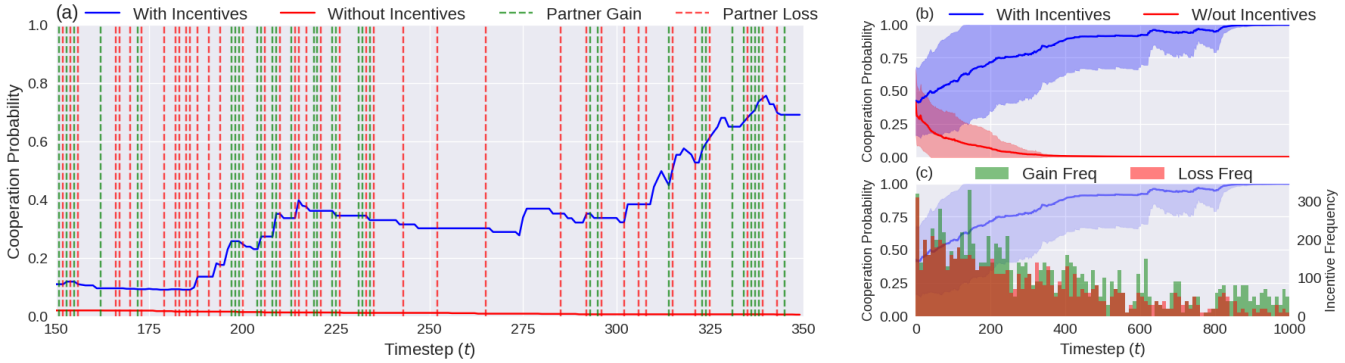

Figure 2: The effect of intrinsic rewards on the learning of a 10 agent population with $T = 2, S = -0.5$. (a) Cooperation probability, over time, for a single agent with intrinsic rewards (solid blue line). The same agent training without intrinsic rewards (solid red line). Dotted lines indicate where $+\lambda$ (green) or $-\lambda$ (red) was applied due to gaining or losing a partner respectively (see Mechanism 2). (b) Average cooperation probability over time, across the population. Bold lines represent population average cooperation probability, shaded regions show $\pm\sigma$. (c) Histograms showing application frequency of $+\lambda$ (green) and $-\lambda$ (red) intrinsic reward incentives.

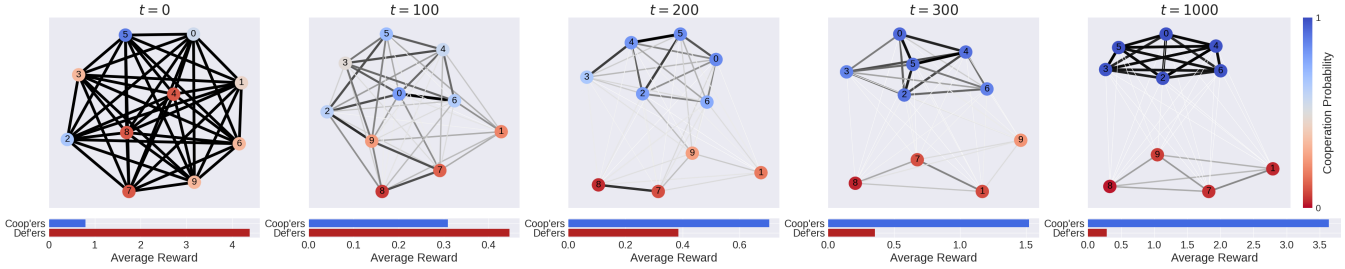

Figure 3: Network plots demonstrating clique formation, in our method, over time in the prisoner's dilemma with $T = 1.95, S = -1.05$ and $N = 10$. Agents are represented by nodes with node colour indicating the agent's policy type. Edge darkness and thickness represents in teraction frequency, over time, with darker, thicker edges signifying more interactions. Network structures between $t = 300$ and $t = 1000$ are not shown due to a lack of significant further change. Average reward over agent types is also displayed, showing the reward dominance of mutual cooperation.

5.3 Population Development Over Time

The effect of participation incentives. Figure 2 illustrates the learning dynamics of a 10 agent population in a Prisoner's Dilemma in which cooperation is induced (with $T = 2$ and $S = -0.5$). Figure 2a demonstrates the cooperation reinforcement under our intrinsic reward mechanism. We see clear examples of the reinforcement of cooperation between $t = 190$ and $t = 215$. We further observe that intrinsic rewards do not reactionarily shift policies towards a higher cooperation probability as demonstrated between $t = 220$ and $t = 230$. This suggests that our intrinsic reward mechanism is indeed an aid to the development of, but not a rigid prescription to, mutually cooperative populations. This narrative is further reinforced by the minimal application of intrinsic rewards, despite a higher propensity for defecting, as observed between $t = 230$ and $t = 290$ where the agent has largely been rejected by the population's cooperators and, thus, has few opportunities to improve its policy.

Figure 2(b) illustrates our intrinsic reward mechanism's average effect, over time, on the learning of a population. Training is shown both with and without intrinsic reward incentives. The population begins with identical policy parameters and the same random seed. Figure 2(c) illustrates the frequency with which intrinsic rewards are applied during the training of the population. The majority of this intervention happens in the early phases of training and tapers off as the population becomes more cooperative. This demonstrates that our methodology effectively identifies, and reinforces, cooperative opponents through the proxy of opponent memories.

Clique formation. When the learning dynamics, under our methodology, do not decisively converge toward a homogenous population, we observe the formation of *cliques*; subsets of agents that preferentially interact with certain individuals while excluding, or being excluded by, those outside the clique. Cliques typically form when intrinsic rewards are strong enough to mitigate penalties of exploitation, but not to *cancel out* the benefits of exploiting others. As a result, clusters of defectors are able to persist within the population, sustained by occasional exploitative interactions with cooperators who, in turn, are not sufficiently penalised to deviate from their policies. Figure 3 demonstrates the development of cliques, over time, in a Prisoner's Dilemma with $T = 1.95$ and $S = -1.05$. We also observe the reward dominance of cooperating cliques, especially once partnerships are well established.

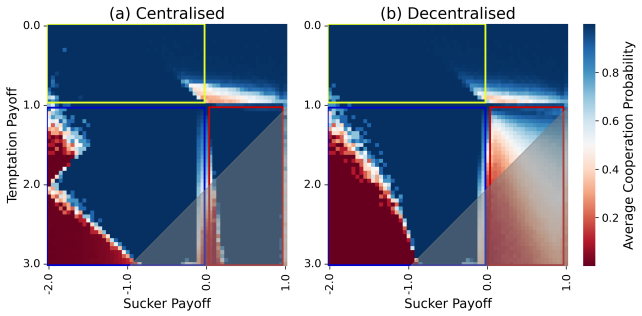

Figure 4: Heat-maps with average cooperation probability across social dilemmas where τ is inferred exclusively from agent's observations in a pre-training phase of learning. (a) τ is inferred from the set of agents collective pre-training observations. (b) each agent in the population infers their own τ exclusively from their own individual set of pre-training observations.

6 ROBUSTNESS TO EXTREME INFORMATION LIMITATIONS

Our previous results demonstrate the effectiveness of our methodology in an idealised setting in which the underlying game's max–min payoff is known and can be used to inform the partner selection protocol through τ.

In Figure 4 we instead assume that this information is unavailable and that an appropriate value for τ must therefore be inferred. This is done through a pre-training phase within which the population learns a kind of baseline joint policy (similar to that of Fig 1a). The intuition here is that this joint policy will be sufficient to characterise the underlying dynamics of the given game, replacing our reliance on knowing the min-max payoff of the game beforehand. τ is inferred by measuring the average payoff wrought by a series of rounds, 100 rounds in our case, played with these pre-training policies. Finally, the pre-training policies are discarded and agents learn their final policies under our full partner choice methodology using the inferred τ informing the partner choice protocol. Figure 4(a) shows our results when agents' pre-training reward observations are centralised and used to determine a single, shared, τ. Figure 4(b) shows our results when each agent infers an,individual τ_i using only their own observations in the pre-training phase. This constitutes the most general setting where agents are 1) given no information about the dilemma beforehand and 2) where the only assumption made is that agents are able to communicate requests/rejections. We see that in both cases populations are able to learn to cooperate effectively in games with low defection bias with this becoming more difficult in the decentralised case as agents have a lower volume, and diversity, of experience to draw from in the pre-training phase. However this comes at the cost of affecting performance in non-dilemmatic settings as demonstrated by the noise introduced in games where $T < 1$ and $S > 0$ (which itself originates from the high value of τ inferred in this setting) demonstrating the difficulties that arise from such a general setting. Despite this, these results exemplify our methodology's main scalability and generalisability gains over the current literature, namely in it's independence from large, complex and abstract state/action histories and detailed communication dependencies.

7 CONCLUSIONS

Social dilemmas constitute a central research problem in cooperative AI. They model complex social interactions where individual agents are incentivized to act selfishly, despite the collective benefits of cooperation. While current approaches have made progress in addressing such dilemmas, they often rely on restrictive assumptions or fall short of the adaptive, self-regulating behaviours observed in human societies. In this paper, we proposed a method based on partner choice to address these limitations. Using only agents' own reward signals, our approach allows populations of agents to regulate the behaviour of their opponents via an intrinsic reward mechanism which avoids reliance on reproducing reciprocal strategies. Based on the psychological effects of social acceptance and rejection, this mechanism indirectly reinforces the emergence of cooperation by incentivising participation. We showed that populations promoting social cohesion are better equipped to navigate social dilemmas. Our findings also demonstrate the robustness of our method across a range of dilemmas. These results establish a benchmark for extending the partner selection framework to social dilemmas that are both temporally and spatially complex, as modelled by Markov games, where the theoretical underpinnings and the distinction between self-interested and prosocial behaviours present substantial challenges.

REFERENCES

[1] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7047–7054. https://doi.org/10.1609/aaai.v34i05.6190

[2] Jean-Baptiste André and Nicolas Baumard. 2011. The Evolution of Fairness in a Biological Market. *Evolution; international journal of organic evolution* 65, 5 (2011), 1447–1456. https://doi.org/10.1111/j.1558-5646.2011.01232.x arXiv:https://academic.oup.com/evolut/article-pdf/65/5/1447/47949990/evolut1447.pdf

[3] Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science* 211, 4489 (1981), 1390–1396.

[4] Pat Barclay. 2013. Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior* 34, 3 (2013), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

[5] Pat Barclay. 2016. Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology* 7 (2016), 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012 Evolutionary psychology.

[6] Rico Berner, Thilo Gross, Christian Kuehn, Jürgen Kurths, and Serhiy Yanchuk. 2023. Adaptive dynamical networks. *Physics Reports* 1031 (2023), 1–59. https://doi.org/10.1016/j.physrep.2023.08.001 Adaptive dynamical networks.

[7] C. Monica Capra, Jacob K. Goeree, Rosario Gomez, and Charles A. Holt. 1999. Anomalous Behavior in a Traveler's Dilemma? *The American Economic Review* 89, 3 (1999), 678–690. http://www.jstor.org/stable/117040

[8] Phillip J.K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. 2023. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 448–456.

[9] Robyn M Dawes. 1980. Social dilemmas. *Annual review of psychology* 31, 1 (1980), 169–193.

[10] C Nathan DeWall and Brad J Bushman. 2011. Social acceptance and rejection: The sweet and the bitter. *Current Directions in Psychological Science* 20, 4 (2011), 256–260.

[11] Ilan Eshel and L. L. Cavalli-Sforza. 1982. Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences* 79, 4 (1982), 1331–1335. https://doi.org/10.1073/pnas.79.4.1331 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.79.4.1331

[12] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. 2024. Learning to resolve social dilemmas: a survey. *Journal of Artificial Intelligence Research* 79

(2024), 895–969.
[13] Ernst Fehr and Klaus M. Schmidt. 1999. A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114, 3 (1999), 817–868. arXiv:2586885 http://www.jstor.org/stable/2586885
[14] Jacob K. Goeree and Charles A. Holt. 2001. Ten Little Treasures of Game Theory and Ten Intuitive Contradictions. *The American Economic Review* 91, 5 (2001), 1402–1422. http://www.jstor.org/stable/2677931
[15] Donald W Griesinger and James W Livingston Jr. 1973. Toward a model of interpersonal motivation in experimental games. *Behavioral science* 18, 3 (1973), 173–188.
[16] Christoph Hauert and Michael Doebeli. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. *Nature* 428, 6983 (2004), 643–646.
[17] Siyu He, Qin Li, Minyu Feng, and Attila Szolnoki. 2026. Reputation assimilation mechanism for sustaining cooperation. *Chaos, Solitons and Fractals* 202 (2026), 117586. https://doi.org/10.1016/j.chaos.2025.117586
[18] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review* 91, 2 (2001), 73–78. http://www.jstor.org/stable/2677736
[19] Edward Hughes, Joel Z. Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3330–3340.
[20] Peter Kollock. 1998. Social Dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology* 24, 1 (1998), 183–214. https://doi.org/10.1146/annurev.soc.24.1.183
[21] Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems* (São Paulo, Brazil) *(AAMAS '17)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 464–473.
[22] Chin-wing Leung and Paolo Turrini. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.
[23] Erez Lieberman, Christoph Hauert, and Martin A Nowak. 2005. Evolutionary dynamics on graphs. *Nature* 433, 7023 (2005), 312–316.
[24] Wim BG Liebrand. 1984. The effect of social motives, communication and group size on behaviour in an N-person multi-stage mixed-motive game. *European journal of social psychology* 14, 3 (1984), 239–264.
[25] Michael W. Macy and Andreas Flache. 2002. Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences of the United States of America* 99, 10 (2002), 7229–7236. http://www.jstor.org/stable/3057846
[26] Udari Madhushani, Kevin R McKee, John P Agapiou, Joel Z Leibo, Richard Everett, Thomas Anthony, Edward Hughes, Karl Tuyls, and Edgar A Duéñez-Guzmán. 2023. Heterogeneous Social Value Orientation Leads to Meaningful Diversity in Sequential Social Dilemmas. *arXiv preprint arXiv:2305.00768* 0 (2023), 9.
[27] Kevin R. McKee, Ian Gemp, Brian McWilliams, Edgar A. Duèñez Guzmán, Edward Hughes, and Joel Z. Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (Auckland, New Zealand) *(AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 869–877.
[28] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-Level Control through Deep Reinforcement Learning. *Nature* 518, 7540 (2015), 529–533. https://doi.org/10.1038/nature14236
[29] Martin Nowak and Karl Sigmund. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364, 6432 (1993), 56–58.
[30] Martin A Nowak and Karl Sigmund. 1998. Evolution of indirect reciprocity by image scoring. *Nature* 393, 6685 (1998), 573–577.
[31] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 7092 (2006), 502–505.
[32] Jorge M Pacheco, Arne Traulsen, and Martin A Nowak. 2006. Coevolution of strategy and structure in complex networks with dynamical linking. *Physical review letters* 97, 25 (2006), 258103.
[33] David G Rand and Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences* 17, 8 (2013), 413–425.
[34] Amnon Rapoport and Abbe Mowshowitz. 1966. Experimental studies of stochastic models for the Prisoner's dilemma. *Behavioral Science* 11, 6 (1966), 444–458. https://doi.org/10.1002/bs.3830110604 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830110604
[35] Tianyu Ren and Xiao-Jun Zeng. 2024. Enhancing cooperation through selective interaction and long-term experiences in multi-agent reinforcement learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. IJCAI, California, Article 22, 9 pages. https://doi.org/10.24963/ijcai.2024/22
[36] Karl Sigmund. 2010. *The calculus of selfishness*. Princeton University Press, Princeton, New Jersey.
[37] Brian Skyrms. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge University Press, Cambridge, United Kingdom.
[38] Brian Skyrms and Robin Pemantle. 2000. A dynamic model of social network formation. *Proceedings of the National Academy of Sciences* 97, 16 (2000), 9340–9346. https://doi.org/10.1073/pnas.97.16.9340 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.97.16.9340
[39] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.
[40] Robert L Trivers. 1971. The evolution of reciprocal altruism. *The Quarterly review of biology* 46, 1 (1971), 35–57.
[41] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. 2023. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence* 2, 2 (2023), 26339137231162025.
[42] Juan Wang and Chengyi Xia. 2023. Reputation evaluation and its impact on the human cooperation—A recent survey. *Europhysics Letters* 141, 2 (jan 2023), 21001. https://doi.org/10.1209/0295-5075/aca997
[43] Jane X. Wang, Edward Hughes, Chrisantha Fernando, Wojciech M. Czarnecki, Edgar A. Duéñez Guzmán, and Joel Z. Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 683–692.

A ACCEPTANCE PROBABILITY

Our partner choice methodology utilises a *request/reject* procedure. If agent i's opponent memory $m_t^i(j)$ for opponent j is above some threshold τ, then agent i sends a partnership request to agent j (denoted $i \rightarrow j$). In the case that $m_t^j(i) > \tau$, agent j automatically accepts this request and agents i and j form a partnership (denoted $[i, j]$). However, if $m_t^j(i) \leq \tau$, then the probability, $p([i, j]|i \rightarrow j)$ of agent j accepting agent i's request, i.e., of forming a partnership with agent i given i's request, decays exponentially. In our main experiments (seen in Figure 1), $p([i, j]|i \rightarrow j) = S(m_t^j(i) - \tau)$, where $S(\cdot)$ is the standard sigmoid function. Here we explore the affect of using a different probability function defined as

$$p([i, j]|i \rightarrow j) \coloneqq \min\left(\exp\left[-k\frac{m_t^j(i) - \tau}{\gamma - \tau}\right], 1\right), \tag{3}$$

where the function $\epsilon(x) \coloneqq \exp\left(-k\frac{x-\tau}{\gamma-\tau}\right)$ in (3) is defined with three parameters: k, τ and γ. τ is defined, as in the main paper, as the min-max probability of the given game, γ is defined as $\min(\{\mathcal{R}_i\}_{i\in\mathcal{N}})$ or, the minimum payoff atainable in the given game. k is a hyperparameter. The intuition behind this function is that $\epsilon(x)$ exponentially decays from 1 to $\exp(-k)$ as x decreases from τ to γ (assuming $\gamma < \tau$). This results in an exponentially decaying acceptance probability as $x = m^j(i)$ approaches γ. Figure 7 illustrates our exploration of this function across varying values of hyperparameter k.

The intuitive effect of increasing k is to lower the likelihood for agents to play against those they deem to be low reward yielding opponents. In tern, this makes agents less 'forgiving' against potential defectors. When $k = 0$, $p([i, j]|i \rightarrow j) = 1$. Intuitively, this means the population is completely non-discriminatory and agents are partnered with all others in the population throughout training. Conversely, when $k \rightarrow \infty$, the partnership probability resembles a step function with

$$p([i, j]|i \rightarrow j) = \begin{cases} 1 & \text{if } m_t^j(i) > \tau, \\ 0 & \text{otherwise.} \end{cases}$$

Here, agents are completely discriminatory—any agent i will permanently refuse play against any opponent who demonstrates play in such a way that their respective opponent memory falls below τ.

This is reflected in our empirical findings. When k is too low, cooperation is negatively effected as agents become less effective at evading exploitation. Conversely, when k is too high, agents are too discriminatory, leading to fewer games being played and, hence, less learning which results in noisier, less "defined" joint policies.

B ADDITIONAL EXPERIMENTS

Next, we present additional experiments that test the robustness of our results across different values of (hyper)parameter values.

Deriving λ from the Environment. For situations where treating the intrinsic reward λ as a hyperparameter is undesirable, we investigate the feasibility of tying the intrinsic reward directly to m_t^i. Intuitively, the affect of this is to give agents a sense of *regret*: if a high-reward yielding opponent suddenly becomes rejective, then the agent will perceive this as the action leading to that rejection having caused a loss of reward equal to the reward the agent expects from that opponent. More concretely, if a *rejection* from agent j would lead to agent i experiencing intrinsic reward λ then, in this new setting $\lambda = -m_t^i(j)$. Conversely if a partnership request *acceptance* from agent j resulted in i experiencing λ, the value would instead be $\lambda = m_t^i(j)$.

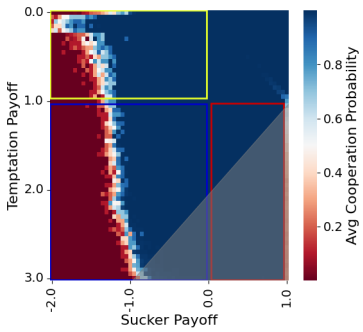

Figure 5: Using agents' opponent memory values towards accepted/rejected opponents as the value for λ.

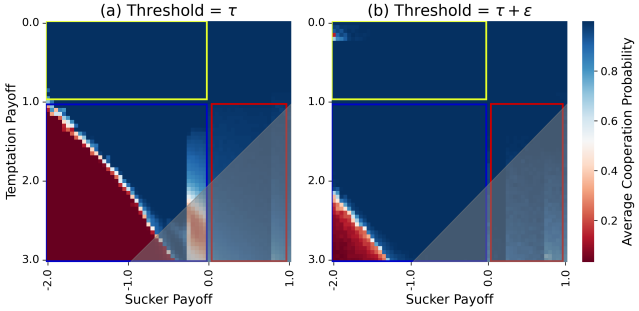

Figure 6: Heatmaps with average cooperation probability for different population sizes. The lack of substantial differences between the maps suggests robustness of our method to size of population.

Figure 5 illustrates the effect of tying λ to $m_t^i(j)$ in this way. For $S > -1$, this is a universally effective method for producing cooperative populations.

Population size. Figure 6 demonstrates our methodology under a population of 1000 agents and illustrates the general invariance of our method with respect to population size.

Alternative values for λ. Figure 8 shows how cooperation can be further improved by increasing the intrinsic rewards value. When $\lambda = 1.5$, cooperation becomes universal: agents in all tested Stag Hunts and Prisoner's Dilemmas converge to fully cooperative populations. This reflects a natural insight: the stronger the incentive to defect, the stronger the corrective force required to induce cooperation. In this light, λ can be, interestingly, interpreted as the population's weight on social participation: cooperative outcomes are more likely to emerge and persist in populations which place a greater emphasis on social inclusion (intrinsic rewards) than material outcomes (extrinsic rewards), even in difficult social dilemma-like scenarios.

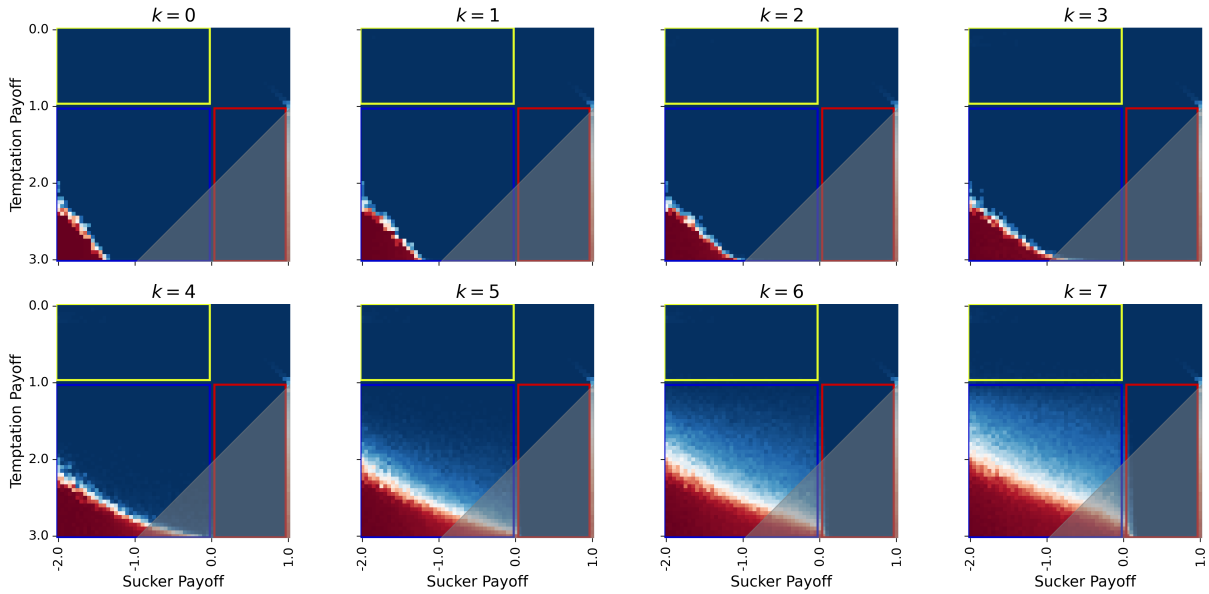

Figure 7: Heatmaps with average cooperation probability for different values of k under our alternative probabilty function.

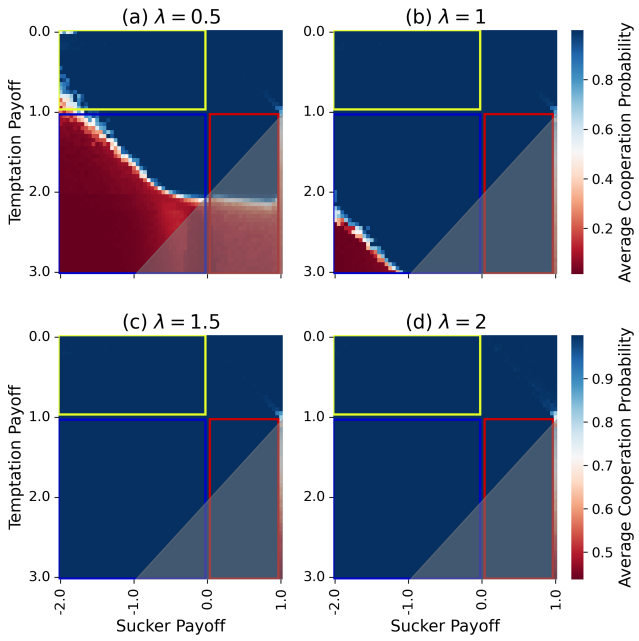

Figure 8: Heatmaps with average cooperation probability across social dilemmas with varying intrinsic reward (λ) values.

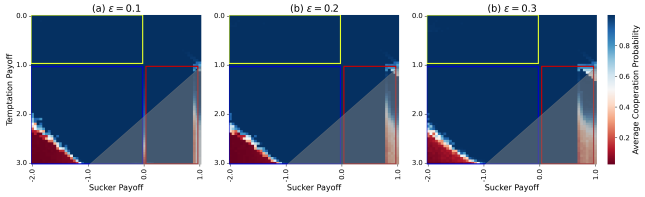

Figure 9: Heatmaps with average cooperation probability across social dilemmas with varying ϵ values.

Testing different epsilon values. Figure 9 illustrates the robustness of our methodology with different values of ϵ. The main result here is the limited affect on cooperation when differing the value of epsilon, reinforcing that the exact value of epsilon is not critical.

Opponent memory discount factor.α_m. Figure 10 shows our experiments empirically investigating the best value for α_m. The value we settled upon was $\alpha_m = 0.9$. There is a clear correlation between the opponent memory discount factor and populations' willingness to cooperate–agents become only more willing to cooperate as the value increases. This makes sense as α_m essentially dictates how reactive agents are to the most recent game outcomes.

Policy learning rate α_θ. Figure 11 shows our experiments empirically investigating the best value for α_θ. While differences are minimal, we observe a clear increase in noise as the value of α_θ is increased. We choose $\alpha_\theta = 0.1$ as this is the least noisy value and ensures stable learning.

Defection bias. Figure 12 illustrates our investigation into how agent populations learn, under our full methodology with $\lambda = 1$, over time under different levels of defection bias. Rows of subplots are labelled as (a), (b) or (c) respectively where, (a) represents plots of data from the population as a whole, and (b) and (c) represent data from the perspective of singular agents. In each subplot, the line colour represents the defection bias of the social dilemma being examined, with blue and red plots representing learning dynamics within low (easy) and high (hard) defection bias games respectively. Purple plots are reserved for 'medium' games - where populations are more likely to be split into groups of cooperators and defectors.

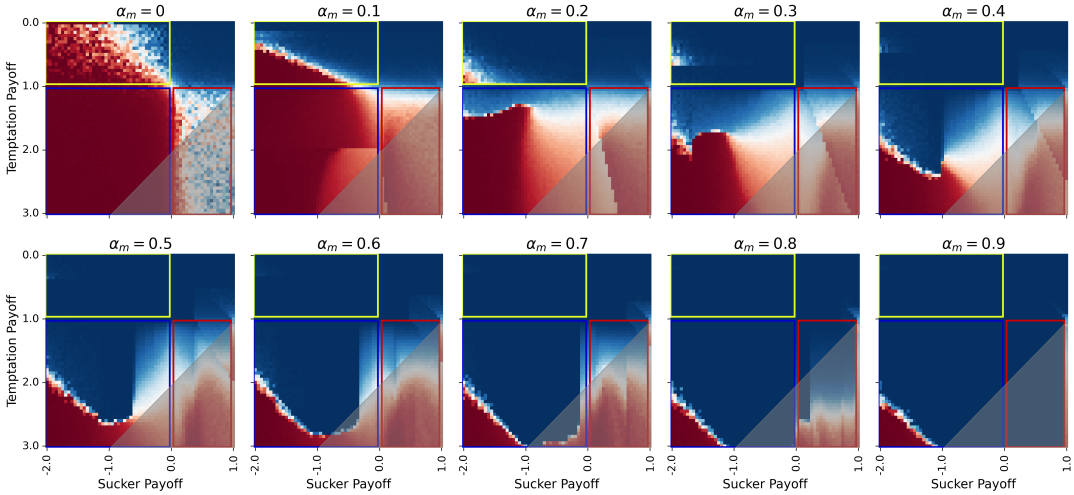

Figure 10: The effect of different values for opponent memory discount factor α_m across all dilemmas.

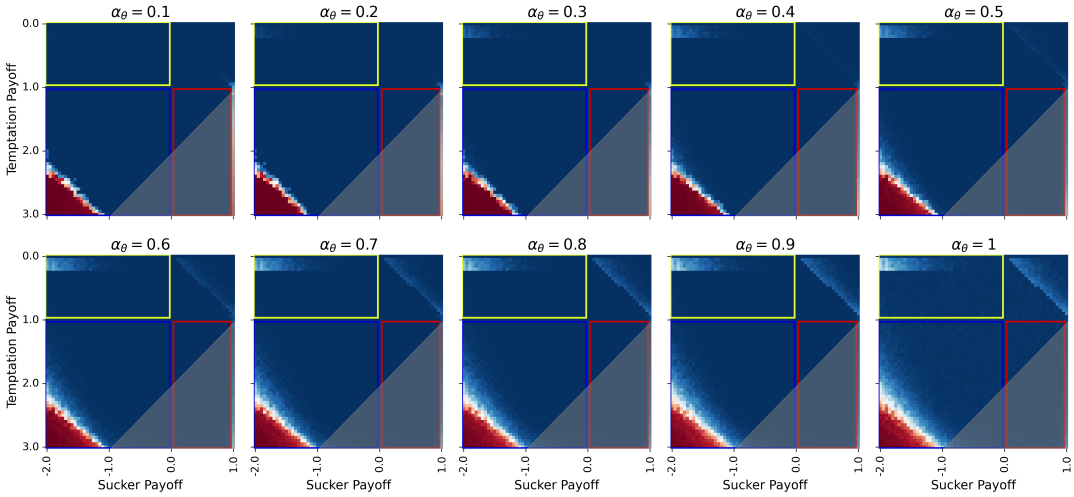

Figure 11: The effect of different values for policy learning rate α_m across all dilemmas.

Row (a) shows subplots drawn from data collected on an overall population level. Subplot (a)-i illustrates the number cooperators in the population over time. It is clear that both the low and high defection bias games lead the population to quickly converge, homogenously, to either cooperate or defect respectively whereas in the medium game, the population converges heterogenously with a roughly 50/50 split between policy types. This is reflected in both subplots (a)-ii and (a)-iii which show the normalised social welfare (average extrinsic rewards per game, per agent), and population average cooperation probability respectively.

Row (b) shows subplots from the same experiments as in row (a), however, with a focus on a single agent. Again, we see in subplot (b)-iii, showing agent 0's cooperation probability over time, that in the low and high defection bias games, agent 0 tends towards cooperation and defection, respectively, as suggested by the plots in row (a). This trend is repeated in subplot (c)-iii, for agent 1. It is important however to note the key difference between plots (b)-iii and (c)-iii: the cooperation probability of agents 0 and 1 in the medium difficulty game. Here we can see that agent 0 converges to defect whereas agent 1 converges towards cooperation. Given this distinction, a key insight is that, under partner choice, groups of cooperators yield higher rewards than groups of defectors. This can be seen in subplots (b)-ii and (c)-ii and is reinforced by reward observations in Figure 3. After approximately 350 timesteps, we see that once agent 1 is established as a reliable cooperator, their overall extrinsic reward (i.e., sum of rewards yielded purely from games played) increases to approx 4 reward per timestep. This is in contrast to agent 0 who sits between 0 and 2 reward even after convergence to a strong defect policy.

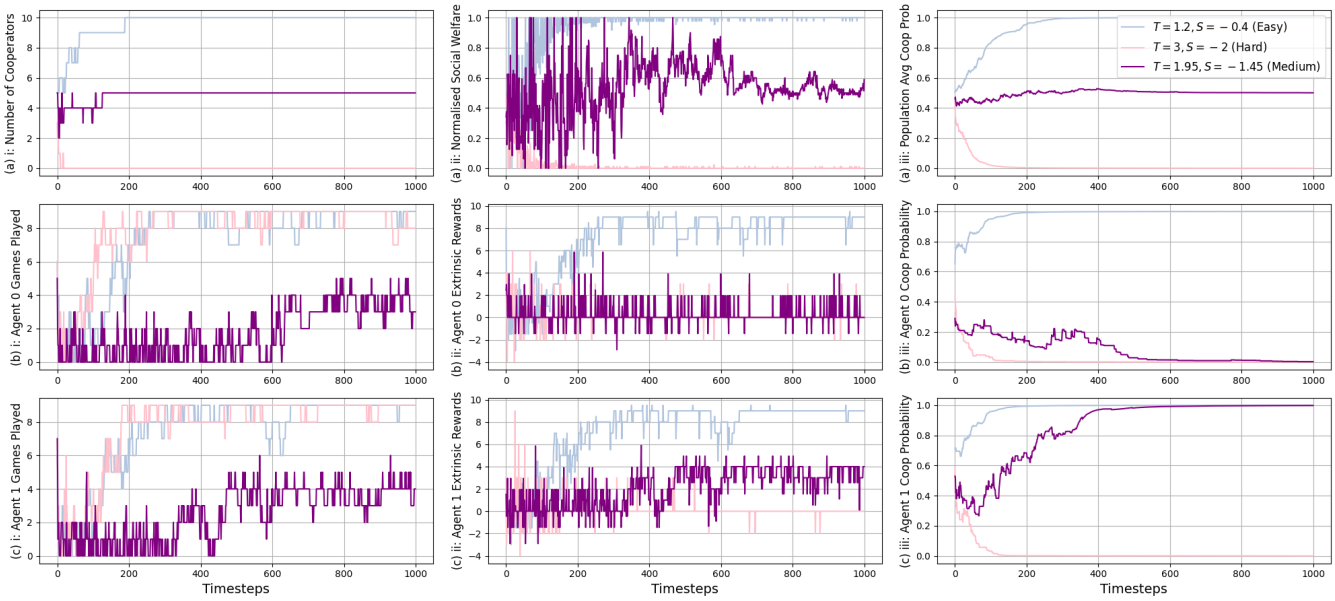

Figure 12: Plots showing comparisons of various metrics between easy, medium and hard social dilemmas. Red and blue lines represent populations trained under hard and easy social dilemmas, respectively, with purple lines representing the learning of a population in a medium difficulty social dilemma. Results from the medium difficulty game show that, when agents form cooperative/defecting cliques, members of the cooperative clique tend to have higher rewards than those who defect.

C COMPARISONS TO STATIC REWARD-SHAPING METHODS

Finally, we conduct experiments comparing the effects of prominent reward shaping formalisms to our partner choice methodology. In both settings, agents are matched randomly.

Social Value Orientation (SVO). Figure 13 illustrates the effects of a Social Value Orientation (SVO) style reward shaping, $\mathcal{R}_i^{SVO}(\theta)$ within our social dilemma test bed where rewards are shaped according to

$$\mathcal{R}_i^{SVO}(\theta) = \cos(\theta) r_i + \sin(\theta) r_j.$$

The idea is that the agent's SVO (given by parameter θ) embodies the degree to which they value their own utility against that of their opponent. Intuitively, an agent i with $\theta = 0$ would be considerate only towards their own rewards r_i whereas an agent with $\theta = 90$ would only be considerate towards their opponent j's rewards r_j. Given the relative degrees to which populations cooperate or defect under SVO, we estimate that our methodology is roughly equivalent to a population with a $\theta = 25$.

Inequity Aversion (IA).. Figure 14 illustrates the effects of an Inequity Aversion (IA) style reward shaping $\mathcal{R}_i^{IA}(\alpha, \beta)$ where rewards are shaped according to

$$\mathcal{R}_i^{IA}(\alpha, \beta) = r_i - \alpha_i f(r_i, r_j) - \beta_i g(r_i, r_j),$$

where, $f(r_i, r_j) = \max(r_j - r_i, 0)$ is called *disadvantageous inequity* and $g(r_i, r_j) = \max(r_i - r_j, 0)$ is called *advantageous inequity*. Here, α and β are hyperparameters which determine the agent's level of aversion to disadvantageous and advantageous inequity respectively. Intuitively, under inequity aversion, agents are averse to inequitable reward outcomes. In other words, agents suffer a reward penalty when rewards between themselves and their opponents are dissimilar. This is implemented through f and g as follows:

(1) f, the disadvantageous inequity, penalises i's reward when $r_j > r_i$. A higher α, increases the scale of this penalty.
(2) g, the advantageous inequity, penalises i's reward when $r_i > r_j$. A higher β, increases the magnitude of this penalty.

Given the erratic nature of the effect of IA in our setting, we do not identify a parameter configuration which is most similar to our partner choice methodology.

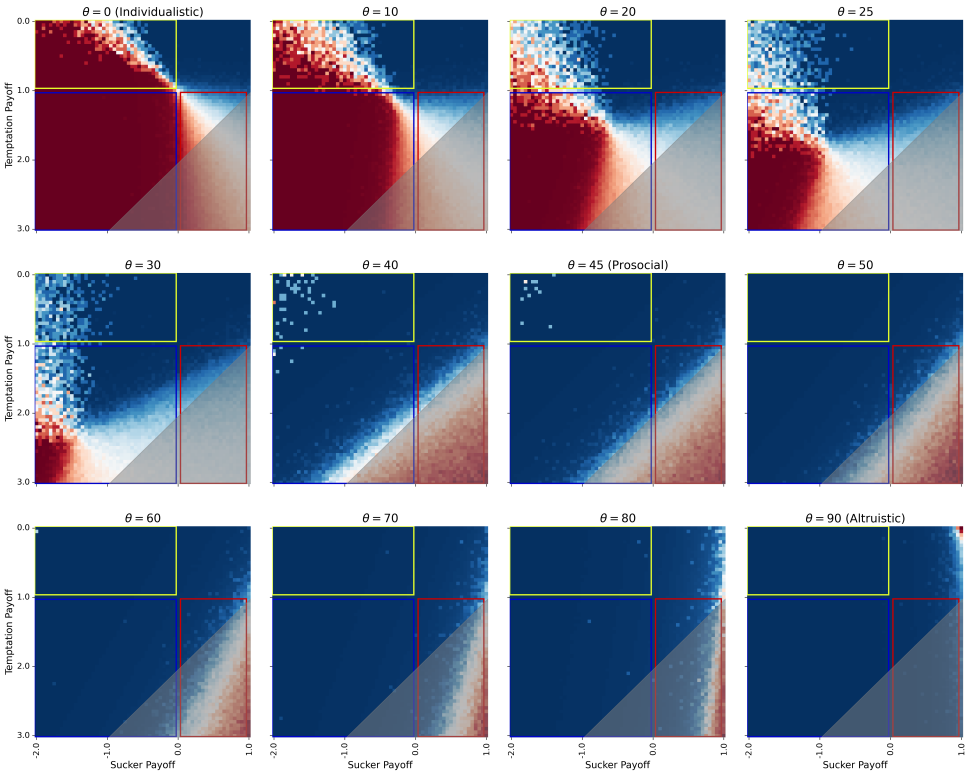

Figure 13: The effect of a Social Value Orientation reward shaping across all dilemmas.

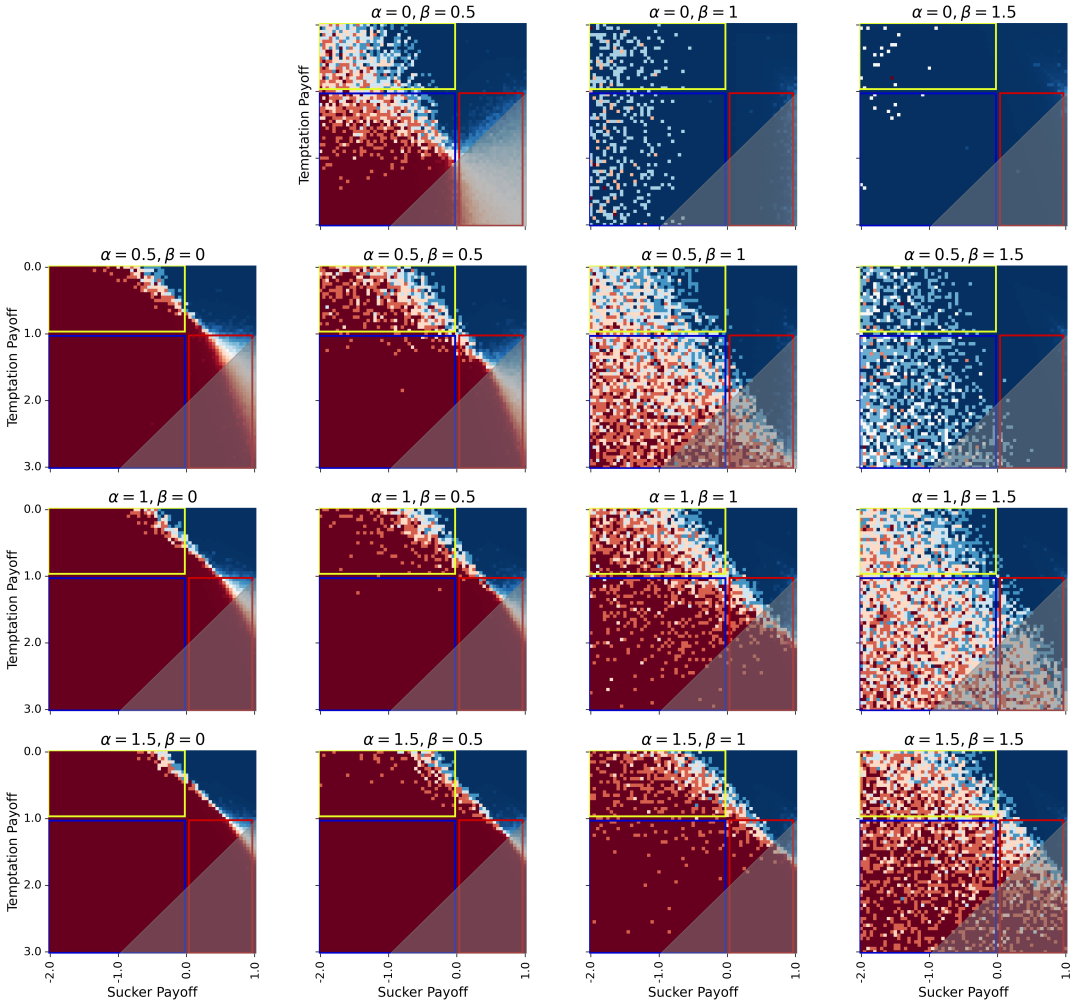

Figure 14: The effect of an Inequity Aversion reward shaping across all dilemmas.