Bottom-Up Reputation Promotes Cooperation with Multi-Agent Reinforcement Learning

Tianyu Ren
University of Manchester
Manchester, United Kingdom
tianyu.ren@manchester.ac.uk

Xuan Yao
Southeast University
Nanjing, China
shiny_yao@yeah.net

Yang Li
University of Manchester
Manchester, United Kingdom
yang.li-4@manchester.ac.uk

Xiao-Jun Zeng
University of Manchester
Manchester, United Kingdom
x.zeng@manchester.ac.uk

ABSTRACT

Reputation serves as a powerful mechanism for promoting cooperation in multi-agent systems, as agents are more inclined to cooperate with those of good social standing. While existing multi-agent reinforcement learning methods typically rely on predefined social norms to assign reputations, the question of how a population reaches a consensus on judgement when agents hold private, independent views remains unresolved. In this paper, we propose a novel bottom-up reputation learning method, *L*earning with *R*eputation *R*eward (*LR2*), designed to promote cooperative behaviour through rewards shaping based on assigned reputation. Our agent architecture includes a dilemma policy that determines cooperation by considering the impact on neighbours, and an evaluation policy that assigns reputations to affect the actions of neighbours while optimizing self-objectives. It operates using local observations and interaction-based rewards, without relying on centralized modules or predefined norms. Our findings demonstrate the effectiveness and adaptability of LR2 across various spatial social dilemma scenarios. Interestingly, we find that LR2 stabilizes and enhances cooperation not only with reward reshaping from bottom-up reputation but also by fostering strategy clustering in structured populations, thereby creating environments conducive to sustained cooperation.

KEYWORDS

Bottom-Up Reputation, Social Norm, Cooperative Intelligence, Multi-Agent Reinforcement Learning, Reward Shaping

1 INTRODUCTION

Cooperation is ubiquitous in almost every form of social interaction and is arguably a key factor in the success of complex social systems,

such as economics [59], evolutionary biology [42], and multi-agent systems (MAS) [4]. With the rise of artificial intelligence (AI), autonomously operating learning agents within MAS are becoming increasingly common. A typical example is autonomous vehicles, which must share the road with both other vehicles and human drivers; without proper coordination, issues such as road congestion may arise [17]. However, achieving effective collective actions among self-interested agents remains a significant challenge.

To address this challenge, mechanisms that facilitate correlated interactions among autonomous, decentralized learning agents are required [6]. These mechanisms shape the interaction structures or rewards within a population to promote cooperation over defection. In evolutionary biology and economics, indirect reciprocity (IR) and direct reciprocity (DR) have been identified as key drivers of cooperation [26]. Notably, IR—offering a compelling explanation for the evolution of cooperative behaviour among MAS agents [27]—differs from DR, where benefits are received directly from those assisted [51]; instead, IR allows individuals to benefit from the broader community. Consequently, sustaining cooperation via IR necessitates social information reflecting the historical behaviours of agents, which is often governed by social norms. These norms establish expected patterns and may impose penalties for violations. Enforcement typically involves rumour [12] or reputation [28], which assess the "goodness" of agents and enable selective altruism, where individuals help those with good reputations.

Studies in MAS have demonstrated the importance of IR by incorporating it into traditional multi-agent reinforcement learning (MARL) tasks [52, 56]. Reputation allows agents to reshape rewards by assigning bonuses or penalties based on observed social information. For instance, agents can integrate moral values [49] and normative punishment [52] to enforce compliance and learn behavioural rules. Reputation also serves as an informative signal for optimizing behaviour [1] or forming selective interaction relationships [23, 33] across different scenarios. This body of research not only promotes cooperation and explains the formalization of artificial ethics and morality among MAS but also provides a feasible pathway for constructing a normative human-AI society.

Despite promising developments, reputation-based strategies and norms in MARL face several challenges. Standard IR theory assumes that reputations are common knowledge, meaning the entire population agrees on the evaluation process [36]. This consensus stabilizes reputation dynamics and promotes cooperation

by setting clear expectations. However, when agents form private assessments, disagreements can arise, leading to perceptions of unfairness that may undermine cooperation [12]. Constraining the norm space through structured interactions appears essential to prevent cooperative collapse [25]. Moreover, while social norms are often enforced exogenously by central institutions, they vary widely across different environments and populations [13]. Establishing appropriate norms thus demands extensive prior knowledge of the environment—a formidable challenge in complex MARL settings. These issues reduce the adaptability of reputation mechanisms and raise two fundamental questions: To what extent do individuals share a consensus on reputations, and how are social norms used to evaluate and assign reputations?

In this paper, we address these problems by developing Learning with Reputation Reward (LR2), a training method that promotes cooperation through reshaped rewards based on bottom-up reputation. Each agent has two policies: a dilemma policy for determining action and an evaluation policy for assigning reputations. LR2 agents aim both to maximise personal benefits without compromising their reputation and to assign reputations that incentivise prosocial behaviour within their local group while furthering their own interests. It reshapes the static reward function in a distributed manner by integrating assigned reputations from neighbours. Unlike similar reputation-based MARL methods, our approach emphasises the heterogeneity and endogeneity of norm formation, framing reputation assignment as a learning process rather than a reactive adaptation to predefined social norms [29, 43, 56]. This approach captures the coevolution of reputation and cooperation, offering a pathway to predict whether, and by what dynamics, a population will achieve coordination under IR. We validate LR2 in various social dilemmas, showing that it not only influences opponents' behavioural strategies through reputation assignment but also promotes strategy clustering in structured populations. Furthermore, we demonstrate that biased social information—where peers assign differing reputations to the same individual—can either facilitate or hinder cooperation, depending on the dilemma's strength. In summary, our work makes three contributions:

- We propose a novel training method LR2, which promotes cooperation through bottom-up reputation. It enables agents to consider their impact on others while optimizing their objectives with learned reputation assignments.
- We reveal how LR2 promotes cooperation by examining the coevolution of cooperation and reputation. LR2 stabilizes prosocial behaviour through emerging normative judgments and further enhances cooperation via strategy clustering.
- We demonstrate that LR2 overcomes the scalability limitations of predefined norms in various social dilemma scenarios. It enables agents to learn and converge on shared norms in a decentralized and sample-efficient manner.

2 RELATED WORKS

Cooperation in MARL. MARL in collaborative settings can be dichotomized into two branches: team-based and mixed-motive environments [5]. In the former, agents coordinate actions and share a single scalar reward. Several recent advanced methods have been

proposed to solve this problem [44, 45, 58]. By contrast, in mixed-motive environments, agents receive individual rewards, necessitating a balance between personal maximisation and social welfare. Such systems often encounter social dilemmas [50], where individually rational decisions lead to collectively suboptimal outcomes. In MAS, these dilemmas extend spatially and temporally, forming sequential social dilemmas [16]. A core challenge is understanding how cooperation among self-interested agents can emerge and remain stable, despite threats like conflict, overconsumption, free-riding, and defection [5]. Previous research has explored solutions through other-regarding preferences [11, 22], reputation mechanisms [1, 43], and anticipation of future behaviours [7, 57]. Notably, reputation serves as an adaptable measure of social standing that can guide subsequent agent interactions [2]. Building on this, our paper focuses on encouraging cooperation in MARL through the auxiliary learning dynamics inherent in reputation assignments.

Reputation and Social Norms. In the IR mechanism, actions are evaluated against social norms that integrate behavioural and reputational information to update an agent's standing [27]. These norms can be imposed top-down by a central authority or emerge bottom-up through interactions [39]. In the top-down approach, norms are designed offline and uniformly applied [37]. For instance, using Boolean inputs for reputation and strategy, norms can be encoded as a 4-bit string, $d = (d_{G,C}, d_{G,D}, d_{B,C}, d_{B,D})_2$. Then, agents update reputations based on second-party reports with imperfect observations [9]. Since reputations are repeatedly evaluated with small errors, this process can be modelled as an ergodic Markov chain with a characterised stationary distribution [28], determining the fitness of an agent's rule π_i. Previous studies indicate that the well-known norm Stern Judging (SJ) effectively drives behavioural dynamics through both imitation processes [37] and Reinforcement Learning (RL) [43]. SJ assigns good reputations to agents who cooperate with good partners and defect against bad ones, while assigning bad reputations to those who act the opposite.

MARL and Bottom-Up Norm. In MARL, the centralized training with decentralized execution (CTDE) framework [20] lacks centralized information for top-down judgment during execution. A bottom-up approach, however, enables agents to adopt decentralized methods for assigning reputations, allowing norms to emerge organically [56]. Recent studies explore how agents learn to assign and respond to reputations. These norms can be learned either through explicit payoff-based methods [56] or by inferring the existence of shared normativity via approximate Bayesian rule induction [29]. Beyond reputation, the bottom-up creation of norms can also manifest as intrinsic rewards [22] or public sanctions [52]. However, maintaining cooperation under learned norm-guided behaviour is constrained by the presence of a predefined set of norms. In this paper, we extend this approach by allowing agents to assign reputations to their interacting neighbours based solely on their own rewards, which does not require knowledge of a predefined set of norms beyond the agents' observations.

3 PRELIMINARIES

3.1 Social Dilemmas

We examine social dilemmas using a spatial model of repeated two-player symmetric games, where agents interact with neighbours

by uniformly choosing to cooperate (C) or defect (D). The game is defined by the payoff matrix M:

$$M = \begin{array}{c} \\ C \\ D \end{array} \begin{pmatrix} R,R & S,T \\ T,S & P,P \end{pmatrix}. \tag{1}$$

Here, mutual cooperation yields reward R and mutual defection yields punishment P. In asymmetric cases, where one agent cooperates and the other defects, the cooperator receives the sucker's payoff S and the defector obtains the temptation payoff T. These payoffs define the social dilemma [21]:

(1) Snowdrift Game (SG): $T > R > S > P$. The optimal strategy depends on the opponent's action, making unilateral cooperation advantageous.
(2) Stag-Hunt Game (SH): $R > T > P > S$. Unilateral cooperation leads to a loss if the other player defects, emphasizing the importance of mutual cooperation.
(3) Prisoner's Game (PD): $T > R > P > S$. Mutual defection is the equilibrium outcome, even though mutual cooperation offers a better payoff for both players.

The PD, which blends aspects of SG and SH, poses the greatest challenge for fostering cooperation. Following the convention [35] of normalising the difference between R and P to 1, we set $R = 1$ and $P = 0$, with T and S constrained to $0 \le T \le 2$ and $-1 \le S \le 1$, respectively.

3.2 Reputation and Social Norms

Reputation is public information shared among neighbouring agents, derived from their strategies to measure cooperativeness. Unlike previous studies using binary reputation values [1, 31], we model reputation as a continuous attribute ranging from 0 to 1, indicating an agent's degree of "Good" (G) or "Bad" (B). This extension enhances the robustness of agent learning by capturing finer variations in social standing regarding agent behaviour. Initially, the G and B labels are nominal with no inherent meaning; their significance emerges through agents' actions during the evolution of the game. An agent's decision to cooperate depends on its current reputation and the observed reputations of its neighbours. In pairwise donor games with binary reputations, common strategies include always cooperate (ALLC), always defect (ALLD), and discriminate (DISC), where DISC agents cooperate with good-reputation recipients and defect against bad ones [37]. However, since reputation in our model is continuous and agents interact with groups of neighbours rather than single recipients, the strategy representations used in previous studies [38] must be expanded. Specifically, the strategy tuple Π becomes an infinitely large bit string $\Pi = (\pi_1, \pi_2, \dots)$, where each $\pi \in \{0, 1\}$ represents whether an agent cooperates for a given combination of self and neighbour reputations. We employ a policy network to specify the behaviour rules for each agent, as detailed in Section 4.3.

3.3 Markov Games

We formalize the evaluation process in social dilemmas as an N-agent, partially observed, general-sum Markov games (POMGs) [18, 19, 30], augmented with the concept of bottom-up reputation and an

evaluation function that assesses the behaviours of neighbourhoods. A POMG is given by the tuple $M = \langle S, \mathcal{A}, O, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{A} := \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ and $O := O^1 \times \cdots \times O^N$ denotes the joint action observations and the joint action space of all N agents, indexed by $i \in [N] := \{1, \dots, N\}$. The game operates on a finite set of states S, with each agent's d-dimensional observation of the state mapped by the function $\mathbb{O} : S \times \{1, \dots, N\} \to \mathbb{R}^d$. In each state, the agent i selects an action from its action set A_i, and the state transitions according to the stochastic function $\mathcal{T} : S \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \to \triangle(S)$ represents the set of probability distributions over S. Each agent then receives a reward based on the reward function $\mathcal{R} : S \times \mathcal{A} \to \mathbb{R}$. The objective of agent i is to learn a policy $\pi^i : O^i \to \triangle(A^i)$ based on its own observation $o^i = \mathbb{O}(s, i)$ and current reward $r^i(s, a^1, \dots, a^N)$ where $s \in S$ represents the current environment state. The goal is to maximize a long-term γ-discounted payoff under the joint policy $\vec{\pi} = (\pi^1, \dots, \pi^N)$ from an initial state s_0:

$$V_{\vec{\pi}}^i(s_0) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r^i(s_t, \vec{a}_t) | \vec{a}_t \sim \vec{\pi}_t, s_{t+1} \sim \mathcal{T}(s_a, \vec{a}_t)\right], \tag{2}$$

where $\gamma \in [0, 1]$ is the discount factor, and T denotes the time horizon. We extend POMGS by incorporating reputation assessments, formalized as the function $p_{\eta_t^i} : \tau_t^i \to \{0, 1\}^{|\Omega^i|}$ where Ω^i represents the neighbour set for agent i and $\tau_t^i := \bigcup_{j \in \Omega^i} ((S \times \mathcal{A}^j)_t \times S)$ captures state-action trajectories of i with its neighbours at timestep t. Accordingly, $\tau_t := \bigcup_{i \in N} \tau_t^i$ aggregates trajectories across all agents. In other words, the reputation assessment p^i evaluates the goodness of each neighbour, assigning 1 for "good" behaviour and 0 for "bad" behaviour. We refer to this framework as *reputation-augmented Markov games*, which provides a basis for reputation assignment and coordination in sequential decision-making. The learning processes of the reputation evaluation and the reputation-augmented reward function will be detailed in the following section.

4 METHODOLOGY

Here, we introduce the Learning with Reputation Reward (LR2) method, which allows agents to learn an evaluation function by explicitly accounting for their impact on neighbours and their own objectives. As shown in Figure 1a, each agent decides whether to cooperate or defect with its neighbours and then assigns assessments based on local interactions. A detailed LR2 framework is provided below, with the algorithm listed in Appendix A.1.

4.1 Learning with Reputation Reward

The core idea of our LR2 approach is to enable agents to learn an evaluation policy that progressively reshapes the rewards of nearby agents by considering the consequences of interactions on their own objectives. This reshaping encourages agents to account for the effects of their behaviour on their neighbours in a more cooperative manner. In the implementation, agent i learns both a dilemma policy and an evaluation policy, parameterised by $\theta^i \in \mathbb{R}^N$ and $\eta^i \in \mathbb{R}^N$, respectively, in different sequences. Let $-i$ denote the set of all neighbours of agent i. As shown in Figure 1b, LR2 agents optimize their dilemma policies by considering their actions on
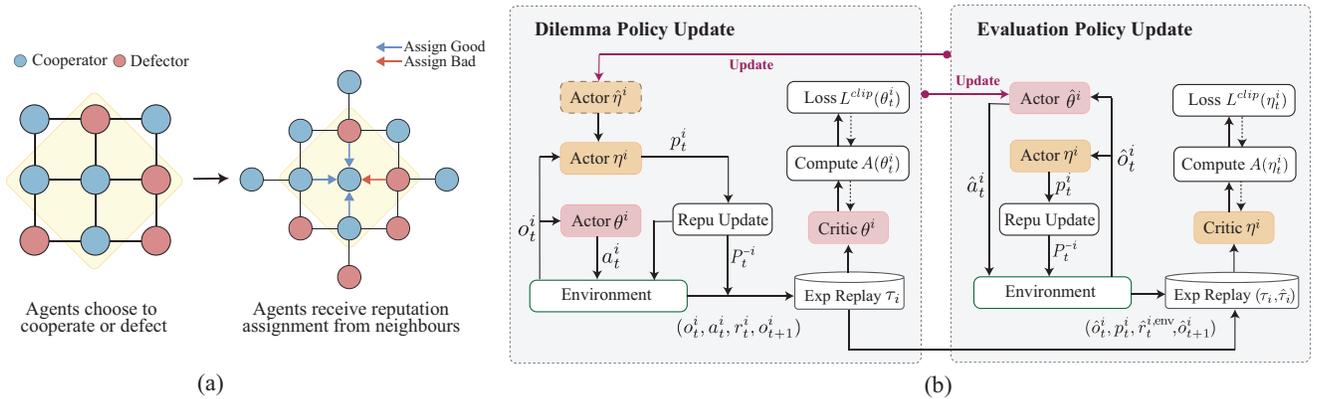
(a)

(b)

Figure 1: Overview of the social dilemma game with reputation and architecture of our LR2 agents. (a) Each agent is connected to four neighbours in a network. Each round consists of two phases: First, agents choose to cooperate or defect based on their reputations and those of their surrounding neighbours. In the second phase, agents receive reputation assignment reflecting how their behaviours are perceived within their neighbours' local group. (b) Agent i updates its dilemma policy θ^i by considering both environmental rewards and assigned reputation. The evaluation policy η^i is then updated based on the rewards accumulated by the updated dilemma policy $\hat{\eta}^i$.

neighbours, while adjusting evaluation policies to influence neighbour behaviour. Each agent occupies a specific spatial coordinate and interacts only within its von Neumann neighbourhood [47]. At each timestep t, agents engage in multiple pairwise social dilemma games involving strategy selection and reputation assignment.

In the first phase, agent i selects either cooperation or defection as its dilemma strategy for the next round, according to $a_t^i \sim \pi_{\theta^i}(\cdot|s_t)$. After participating in one round of game pairwise interactions with each neighbour, the agent i receives an environmental reward given by:

$$r_t^{i,\text{env}} = \sum_{j \in \Omega^i} a^{i\top} \mathcal{M} a^j, \tag{3}$$

We refer to this aggregate reward as the environmental reward that the agent receives from participating in social dilemma games. In the second phase, the neighbours of agent i assign assessments based on the evaluation function $p_{\eta^i} : O \times A^{-i} \to \mathbb{R}^{|\Omega^i|}$, which maps the agent's observation o^i and the actions of its neighbours to a vector of reputation assessments for all surrounding agents. Finally, the reputation of agent i at the timestep t can be updated according to a running average:

$$P_t^i = \alpha P_{t-1}^i + \frac{1-\alpha}{|\Omega^i|} \sum_{j \in \Omega^i} p_{\eta^i}^i(o_t^j, a_t^{-j}), \tag{4}$$

where α is a smoothing parameter that quantifies how quickly the agent's reputation is reshaped by nearby assessments and $p_{\eta^j}^{i,t}$ is the reputation assessment assigned to agent i by its neighbour j.

During training, LR2 agents update their dilemma policies by taking into account both environmental rewards and the reputations assigned by their neighbours. They then apply a separate policy gradient to the evaluation policy, based on the rewards generated by the updated dilemma actions (see Figure 1b). Following the implementation in [57], we structure the agents update procedure in an online cross-validation manner [46], to account for the fact that reputation assessments toward neighbours have a measurable effect

only after those neighbours have updated their dilemma policies. We next describe the LR2 method in a step-by-step process below.

4.2 Dilemma Policy Update

An LR2 learner optimizes its dilemma policy by considering the additional effect of reputation assessments received from neighbouring agents. Instead of optimizing the expected return under current dilemma parameters $V^i(\theta^i, \theta^{-i})$, an LR2 agent optimizes $V^i(\theta^i, \theta^{-i}, \eta^{-i})$, which accounts for the neighbours' evaluation policies η^{-i}. Specifically, it modifies the immediate reward of an agent so that, at each timestep t, the aggregate reward of agent i is divided into two components:

$$r^i(s_t, a_t, \eta^{-i}) := \beta r^{i,\text{env}}(s_t, a_t) + P_t^i(1-\beta)r^{i,\text{env}}(s_t, a_t), \tag{5}$$

where β is a weighting parameter that balances the influence of the environmental reward and the reputation-based reward. This formulation enables agents to adapt their strategies based not only on direct payoffs but also by accounting for how their actions affect their reputation among neighbours, effectively incorporating the influence of neighbouring agents' assessments into their own policy optimization. When $\beta = 1$, agents degenerate into selfish entities, learning independently and driven solely by their own interests.

Given this formulation, the reshaped reward r_t^i depends on the average reputation derived from the neighbour assessments vector $p_{\eta^{-i}}$. Accordingly, agent i learns a dilemma policy π^i, parameterised by θ^i, to maximize the objective:

$$\max_{\theta^i} J^{\text{dilemma}} := \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r^i(s_t, a_t, \eta^{-i}) + \omega \mathcal{H}^\pi(\cdot|s_t) \right]. \tag{6}$$

We incorporate an additional entropy bonus, controlled by the hyperparameter ω, to encourage exploration and mitigate the issue of early convergence [8]. The LR2 agent i updates its policy

parameter θ^i using a policy gradient approach:

$$\hat{\theta}^i \leftarrow \theta^i + \lambda \sum_{t=0}^{T} \left[\nabla_{\theta^i} \log \pi^i(a_t^i|o_t^i) G_t^i(\tau^i, \eta^{-i}) \right], \quad (7)$$

where λ is the decay learning rate, and G^i represents the discounted return starting from time t, which depends on the trajectory τ^i and the evaluation policies of the neighbours η^{-i}.

4.3 Evaluation Policy Update

With the updated dilemma policy $\pi_{\hat{\theta}}$, the system then generates a new trajectory $\hat{\tau} := (\hat{s}_0, \hat{a}_t, \hat{r}_0, \ldots, \hat{s}_T)$ based on the updated dilemma policy $\pi_{\hat{\theta}}$. To evaluate the quality of their neighbours, agents primarily compare the average performance of their neighbours within their local group. Hence, the evaluation reward r^{eval} is designed to assess how effectively agent i interacts with a specific neighbour j in comparison to the average interaction with all its neighbours:

$$r^{i,\text{eval}}(\hat{s}_t, \hat{a}_t) := \sum_{j \in \Omega^i} \left(\hat{r}^{i,j,\text{env}}(\hat{s}_t, \hat{a}_t) - \frac{\hat{r}^{i,\text{env}}(\hat{s}_t, \hat{a}_t)}{|\Omega^i|} \right). \quad (8)$$

This comparison allows agents to influence influencing their neighbours' behaviour so as to maximize their own extrinsic rewards by assigning reputation scores to neighbours. Agent i then uses this new trajectory $\hat{\tau}$ to optimize its evaluation policy, parameterised by η^i, with the following objective:

$$\max_{\eta^i} J^{\text{eval}} := \mathbb{E}_{\hat{\pi}} \left[\sum_{t=0}^{T} \gamma^t \left(r^{i,\text{eval}}(\hat{s}_t, \hat{a}_t) - \mu \mathcal{D}^i(o_t, a_t) \right) \right], \quad (9)$$

where $\mathcal{D}^i(o_t, a_t)$ represents the mean-squared-error, controlled by the sensitive parameter μ:

$$\mathcal{D}^i(o_t, a_t) := \sum_{j \in \Omega^i} \sum_{k \in \Omega^j} \left[p_{\eta^i}^j(o_t^i, a_t^{-i}) - p_{\eta^k}^j(o_t^k, a_t^{-k}) \right]^2. \quad (10)$$

Unlike conventional RL settings where penalty terms like behaviour costs are used, we assume that agents have mixed motivations so that they also care about the alignment between their own assessments and others' assessments (i.e., gossip from neighbours' neighbours [12]) toward the same agent. However, we argue that this negative term incurred by Equation (9) should not be included in the total reward, as the dilemma policy and evaluation policy are two separate modules. With the new trajectory $\hat{\tau}$, agent i carries out an update regarding the evaluation policy accordingly:

$$\hat{\eta}^i \leftarrow \eta^i + \lambda f(\hat{\tau}^i, \tau^i, \hat{\theta}, \eta^i), \quad (11)$$

with the same learning rate used in Equation (7). Note that when taking the gradient with respect to η^i, we need to account for the fact that $\hat{\pi}^{-i}$ depends on η^i. Finally, in alignment with the loss function proposed in [57], the update function for the evaluation policy π_{η}^i can be expressed as follows:

$$f(\hat{\tau}^i, \tau^i, \hat{\theta}, \eta^i) = \sum_{j \in \Omega^i} \sum_{t=0}^{T} \nabla_{\eta^i} \log \pi_{\hat{\theta}^j}(\hat{a}_t^j|\hat{o}_t^j) {G'}_t^i(\hat{\tau}^i, \tau, \eta^i), \quad (12)$$

where

$${G'}_t^i(\hat{\tau}, \tau, \eta^i) = \sum_{l=t}^{T} \gamma^{l-t} \left[r^{i,\text{eval}}(\hat{s}_t, \hat{a}_t) - \mu \mathcal{D}^i(o_t, a_t) \right]. \quad (13)$$

Here, the first term on the right side of Equation (13) reflects how changes in the agent's evaluation policy η^i affect its own expected comparable extrinsic reward $r^{i,\text{eval}}$ through the impact on neighbouring agent's policies $\hat{\theta}^j$. Note that there is no recursive dependence of θ^{-i} on η^i in the second term, as it is included in the previous evaluation episode.

5 EXPERIMENTAL SETUP

5.1 Evaluation Domain

In our experiment, agents are situated on an $L \times L$ square lattice with periodic boundary conditions. Vertices represent agents, and edges indicate relationships with their four neighbours. At the start of each episode, agents are randomly assigned as either cooperators or defectors, with equal probability, and given a random reputation.

In all experiments, we simulate a population of $N = 400$, each participating in all episodes. The agent parameters are distributed across 100 learner processes, each handling policy gradient updates for four agents. To generate experience, 10 parallel arenas are created, where agents interact with neighbours over $10,000$ episodes of 20 timesteps each, totalling $200,000$ steps. After each episode, sampled trajectories are aggregated and sent to learner processes for updates. Cooperation levels, representing the average frequency of cooperative strategies, are measured across all scenarios.

5.2 Practical Algorithm

To maximize long-term γ-discounted payoff, a common strategy is for each agent to optimize its policy independently using policy gradient techniques [46], such as REINFORCE [54] or Proximal Policy Optimization (PPO) [41]. In this paper, we deploy PPO as the learning algorithm. Similar to the asynchronous advantage actor-critic method [24], PPO maintains both value (critic) and policy (actor) estimates using deep neural networks and replaces the return G_t^i by $G_t^i - V^i(s_T)$. However, it improves stability by employing a clipped surrogate objective to limit policy updates. We also implement common practices like Generalize Advantage Estimation (GAE) [40] with advantage normalization and value clipping. Gradients are computed using the Adam optimizer [14] with a linear annealing learning rate.

5.3 Design of Network Architectures

In our implementation, each agent contains two independent neural networks trained on environmental experience batches to update their weights. These networks operate separately, formulating policies without parameter sharing. Each learns abstract representations from observations to maximize rewards. The architecture consists of a dual-layer perception with 32 hidden units and employs the activation function *tanh* for non-linear transformations.

For the dilemma policy network, inputs include the agent's received reputation assignments and those of its neighbours. The network outputs a policy—a probability distribution over the next binary dilemma action (cooperation or defection)—and a value function estimating discounted future returns. The evaluation policy network takes each neighbour's dilemma action and reputation assessments as input, generating a value function and an updated
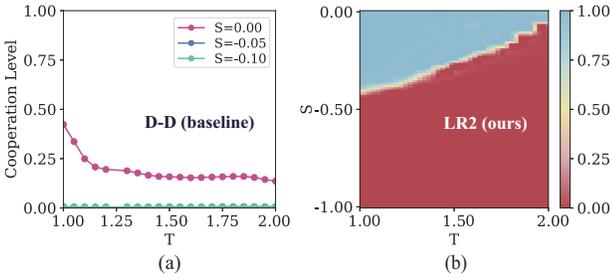
Figure 2: Comparison of cooperation levels between LR2 and the D-D baseline across different T and S values. LR2 more effectively promotes cooperation in the PD game. (a) D-D baseline: agents optimize dilemma policies based solely on environmental rewards. (b) LR2 (ours): agents utilize both dilemma and evaluation policies, with rewards reshaped by reputation. The colour gradient from red to blue represents cooperation levels ranging from 0 to 1.

assessment vector for each neighbour. While agents learn independently, they coexist in a shared environment, influencing one another's experiences and learning.

6 RESULTS

To ensure robustness, we average results over the final ten episodes and replicate each experiment five times. In the D-D baseline, agents base decisions solely on neighbors' past actions, without reputation. To isolate LR2's components, we test ablated versions where policies train separately (e.g., Independent PPO (IPPO) [3]) and reputation is excluded from rewards. While focusing on LR2's performance in the Prisoner's Dilemma (PD), we also evaluate it in Stag Hunt (SG) and Snowdrift (SH). Appendix B.2 further examines hyperparameter effects and reputation-based intrinsic rewards on cooperation.

6.1 Effectiveness of Introducing LR2

We begin by comparing the performance of our LR2 method against the D-D baseline across various PD scenarios. Figure 2 depicts cooperation levels as functions of the temptation payoff T and sucker's payoff S, with and without reputation assignment and reward reshaping. In Figure 2a, we observe that the well-established spatial reciprocity effect in evolutionary game theory [53]—which promotes cooperation through the network structure—fails to materialize in MARL settings using solely dilemma policies: even when exploitation carries no negative consequences (e.g., $S = 0$), cooperation remains below 0.25 in the D-D group. Moreover, as detailed in Appendix B.3, predefined social norms do not foster cooperation in structured MARL populations, which is also contrary to recent findings in evolutionary dynamics [25]. In contrast, Figure 2b demonstrates that agents using the LR2 method outperform the D-D baseline, significantly expanding the "wave of cooperation" in the contour plot. Moreover, LR2 agents display high sensitivity to reward changes, with a sharper transition from full cooperation (blue area) to full defection (red area). Interestingly, contrary to previous studies suggesting that cooperation is difficult to sustain under conditions of extreme dilemma strength [32], LR2 agents
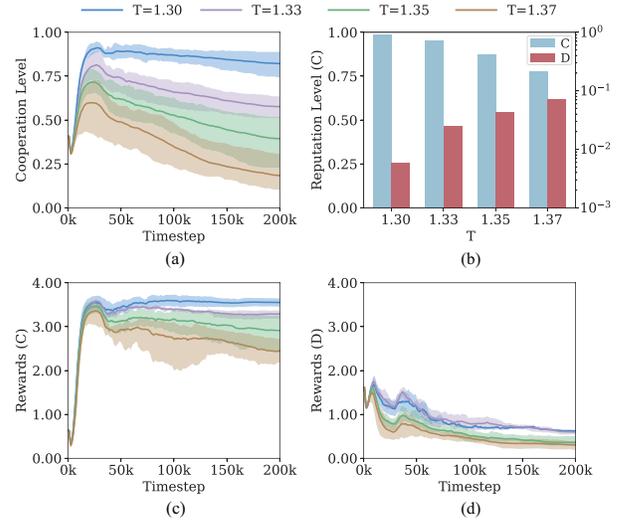

Figure 3: The evolution of cooperation with associated rewards and reputations. LR2 agents learn to assess their neighbours' behaviours to reshape rewards, fostering cooperative evolution. The evaluation includes (a) the evolutionary trajectory of cooperation levels, (b) the average reputation of cooperative and defective agents at the end of training, and (c)-(d) the rewards of cooperator and defector over time. Results are presented with the parameter T varying from 1.30 to 1.37, while S is fixed at -0.33.

maintain cooperation even when dilemma strength is very high (e.g., $T = 2.0$). This suggests the robustness of our approach in promoting the evolution of cooperation. We also assess LR2's robustness under alternative interaction structures in Appendix B.1.

6.2 Cooperation and Reputation Dynamics

We next illustrate the interplay between the evolution of cooperation and the reputation formulation process, providing a more intuitive explanation of how agents encourage prosocial behaviours by influencing others' learning dynamics. This relationship is detailed in Figure 3, which consists of four subgraphs representing the evolution of cooperation, the corresponding average rewards for cooperative and defective actions, and the reputation distribution for both actions. Each subgraph contains four parameter combinations, reflecting varying dilemma strengths in PD scenarios. As shown in Figure 3a, in all scenarios, the cooperation level exhibits a consistent trend: it briefly declines at the beginning, rises to a peak, and then either stabilizes or gradually decreases.

Since agents' rewards are reshaped by neighbour-assigned reputations, Figure 3b examines the effectiveness of the learned reputation assignment policy by comparing the average reputations of cooperators and defectors at the end of the training process. To highlight the differences, cooperative reputations are shown on a linear scale, while defective ones remain consistently below 0.1 on a log scale. The group is more sensitive to negative feedback, consistently assigning low reputations to underperforming agents regardless of the dilemma scenario. However, they possess a conservative
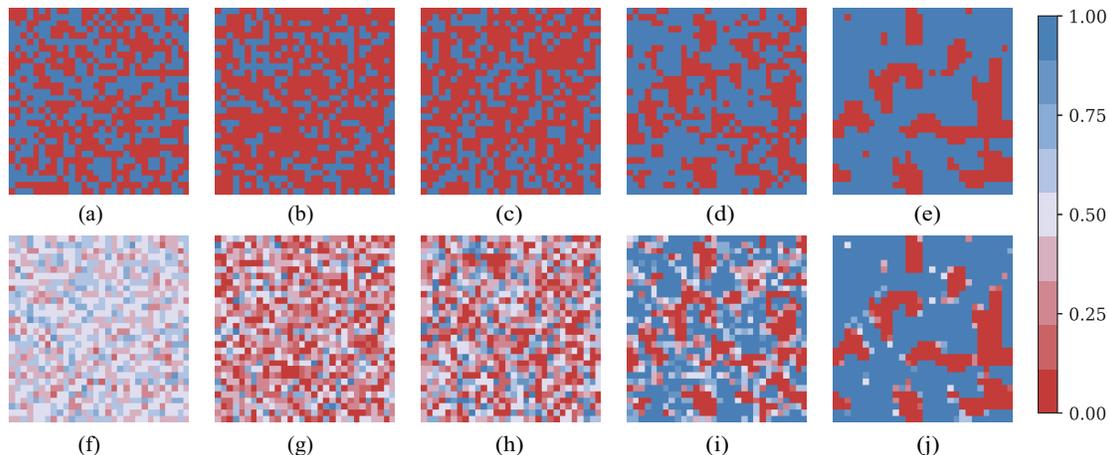
Figure 4: Representative snapshots showing the spatial distribution of learned dilemma actions and assigned reputations on a square lattice. The LR2 method fosters the formation of cooperator clusters through spatial effects, thereby enhancing the overall level of cooperation. Panels (a)-(e) display the evolutionary trajectories of two competing dilemma strategies at the timesteps t = 1k, 5k, 10k, 20k, and 50k. The corresponding panels (f)-(j) illustrate the average reputations assigned by neighbours at the same timesteps. Pixels represent agents as cooperators (blue) and defectors (red), with reputation levels ranging from 0 to 1. Results are obtained for T = 1.33 and S = −0.33.

approach to positive ratings. Notably, even at T = 1.37, the average reputation of cooperators remains around 0.75. To further illustrate how these established evaluation strategies guided the LR2 agents' learning, Figures 3c and 3d present the evolutionary trajectories of the average rewards for cooperative and defective actions. Here, the reshaped reward functions as a teaching signal, encouraging agents to act prosocially while discouraging free-riding behaviour. Initially, cooperative rewards are lower than those of defectors, but as agents learn to assign appropriate reputations to their neighbours, the rewards for cooperation gradually surpass those for defection. This demonstrates the synergy between learning reputation evaluations and dilemma-based behaviours. Our findings suggest that effective reputation assignment is crucial; as T increases, agents struggle to identify cooperators, leading to a rapid decline in cooperative rewards and the eventual extinction of cooperators.

6.3 Formation of Spatial Strategy Patterns

The above analysis provides an intuitive explanation of the synergy between the established reputation evaluation policy and the evolution of cooperation. To better understand why cooperation emerges within the LR2 framework, Figure 4 presents characteristic snapshots of agent dilemma action patterns and received reputation assignment, allowing us to further investigate the strategy distribution in the learning dynamics of LR2 agents. Interestingly, we find that LR2 not only reshapes neighbours' payoffs and influences their learning dynamics by effectively identifying dilemma strategies and assigning reputations, but also promotes the formation of co-operative clusters (abbreviated as C-clusters). This spatial pattern formation highlights the evolutionary advantage of the learned evaluation policy, as it helps resist defector invasion and facilitates the emergence of cooperation [48].

Furthermore, the evolutionary path, starting from an initial random state (Figures 4a and 4f) and progressing to a final equilibrium,

can be divided into two distinct phases, consistent with the reported spatial reciprocity effect [53]. The first phase is the enduring period (END), during which cooperators resist the invasion of defectors. This is followed by the second phase called the expansion period (EXP), where the fraction of cooperators begins to increase, sig-nalling the growth and spread of cooperative behaviour. As shown in Figures 4(a-b) and 4(f-g), during the END period, LR2 agents have not yet learned to effectively evaluate the behaviour of others, lead-ing defectors to avoid reputational damage despite their free-riding behaviour. Consequently, D-clusters expand rapidly, constraining the space available for cooperators. However, once agents learn to assign reputations correctly, they can influence the learning dynamics of others and indirectly protect their own interests. Al-though cooperation still declines during this phase, the decrease is progressively restrained as C-clusters form incrementally, allow-ing cooperators to survive and maintain their presence (Figure 4c). After a suitable transient period, the population enters the EXP phase, where agents begin to assign reputations accurately. This causes the fragmented D-clusters to gradually disintegrate, while the more cohesive C-clusters expand in size. Moreover, the finding that agents are more sensitive to losses in Figure 3, is confirmed here. As observed in the second row, LR2 agents consistently assign a reputation level of 0 to defective neighbours but are slower in providing positive evaluations to cooperative behaviours.

6.4 Architecture Ablations and Robustness

To better understand LR2, we modified its architecture to address two questions: (1) Does LR2's performance depend on agents form-ing heterogeneous evaluations of the same behaviour? (2) How is performance affected when agents merely observe reputation without integrating it into reward reshaping? To answer these, we conducted ablation experiments to illustrate the effects of reputa-tion alignment and reward on LR2's performance (Figure 5). Finally,
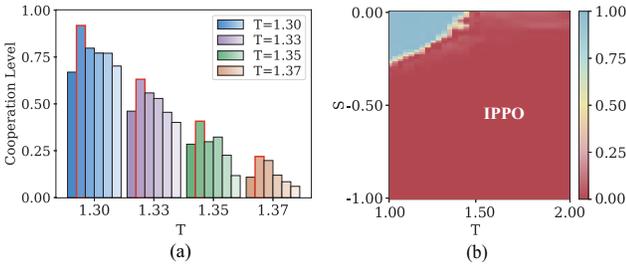
Figure 5: Ablations on LR2 architecture components. Considering others' reputation evaluations and the effects of assigned reputations most effectively promotes cooperation. (a) Reputation alignment varying importance, shown by colours from dark to light representing μ from 1 to 0. Parameter S is fixed at -0.33. (b) Performance of the IPPO training method without reputation reward.

we verified the robustness of the LR2 method in the SG and SH dilemma settings, with the results summarized in Table 1.

Dual impact from local assignment. In LR2, agents consider both the rewards from interactions and the consistency of their evaluations with others. Since reputation assignment is a biased judgment based on private observation, isolated disagreements may arise. Figure 5a compares six cases, ranging from completely ignoring reputation alignment ($\mu = 0$, darkest bar) to treating it as equally important as rewards ($\mu = 1$, lightest bar) in 0.2 intervals. As shown in all cases, moderate consideration of evaluation consistency ($\mu = 0.2$, red border) consistently leads to optimal cooperation. While aligning evaluations with neighbours can help agents learn prosocial behaviour, placing too much emphasis on alignment can be detrimental. For example when $T = 1.37$, treating alignment and rewards equally results in the worst performance.

Decoupling reputation from LR2 rewards. In Figure 5b, we consider an IPPO ablation where each agent still trains two policies, dilemma and evaluation, but reputation is used only as observational information and does not affect rewards, by setting $\beta = 1$. Compared to the rewards in Figure 2, IPPO still outperforms the baseline by using reputation to assess the quality of an individual's past behaviour, since it enriches observational information. However, the "wave of cooperation" shrinks sharply compared to LR2 training. It suggests that reward shaping proves crucial for sustaining cooperative behaviour, particularly when dilemma strength intensifies, as seen in the right top corner of both contour plots.

Table 1: Comparison of LR2 agents, the D-D baseline, and an ablated IPPO method across three dilemmas. Performance is measured by the cooperation level post-training. LR2 agents (ours) achieve higher cooperation than both alternatives in all scenarios.

Method	Prisoner's Dilemma		Snowdrift Game		Stag-Hunt Game	
	(1.1,−0.1)	(1.3,−0.3)	(1.1,0.1)	(1.3,0.3)	(0.9,−0.1)	(0.7,−0.3)
D-D	0.00	0.00	0.50	0.49	0.10	0.28
IPPO	0.95	0.00	0.99	0.98	0.98	0.99
LR2 (ours)	1.00	0.98	1.00	0.99	1.00	1.00

Robustness of LR2 in different dilemma types. We conclude the Results section by evaluating the robustness of the proposed LR2 method in promoting cooperation across different dilemma types, comparing it to the baseline D-D method and ablation IPPO version. As shown in Table 1, we assess two parameter combinations in each of the SD, SG, and SH scenarios, representing weak and strong dilemma strengths. LR2 consistently exhibits strong performance in both cases, highlighting the effectiveness of our approach. While IPPO approaches similar cooperation levels to LR2 under weak dilemmas, it fails to maintain this effect as the dilemma intensity increases, emphasizing the importance of incorporating reputation rewards and assignments into the learning process.

7 CONCLUSION AND FUTURE WORK

We introduce Learning with Reputation Reward (LR2), a method that leverages reshaped rewards based on bottom-up reputation to promote cooperative behaviour. LR2 incorporates reputation assignment into the dilemma policy learning, while agents concurrently learn an evaluation policy that assigns reputations to incentivise prosocial actions within local groups and maximise their extrinsic objectives. Unlike previous IR methods in MARL that often rely on predefined social norms to guide reactive learning, LR2 allows agents to privately assess each other's reputations based on their own interactions and local observations, further optimizing behaviour using assigned reputation information. This approach provides a pathway to predict the coevolutionary dynamics of reputation and cooperation in structured populations.

Our findings show that LR2 not only stabilizes but also enhances cooperation in spatial social dilemmas. LR2 agents are more sensitive to poor performance, allowing them to effectively identify and penalise defective neighbours. Because reshaped rewards force agents to consider the causal impact of their actions on neighbours, they adopt myopic best responses that lead to improved collective performance. This supports the assertion that even when reputations are privately assessed without enforcement from a top-down institution, emerging reputations can still facilitate cooperation through IR methods [12, 13, 31]. Moreover, LR2's promotion of cooperative behaviour is not solely due to reward reshaping but also to its ability to foster the clustering of cooperative strategies, further enhancing spatial reciprocity [10, 34, 53].

Our results raise several unresolved questions regarding the bottom-up reputation in cooperative MARL. While LR2 emphasizes the importance of local observations, agents may receive conflicting information from multiple sources with different weights. Future research could explore how agents process such information in structured populations [55] and consider extending LR2 to more complex evaluation frameworks, such as Melting Pot [15].

SUPPLEMENTARY MATERIAL

The full version with Appendix at https://arxiv.org/abs/2502.01971. Source code is available at https://github.com/itstyren/LR2.

REFERENCES

[1] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '21)*. Richland, SC, 115–123.

[2] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7047–7054.

[3] Matteo Bettini, Amanda Prorok, and Vincent Moens. 2024. Benchmarl: Benchmarking Multi-Agent Reinforcement Learning. *Journal of Machine Learning Research* 25, 217 (2024), 1–10.

[4] Vincent Conitzer and Caspar Oesterheld. 2023. Foundations of Cooperative AI. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23, Vol. 37)*. 15359–15367.

[5] Yali Du, Joel Z Leibo, Usman Islam, Richard Willis, and Peter Sunehag. 2023. A Review of Cooperation in Multi-Agent Learning. *arXiv preprint arXiv:2312.05162* (2023).

[6] 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.

[7] Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. Richland, SC, 122–130.

[8] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement Learning with Deep Energy-based Policies. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. 1352–1361.

[9] Chris Haynes, Michael Luck, Peter McBurney, Samhar Mahmoud, Tomáš Vítek, and Simon Miles. 2017. Engineering the Emergence of Norms: A Review. *The Knowledge Engineering Review* 32 (2017), e18.

[10] Yujie He, Tianyu Ren, Xiao-Jun Zeng, Huawen Liang, Liukai Yu, and Junjun Zheng. 2024. Temporal interaction and its role in the evolution of cooperation. *Physical Review E* 110, 2 (2024), 024210.

[11] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. *Advances in Neural Information Processing Systems* 31 (2018).

[12] Mari Kawakatsu, Taylor A Kessinger, and Joshua B Plotkin. 2024. A Mechanistic Model of Gossip, Reputations, and Cooperation. *Proceedings of the National Academy of Sciences* 121, 20 (2024), e2400689121.

[13] Taylor A Kessinger, Corina E Tarnita, and Joshua B Plotkin. 2023. Evolution of Norms for Judging Social Behavior. *Proceedings of the National Academy of Sciences* 120, 24 (2023), e2219480120.

[14] Diederik P Kingma. 2015. Adam: A method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[15] Joel Z Leibo, Edgar A Dueñez-Guzman, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charlie Beattie, Igor Mordatch, and Thore Graepel. 2021. Scalable Evaluation of Multi-Agent Reinforcement Learning with Melting Pot. In *Proceedings of the 38th International Conference on International Conference on Machine Learning - Volume 139 (ICML'21)*. 6187–6199.

[16] 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 (AAMAS '17)*. Richland, SC, 464–473.

[17] Xinle Liang, Yang Liu, Tianjian Chen, Ming Liu, and Qiang Yang. 2022. Federated Transfer Reinforcement Learning for Autonomous Driving. In *Federated and Transfer Learning*. Springer, 357–371.

[18] Michael L. Littman. 1994. Markov Games as A Framework for Multi-Agent Reinforcement Learning. In *Machine Learning Proceedings 1994*. Elsevier, 157–163.

[19] Qinghua Liu, Csaba Szepesvári, and Chi Jin. 2022. Sample-Efficient Reinforcement Learning of Partially Observable Markov Games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22, Vol. 35)*. Red Hook, NY, USA, 18296–18308.

[20] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA, 6382–6393.

[21] Michael W. Macy and Andreas Flache. 2002. Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences* 99, suppl_3 (2002), 7229–7236.

[22] Kevin R McKee, Edward Hughes, Tina O Zhu, Martin J Chadwick, Raphael Koster, Antonio Garcia Castaneda, Charlie Beattie, Thore Graepel, Matt Botvinick, and

Joel Z Leibo. 2021. A Multi-Agent Reinforcement Learning Model of Reputation and Cooperation in Human Groups. *arXiv preprint arXiv:2103.04982* (2021).

[23] Kevin R McKee, Andrea Tacchetti, Michiel A Bakker, Jan Balaguer, Lucy Campbell-Gillingham, Richard Everett, and Matthew Botvinick. 2023. Scaffolding Cooperation in Human Groups with Deep Reinforcement Learning. *Nature Human Behaviour* 7, 10 (2023), 1787–1796.

[24] Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P. Lillicrap, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48 (ICML'16)*. 1928–1937.

[25] Yohsuke Murase and Christian Hilbe. 2024. Computational evolution of social norms in well-mixed and group-structured populations. *Proceedings of the National Academy of Sciences* 121, 33 (2024), e2406885121.

[26] Martin A Nowak. 2006. Five Rules for the Evolution of Cooperation. *Science* 314, 5805 (2006), 1560–1563.

[27] Martin A Nowak and Karl Sigmund. 2005. Evolution of Indirect Reciprocity. *Nature* 437, 7063 (2005), 1291–1298.

[28] Hisashi Ohtsuki and Yoh Iwasa. 2004. How Should We Define Goodness? —Reputation Dynamics in Indirect Reciprocity. *Journal of Theoretical Biology* 231, 1 (2004), 107–120.

[29] Ninell Oldenburg and Tan Zhi-Xuan. 2024. Learning and Sustaining Shared Normative Systems via Bayesian Rule Induction in Markov Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24)*. Richland, SC, 1510–1520.

[30] Julien Perolat, Joel Z. Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A Multi-Agent Reinforcement Learning Model of Common-Pool Resource Appropriation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Red Hook, NY, USA, 3646–3655.

[31] Shirsendu Podder, Simone Righi, and Károly Takács. 2021. Local Reputation, Local Selection, and the Leading Eight Norms. *Scientific Reports* 11, 1 (2021), 16560.

[32] Tianyu Ren and Xiao-Jun Zeng. 2023. Reputation-Based Interaction Promotes Cooperation with Reinforcement Learning. *IEEE Transactions on Evolutionary Computation* 28, 1177–1188 (2023), 042145.

[33] 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 (IJCAI '24)*. 193–201.

[34] Tianyu Ren and Junjun Zheng. 2021. Evolutionary dynamics in the spatial public goods game with tolerance-based expulsion and cooperation. *Chaos, Solitons & Fractals* 151 (2021), 111241.

[35] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. 2006. Evolutionary Dynamics of Social Dilemmas in Structured Heterogeneous Populations. *Proceedings of the National Academy of Sciences* 103, 9 (2006), 3490–3494.

[36] Fernando P Santos, Jorge M Pacheco, and Francisco C Santos. 2021. The Complexity of Human Cooperation under Indirect Reciprocity. *Philosophical Transactions of the Royal Society B* 376, 1838 (2021), 20200291.

[37] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2016. Social Norms of Cooperation in Small-Scale Societies. *PLOS Computational Biology* 12, 1 (2016), 1–13.

[38] Fernando P Santos, Francisco C Santos, and Jorge M Pacheco. 2018. Social Norm Complexity and Past Reputations in the Evolution of Cooperation. *Nature* 555, 7695 (2018), 242–245.

[39] Bastin Tony Roy Savarimuthu and Stephen Cranefield. 2011. Norm Creation, Spreading and Emergence: A Survey of Simulation Models of Norms in Multi-Agent Systems. *Multiagent and Grid Systems* 7, 1 (2011), 21–54.

[40] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[42] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. 2010. Social Learning Promotes Institutions for Governing the Commons. *Nature* 466, 7308 (2010), 861–863.

[43] Martin Smit and Fernando P. Santos. 2024. Learning Fair Cooperation in Mixed-Motive Games with Indirect Reciprocity. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24)*. 220–228.

[44] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 97*. ICML '19, 5887–5896.

[45] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-Decomposition Networks for Cooperative

Multi-Agent Learning Based on Team Reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '18)*. Richland, SC, 2085–2087.

[46] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. Cambridge, MA, USA, 1057–1063.

[47] György Szabó and Gabor Fath. 2007. Evolutionary Games on Graphs. *Physics Reports* 446, 4-6 (2007), 97–216.

[48] Attila Szolnoki and Xiaojie Chen. 2017. Alliance Formation with Exclusion in the Spatial Public Goods Game. *Physical Review E* 95, 5 (2017), 052316.

[49] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2023. Modeling Moral Choices in Social Dilemmas with Multi-Agent Reinforcement Learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)*. 317–325.

[50] Paul AM Van Lange, Jeff Joireman, Craig D Parks, and Eric Van Dijk. 2013. The Psychology of Social Dilemmas: A Review. *Organizational Behavior and Human Decision Processes* 120, 2 (2013), 125–141.

[51] Matthijs Van Veelen, Julián García, David G Rand, and Martin A Nowak. 2012. Direct Reciprocity in Structured Populations. *Proceedings of the National Academy of Sciences* 109, 25 (2012), 9929–9934.

[52] 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.

[53] Zhen Wang, Satoshi Kokubo, Jun Tanimoto, Eriko Fukuda, and Keizo Shigaki. 2013. Insight into the So-Called Spatial Reciprocity. *Physical Review E* 88, 4 (2013), 042145.

[54] Ronald J Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Machine learning* 8 (1992), 229–256.

[55] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 32, 1 (2020), 4–24.

[56] Jason Xu, Julian García, and Toby Handfield. 2019. Cooperation with Bottom-up Reputation Dynamics. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '19)*. Richland, SC, 269–276.

[57] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to Incentivize Other Learning Agents. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*. Red Hook, NY, USA, Article 1275, 12 pages.

[58] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22, Vol. 35)*. Red Hook, NY, USA, 24611–24624.

[59] Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C Parkes, and Richard Socher. 2022. The AI Economist: Taxation Policy Design via Two-Level Deep Multiagent Reinforcement Learning. *Science Advances* 8, 18 (2022), eabk2607.