

Matching Gains with Pays: Effective and Fair Learning in Multi-Agent Public Goods Dilemmas

Yitian Chen^{a,b,1}, Xuan Liu^{b,c,*}, Shigeng Zhang^a, Xinning Chen^b and Song Guo^d

^aSchool of Computer Science and Engineering, Central South University, China

^bCollege of Computer Science and Electronic Engineering, Hunan University, China

^cThe Ministry of Education Key Laboratory of "Fusion Computing of Supercomputing and Artificial Intelligence", China

^dDepartment of Computer Science and Engineering, Hong Kong University of Science and Technology, China

Abstract. The training of multi-agent reinforcement learning (MARL) tasks with the public goods dilemma (PGD) is difficult because the selfish actions of individual agents for high personal rewards may reduce the collective utility of the whole group. Existing solutions to this problem, e.g., reward gifting or intrinsic rewards, although inducing cooperation among agents in small groups, cannot guarantee fairness among agents' policies and fail to achieve optimal group utility in large-scale systems. In this paper, we propose F4PGD, an effective method to train large-scale MARL tasks with PGD in a decentralized manner, which is inspired by Adam's equity theory that the match between a person's payoff and his contribution is the key incentive for people to contribute to the common good. In F4PGD, a mechanism is designed to match an agent's reward with its contribution, which suppresses agents from taking a free ride and meanwhile encourages well-learned agents to contribute to public goods. Experimental results show that F4PGD effectively learns optimal policies for the whole group and guarantees fairness among agents in several typical MARL tasks with PGD.

1 Introduction

Multi-agent reinforcement learning (MARL) has shown great potential in cooperative and competitive multi-agent decision control problems [21, 15, 5, 23, 7]. In MARL, agents are trained to optimize their policies for completing tasks by maximizing some pre-defined rewards [11, 34, 10]. Training in MARL can be either centralized or decentralized. The centralized training mechanism is suitable for small-scale collaborative tasks as it can effectively mitigate the non-stationarity problem in MARL. However, training a central controller in complex tasks with large state and action spaces requires significant computational resources, and it also faces the risk of single-point failure. In contrast, the decentralized training mechanism uses local observations of agents to train the policy and thus is more suitable for large-scale multi-agent systems. In decentralized training, a well-designed reward function is crucial for agents to learn how to cooperate since agents tend to act selfishly for high individual rewards.

When a MARL task requires some agents to sacrifice individual payoff for the utility of the whole group, it will face the public goods dilemma (PGD) [18], making it difficult to train collaborative agents

in a decentralized manner. PGD is a form of social dilemma, which is caused by the conflict between the short-term reward of individual agents and the long-term collective utility of the whole group. The dilemma widely exists in human societies, such as wearing masks during the pandemic, vaccinating, unpaid blood donation, and constructing roads and schools. In PGD, selfish individuals can obtain high payoffs in the short term by taking free rides, which however might reduce the collective utility of the whole group in the long term or even fail the task. Moreover, the fear of being exploited by others also discourages individuals from providing public goods.

There have been many works on training cooperative multi-agent systems and eliminate social dilemmas in MARL [2, 6, 29]. However, they mainly focus on simple matrix games. Considering that real-world social dilemmas are temporally extended, Leibo et al. [19] propose a sequential social dilemma (SSD) model where cooperating and defecting are not simple one-step actions but properties of agents' strategies. Based on this model, there have been many recent works on solving complex and realistic social dilemmas, which can be classified into two categories: reward gifting [22, 32, 31] and intrinsic reward [14, 30, 16, 12, 24]. However, these works either neglect fairness among agents, which is essential for system robustness, or introduce extra overhead in training agents to reward others, thus failing to train large-scale systems in MARL tasks with PGDs.

In this paper, we propose **F**airness-based training for **P**ublic Goods **D**ilemmas (F4PGD) to effectively train MARL tasks with PGDs in a decentralized manner. We reveal that the mismatch between an agent's payoff and its contribution is one of the key problems that hinders agents from collaborating for optimal group utility. Existing works [17, 35, 13] tend to assign equal awards to individual agents, which might suppress well-learned agents to contribute to the public goods. To address this problem, we resort to the *equity theory* proposed by Adams [1], which suggests that individuals prefer to use relative comparison between outcome and input rather than the absolute outcome to judge whether the compensation is fair or not. Inspired by this theory, we propose to use the matching degree between an agent's payoff and its contribution as the fairness metric to guide the training of agents, which can encourage well-learned agents to contribute to public goods and meanwhile suppress selfish agents from benefiting from the public goods by taking free rides. Moreover, the agents with poor performance can also access the valuable experience to optimize their policies because of the sufficient public goods

* Corresponding Author. Email: xuan_liu@hnu.edu.cn

¹ Equal contribution.

provided by well-learned agents.

To summarize, we make the following contributions in this paper:

1) We reveal that the matching between an agent's payoff and its contribution, rather than equal reward, is the key incentive to solve the PGD problem in MARL training.

2) We propose F4PGD, a decentralized training method that effectively trains cooperative agents in MARL tasks with PGDs and finds optimal policies even when the number of agents is large.

3) We evaluate the performance of F4PGD and compare it with state-of-the-art solutions in different types of PGD tasks. Experimental results show that agents trained with F4PGD develop efficient cooperative policies and balance contributing and benefiting behaviors well in PGDs. Moreover, our work proves to be effective even in the more complex and realistic sequential public goods dilemma.

2 Preliminaries

2.1 Multi-agent Reinforcement Learning

Consider an N -player Markov game \mathcal{M} which is defined as a tuple $\langle N, S, A, R, \mathcal{T} \rangle$, where S and A are the finite sets of states and actions respectively. $\mathcal{T} : S \times (A_1 \times A_2 \times \dots \times A_N) \rightarrow \Delta(S)$ is the transition probability function describing the probability distribution of the next state given the current state and the joint actions. For each player $i \in N$, $R_i : S \times (A_1 \times A_2 \times \dots \times A_N) \rightarrow \mathbb{R}$ is the reward it obtains after the joint actions are executed in a certain state.

When discussing social dilemmas in multi-agent reinforcement learning, each agent is trained and rewarded independently. At time step t , agent i takes an action $a_{i,t}$ based on its observation $o_{i,t} = (s_t, i)$ and receives an extrinsic reward $r_{i,t}(s_t, \mathbf{a}_t)$. Each agent aims to learn a policy $\pi_i(o_i)$ which maximizes its long-term accumulative discount reward $\mathbb{E} \left[\sum_{t=0}^T \gamma^t r_{i,t} \right]$, where T is the length of an episode and $\gamma \in [0, 1)$ is the discount factor. In this work, we deploy advantage actor-critic (A2C) algorithm to train each agent independently, and we induce cooperation for higher social welfare by adding an intrinsic reward which is defined in Section 3.

2.2 Public Goods Dilemmas

The public goods dilemma, a common issue in behavioral economics and social science, also exists in MARL. This problem intensifies in large-scale systems as individual motivation to engage in public affairs decreases with an increasing number of agents. To illustrate this, we introduce two different types of PGDs.

Donation Game Consider a basic PGD scenario involving N players where each player can choose to donate part of its bonus to the common pool, as Fig.1(a) shows. The bonuses invested in the common pool will be multiplied by m to produce social welfare, $m \in (1, N)$. Then all the players share the social welfare equally. Denote the donation of player i as d_i . The final payoff of player i can be defined as $p_i = \frac{m \sum_{j=1}^N d_j}{N} - d_i$. We set m to 1.5 and allow the players to donate none/few/lots of their bonuses (i.e. 0, 2, 5) at each time step. Despite that the more agents donate, the more social welfare will be generated in this scenario, the agents face the temptation of taking free rides by making no donation. A large group scale intensifies this problem because the social welfare each agent can obtain from a shared common pool becomes less. Moreover, the donation of each agent has less influence over the average social welfare.

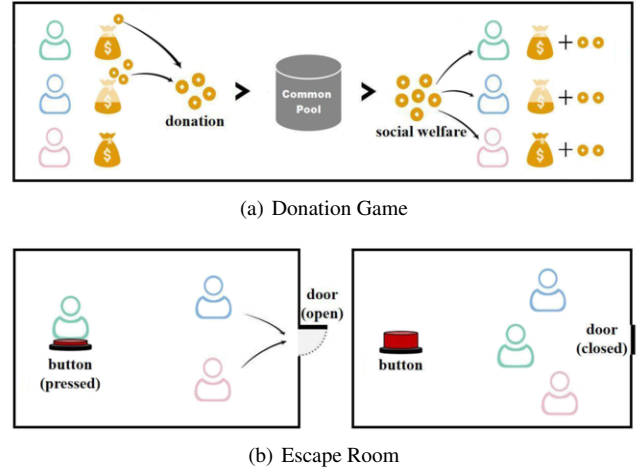


Figure 1. Sketch maps of multi-agent public goods dilemmas.

Escape Room The Escape Room game [32] shown in Fig.1(b) is a threshold public goods dilemma [4], where social welfare will not be generated before enough contributions are made by the individuals. In this scenario, N players aim to escape from the room through a door that opens only if no less than M agents press the button simultaneously. Those who press the button will not be able to escape. An agent gains a reward of 2 for escaping and loses 1 for pressing the button. If an agent stays still or approaches a closed door, it obtains no reward. Agents lack prosocial motivation in Escape Room for two reasons. First, if fewer than M agents make efforts to open the door, those who choose to press the button will be punished without generating any collective benefit. Second, individuals can only benefit from others' contribution, while providing public goods cannot generate any direct benefits for themselves, so the contradiction between individual payoffs and collective benefits is more prominent.

3 Methodology

In this section, we introduce F4PGD in three parts which mainly address the following issues: 1) How to define fairness in public goods dilemmas? 2) How to evaluate fairness in the training of agents since it varies with time? 3) How to set parameters reasonably and train agents to cooperate efficiently in PGDs?

3.1 Fairness Reward Modeling

Equity theory studies the rationality and fairness of compensation distribution and its influence on workers' enthusiasm for production. As suggested by equity theory, individuals generally make social comparisons in terms of inputs and outcomes. Denote one's personal input and outcome as I_P and O_P respectively, and his estimation of the colleagues' inputs and outcomes are I_C and O_C . According to equity theory, an individual feels fair if $\frac{O_P}{I_P} \approx \frac{O_C}{I_C}$. In multi-agent tasks involving PGDs, we follow the notion that if an agent receives a disproportionately high payoff relative to its contribution, it introduces unfairness to the group. Consequently, we define an agent's fairness reward, or in other words, the penalty for unfairness, as a non-positive value computed as the difference between its relative contribution and relative payoff. Relative contribution (payoff) is the ratio of an agent's own contribution (payoff) to the group's average contribution (payoff). Although the units of contribution and payoff may vary across different tasks, computing the ratio can eliminate

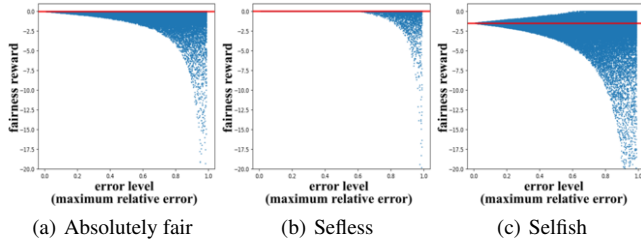


Figure 2. Simulation experiments. Estimation errors δ_c, δ_p are sampled from a uniform distribution $U(0, \delta_{max})$ independently. The horizontal axis is the maximum relative error δ_{max} , and the vertical axis is the fairness reward. The red lines represent the true fairness rewards in each case.

Table 1. The proportion of points that have significant deviation (Deviating from the true fairness reward by more than 0.5) within each error interval.

Maximum Relative Error	(0, 10%)	[10%, 20%)	[20%, 30%)	[30%, 40%)	[40%, 50%)
Fair	0.00%	0.00%	0.77%	6.36%	13.85%
Selfless	0.00%	0.00%	0.00%	0.00%	0.00%
Selfish	0.00%	0.12%	10.57%	28.40%	41.54%
Maximum Relative Error	[50%, 60%)	[60%, 70%)	[70%, 80%)	[80%, 90%)	[90%, 1)
Fair	18.21%	22.45%	25.65%	28.45%	29.60%
Selfless	0.00%	0.00%	0.69%	3.65%	7.02%
Selfish	53.39%	60.89%	66.51%	71.70%	74.92%

this issue. Let C_i^t and P_i^t denote the contribution and payoff of agent i , \bar{C}_g^t and \bar{P}_g^t denote the group's average contribution and payoff levels at time t . We define agent i 's fairness reward at time t as:

$$f_i^t = \min\left(0, \frac{C_i^t}{\bar{C}_g^t} - \frac{P_i^t}{\bar{P}_g^t}\right). \quad (1)$$

Note that the fairness rewards of agents who have sufficient contributions are 0 according to Eq. 1, because providing public goods is only the way to produce social welfare rather than agents' primary goal. It is illogical to allow agents to gain benefits directly from providing public goods, so we take the minimum of $\frac{C_i^t}{\bar{C}_g^t} - \frac{P_i^t}{\bar{P}_g^t}$ and zero in order to avoid agents from acquiring absolutely selfless behaviors which may greatly reduce the overall efficiency of the system.

According to Eq. 1, it is essential that \bar{C}_g^t and \bar{P}_g^t are available to each agent, which requires centralized information. Nevertheless, this does not introduce significant cost to the training process since each agent is trained independently. Additionally, such information can be estimated in decentralized manners through applying distributed computing methods that focus on aggregation calculation problems or enabling agents to exchange local information with each other. However, estimating the average values \bar{C}_g^t and \bar{P}_g^t will inevitably introduce errors to the calculation of fairness reward, thus we study the robustness of our method against estimation errors in two steps. First, we discuss how different levels of errors affect the fairness rewards of agents exhibiting different behaviors by simulating 3 specific cases of agents in this section. Second, we add different levels of errors in the training of agents to study how they affect the practical performance of our method in 4.2 and 5.2.

Here we simulate 3 specific cases to infer how estimation errors affect agents with different behaviors in calculating fairness rewards:

- Absolutely fair: $C_i^t = \bar{C}_g^t, P_i^t = \bar{P}_g^t$. In this case, the fairness reward is calculated as $\min\left(0, \frac{1}{1 \pm \delta_c} - \frac{1}{1 \pm \delta_p}\right)$.
- Selfless: $C_i^t = 2 \cdot \bar{C}_g^t, P_i^t = 0.5 \cdot \bar{P}_g^t$. In this case, the fairness reward is calculated as $\min\left(0, \frac{2}{1 \pm \delta_c} - \frac{0.5}{1 \pm \delta_p}\right)$.

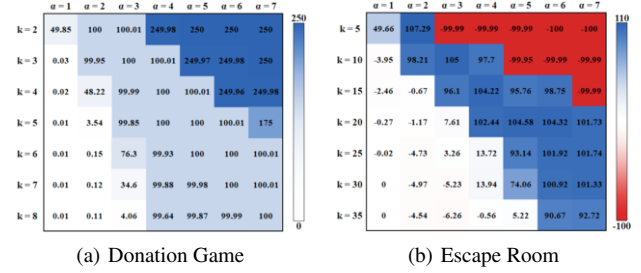


Figure 3. Average individual rewards of 10 F4PGD agents with different combinations of parameters α and k .

- Selfish: $C_i^t = 0.5 \cdot \bar{C}_g^t, P_i^t = 2 \cdot \bar{P}_g^t$. In this case, the fairness reward is calculated as $\min\left(0, \frac{0.5}{1 \pm \delta_c} - \frac{2}{1 \pm \delta_p}\right)$.

We use scatter diagrams to display how the calculation results of fairness rewards are affected by errors δ_c and δ_p , which are sampled from a uniform distribution independently, i.e. $\delta_c, \delta_p \sim U(0, \delta_{max})$. The parameter $\delta_{max} \in (0, 1)$ is the maximum relative error and different values of δ_{max} represent different error levels. As Fig. 2 shows, in each case, the range of fairness reward deviation (i.e. the vertical distances from the data points to the red lines) expands with the increase of error levels. Only when the errors are quite large can the calculation deviation of fairness reward be significant. To further look into the probability of significant deviations in fairness rewards, we divide the points in Fig. 2 into 10 intervals according to the error levels, and then within each interval, we calculate the proportion of points whose deviation from the true fairness reward exceeds 0.5. Table 1 shows the probability of significant deviations in fairness rewards within each error interval. The results show that the fairness rewards of selfless agents can hardly be affected by errors in the estimation of the group's average contribution and payoff. This is attributed to the fact that selfless agents typically exhibit higher relative contributions and lower relative payoffs, resulting in a fairness reward of 0 after the minimum value operation. Consequently, only when the error is significant enough to reverse $\frac{C_i^t}{\bar{C}_g^t} - \frac{P_i^t}{\bar{P}_g^t}$ from positive to negative can it affect the calculation result. In contrast, the errors have much more significant impact on selfish agents' fairness rewards. However, the impact is twofold, as it can either increase or decrease the punishment for free-riding behaviors. Overall, prosocial agents are less affected by estimation errors and agents who free-ride face a probability of receiving heavier punishments. Since the elimination of public good dilemmas mainly depends on the agents who provide sufficient public goods, our fairness reward is robust to the estimation errors of the group's average contribution and payoff.

3.2 Fairness-based Training for PGDs

Since RL agents are designed to handle sequential decision making problems, it is of vital importance to reasonably assess agents' contribution and payoff levels that vary with time before applying the fairness reward as the intrinsic reward of agents. In view of this, we use an accumulative variable buffer to record the contributions and rewards of each agent for the most recent k steps. Specifically, at step t , new data are involved in the buffer and the most distant data (i.e. the data recorded at step $t - k$) are removed. The parameter k is the length of the buffer, which can be adjusted flexibly according to practical scenarios. Thus, at time t , we compute the contribution and payoff of agent i by the sums of one-step contributions and payoffs recorded in its buffer respectively, i.e. $C_i^t(k) = \sum_{\tau=t-k+1}^t c_i^\tau$,

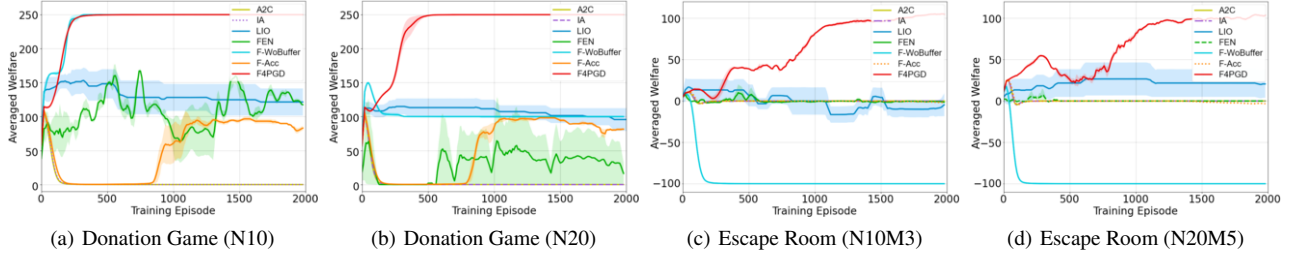


Figure 4. Episode average welfare (i.e. average net profit of agents within an episode) of different methods in Donation Game and Escape Room.

$P_i^t(k) = \sum_{\tau=t-k+1}^t r_i^\tau$. The key point to design a proper accumulative variable buffer is to include a reasonable amount of valid data in the buffer to assess an agent’s payoff and contribution. If the buffer is too short, the fairness reward will change dramatically over time because each single-step action influences the assessment of contribution and payoff levels greatly. On the contrary, an excessively long buffer will dilute the effect of single-step actions and further invalidate the fairness reward.

Then we combine the accumulative variable buffer with the fairness reward to obtain the intrinsic rewards of agents. The utility of agent i at time step t is the weighted sum of its environmental reward r_i^t and its intrinsic reward $f_i^t(k)$, which can be calculated as follows:

$$u_i^t = r_i^t + \alpha \cdot f_i^t(k) = r_i^t + \alpha \cdot \min(0, \frac{C_i^t(k)}{C_g^t(k)} - \frac{P_i^t(k)}{P_g^t(k)}). \quad (2)$$

The parameter α controls the weight of intrinsic rewards. The principle behind the intrinsic reward is that selfish agents who try to obtain higher payoffs by free-riding will face heavier punishments, as such behavior generally leads to high payoff levels with low contribution levels. In contrast, the punishments for prosocial agents are negligible or non-existent owing to their high contribution levels. In this way, an agent can reduce or even avoid punishments by making a sufficient contribution to the public welfare.

3.3 Parameter Analysis

In order to investigate how different parameter settings for F4PGD affect the behaviors agents acquire, we train populations of 10 agents using different combinations of intrinsic reward weight α and buffer length k in Donation Game and Escape Room respectively. In Escape Room, only when no less than 3 agents press the button will the door be open. Fig.3 shows the average individual rewards in different settings. Each episode includes 100 steps and the reported values are calculated by averaging the results of the last ten episodes of training.

In spite of the difference in the upper right corners of Fig.3(a) and Fig.3(b), the results actually reflect the same pattern, namely, when the weight of intrinsic reward is heavier and the accumulative variable buffer is shorter, agents tend to behave more selfishly. The difference in the upper right corners between these two figures derives from the fact that selfish behaviors lead to different average individual rewards in these two games. In Donation Game, the more selfishly agents behave, the more social welfare they produce, resulting in higher average rewards. However, in Escape Room, if all the agents exhibit absolute selflessness by pressing the button without anyone escaping, they will pay the cost in vain without generating any social welfare, thus the average rewards are negative.

The results provide empirical guidance on how to adjust the parameters of F4PGD to obtain different behaviors of agents. The value

of hyperparameter α directly controls the weight of intrinsic reward. In scenarios where agents tend to behave selfishly, increasing the weight of intrinsic rewards will generate more incentives for agents to adopt prosocial behaviors. For example, if providing public goods takes considerable costs, the weight of intrinsic rewards should be heavy enough for agents to acquire prosocial behaviors. It is also worth noting that an excessively heavy weight of intrinsic reward will probably lead to absolutely selfless behaviors, which rather reduces the collective efficiency greatly in some cases. The buffer length k can be set according to the density of environmental rewards. In tasks with dense rewards, the values of payoffs and contributions accumulate rapidly over time, and an excessively long buffer will dilute the impact of an agent’s immediate actions and invalidate the fairness reward. On the contrary, if the buffer is too short in tasks with sparse rewards, the values of payoffs and contributions will be too small to reasonably evaluate fairness.

4 Performance Evaluation

In this section, we introduce the baselines and present the results to show that F4PGD effectively eliminates the public goods dilemmas and improves the overall efficiency in large-scale agent groups. We evaluate the efficiency of different agent groups by the average welfare in an episode. Moreover, we show the distribution of each individual’s payoff and contribution to demonstrate that F4PGD agents strike a good balance between contributing and benefiting. Finally, we show how an agent’s behavior evolves with training and analyze how F4PGD induces cooperation in large scale systems.

4.1 Baselines

We compare F4PGD against several representative methods based on our categorization of related work. Additionally, we conducted comparisons with several ablation baselines.

- Advantage Actor-Critic (A2C) [26]: Each agent aims to optimize its individual reward generated by the environment.
- Inequity Aversion (IA) [14]: A representative method for social dilemmas using intrinsic rewards. As is concluded by previous work, it is advantageous inequity aversion that helps to resolve public goods dilemmas. We use the same parameter for that as IA.
- Learning to Incentivize Others (LIO) [32]: A representative method for social dilemmas using reward gifting mechanism. Each agent is equipped with an incentive function that maps its observation and others’ actions to a vector of rewards for others.
- Fair-Efficient Network (FEN) [17]: A representative work studying fairness in MARL, which, however, takes fairness as reward equity and focuses on resource occupation tasks.
- F-WoBuffer: An ablation variant of F4PGD only using payoffs and contributions of the current step to calculate intrinsic rewards.

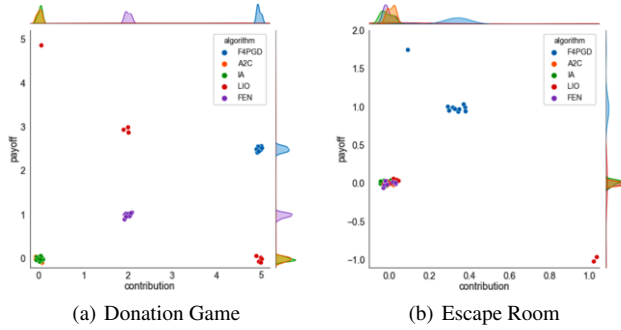


Figure 5. Payoff and contribution distributions of individuals.

- F-Acc: Another ablation variant using accumulative payoffs and contributions up to the current step to calculate intrinsic rewards.

4.2 Results

Efficiency We train groups of 10 and 20 agents in Donation Game and Escape Room respectively, with each episode consisting of 100 steps. Fig.4(a) and 4(b) show the learning curves of averaged welfare in Donation Game. In general, F4PGD outperforms the baselines across different group sizes, achieving Pareto efficiency where each agent donates 5 at each step. The ablation variant F-WoBuffer reaches optimal results with 10 agents but only sub-optimal results with 20 agents. The averaged welfare of another ablation variant F-Acc converges to about 100, where each agent chooses to donate 2 in the majority of the time steps. A2C agents, aiming to optimize individual rewards, fall into a public goods dilemma. The IA agents take advantageous inequity aversion as the intrinsic reward, a punishment to those who obtain high rewards. However, compared to punishing agents with high rewards regardless of whether their behavior is prosocial or selfish, it is more reasonable to punish those who selfishly free ride. Punishment for inequity is more likely to affect the efficiency of agents since it prevents them to reach higher scores. LIO agents outperform A2C and IA agents, but they still fall short of achieving the optimal result. Worse still, it takes more time to train LIO agents when the scale of the population is large, because the incentive function becomes more complex. FEN achieves the sub-optimal result in the 10-agent-group, but it fails to train the larger group to cooperate effectively.

In Escape Room, the minimum number of agents needed to open the door is 3 in a group of 10 and 5 in a group of 20. Fig.4(c) and 4(d) show how the averaged welfare varies with the training episode. As the results show, the average welfare that F4PGD agents achieve in different settings approaches the upper bounds (110 and 125 in groups of 10 and 20 respectively), which demonstrates that agents acquire approximate optimal joint policies with different group scales. However, the ablation baseline F-WoBuffer even leads to negative values of social welfare as every agent would rather press the button than escape. This indicates that one-step actions are insufficient to reflect the overall levels of agents' contributions and payoffs. Even if an agent chooses to escape at only one step, its relative contribution and payoff will change dramatically, thus resulting in a heavy punishment. So the price of escaping forces it to provide public goods consistently. The performance of LIO is unsatisfactory in general and varies with different random seeds. Agents trained by other baselines fail to escape, for these methods cannot generate any incentive for agents to press the button and the door always keeps closed.

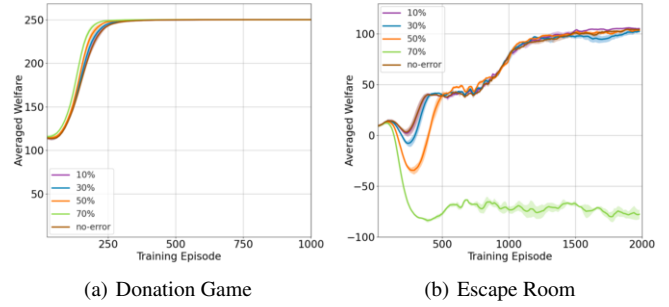


Figure 6. Error experiment in Donation Game and Escape Room.

Individual distribution To analyze the distribution of agents' payoffs and contributions and verify whether fairness is guaranteed, we test populations of 10 agents in the two games using F4PGD and four baselines. Scatter diagrams in Fig.5 show each agent's payoff and contribution. Data points in different colors represent agents trained with different algorithms. Each data point is obtained by averaging the values in a segment of the convergent period. To distinguish overlapping data points, we add a tiny random disturbance to each data point on the premise of not changing the overall distribution patterns.

Fig.5(a) shows the individuals' payoff and contribution distributions in Donation Game. Although A2C and IA agents are densely distributed, they are mainly near the coordinate origin, indicating that they do not provide public goods and thus cannot produce social welfare. It is pointless to pursue fairness with extremely low efficiency. In comparison, FEN agents show higher contributions and payoffs while maintaining a concentrated distribution. For the population trained by LIO, five of the ten agents behave selflessly, tending to donate as much as possible; four of them donate less thus obtaining higher payoffs; one of them behaves selfishly with the highest payoff. Generally, LIO agents exhibit poor fairness in Donation Game. In contrast, agents trained by F4PGD achieve higher individual payoffs than A2C, IA, and FEN agents while maintaining fairness. Overall, F4PGD performs well both in terms of fairness and efficiency. In Escape Room, each agent's level of contribution is evaluated by the probability of pressing the button. As Fig.5(b) shows, all the data points of A2C, IA, and FEN are distributed near the coordinate origin. This indicates that none of these agents take action to open the door, resulting in it remaining closed and preventing agents from escaping. Two of the LIO agents tend to press the button and have high contribution levels, but they fail to open the door. For F4PGD agents, the probabilities of pressing the button are mostly around 0.3, which matches the minimum proportion needed to open the door. All the F4PGD agents have higher individual payoffs than the baselines, indicating that most agents can escape successfully. Overall, F4PGD performs well on balancing efficiency and fairness in Escape Room.

Robustness We also evaluate the robustness of F4PGD in practical training by adding relative errors sampled from a uniform distribution $U(0, \delta_{max})$. We test F4PGD with 5 levels of errors, i.e. $\delta_{max} = 0, 10\%, 30\%, 50\%, 70\%$. The results in Fig.6 show that F4PGD is quite robust in Donation Game and remains robust to errors within a range of 50% in Escape Room. It is worth noting that in the case of $\delta_{max} = 70\%$, the averaged welfare drops to negative values, indicating that the majority of agents are inclined to press the button rather than to escape. This evidences that a larger error tends to make the group more selfless, which also confirms our analysis in Section 3.1, namely that the impact of estimation errors is minor for prosocial agents and dual-sided for selfish ones.

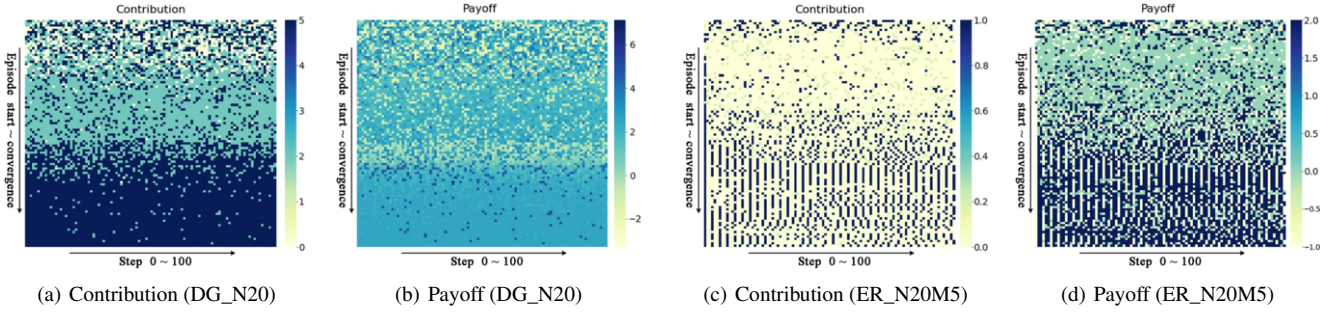


Figure 7. Behavior analysis of a randomly chosen agent. (a) and (b) show its contribution and payoff at each step in Donation Game, (c) and (d) show those in Escape Room. Each figure includes 100×100 pieces of data, i.e. 100 steps per episode \times 100 episodes sampled uniformly from the whole training process.

4.3 Behavior Analysis

To study how F4PGD agent’s behavior evolves during training, we randomly select one agent from the group and record its contribution and payoff values at each step in an episode. We uniformly sample 100 episodes from the training period for recording. The study of agents’ behavior evolution is conducted in 20-agent Donation Game and Escape Room respectively. As shown in Fig.7, in Donation Game, the agent randomly selects actions at the start of training. However, as training progresses, its strategy gradually evolves towards contributing as much as possible and consistently achieving high payoffs. In Escape Room, the agent also starts with random actions but it soon tends to approach the door to avoid the cost of pressing the button. As each agent within the group behaves selfishly, the door remains closed, and the agent’s payoff is mostly zero (light green grids in Fig.7(d)). With further exploration, the intrinsic reward penalizes those who take free rides. Finally, the agent’s behavior converges to a periodic pattern, which indicates that group members have learned to cooperate by pressing the button in turn. When training converges, the group’s joint strategy is Pareto efficient in Donation Game and approximately Pareto efficient in Escape Room. We present the proof below.

Definition 1. Denote the payoffs of a joint strategy as $s = (p_1, \dots, p_N)$. A joint strategy $s_i = (p_1^i, \dots, p_N^i)$ is Pareto efficient if and only if there is no other joint strategy $s_j = (p_1^j, \dots, p_N^j)$ satisfying $p_1^j \geq p_1^i, \dots, p_l^j > p_l^i, \dots, p_N^j \geq p_N^i$.

Proposition 1. In Donation Game, the joint strategy s_i where each agent donates as much as possible is Pareto efficient.

Proof. We prove by contradiction. Assume s_i is not Pareto efficient, then there exists another joint strategy s_j that satisfies $p_1^j \geq p_1^i, \dots, p_l^j > p_l^i, \dots, p_N^j \geq p_N^i$. In Donation Game, an agent l can only increase its individual payoff by reducing its donation, while this leads to the decrease in public welfare and the payoffs of other agents, i.e., $p_l^j < p_l^i, \dots, p_N^j > p_N^i, \dots, p_N^j < p_N^i$. This contradicts the assumption, so proposition 1 holds. \square

Proposition 2. In Escape Room, the joint strategy s_i where M agents press the button while the other $N - M$ escape is Pareto efficient (M is the minimum number of agents required to open the door, N is the total number of agents).

Proof. We prove by contradiction. Assume s_i is not Pareto efficient, then there exists another joint strategy s_j that satisfies $p_1^j \geq p_1^i, \dots, p_l^j > p_l^i, \dots, p_N^j \geq p_N^i$. In Escape Room, agents who successfully escape have attained the maximum achievable payoffs,

meaning that only the payoffs of agents who press the button have the potential for an increase. However, the door will turn closed once one of them changes its action, leading to a payoff decrease in the agents who could escape with strategy s_i . This contradicts the assumption, so the assumption is not valid, and proposition 2 holds. \square

5 Applicability in Sequential Social Dilemma

In realistic multi-agent tasks, the action and state spaces are generally much larger, and agents’ behaviors are reflected in the effects of actions over a period of time. Since Donation Game and Escape Room are simple tasks where providing public goods and taking free rides are both one-step actions, we further verify the applicability of F4PGD in a realistic sequential public goods dilemma - Cleanup. We compare F4PGD against several baselines, including representative algorithms for sequential social dilemmas and an ablation variant of F4PGD, in this grid world game with RGB observations for agents.

5.1 Cleanup Setting

Cleanup is a grid world game with sparse rewards and much larger state spaces, which poses a greater challenge for agents to acquire cooperative behaviors. In this scenario, agents aim to collect apples from a field as much as possible. However, the spawn rate of apples is inversely proportional to the amount of waste in the aquifer. Public goods can be provided by cleaning the aquifer, which does not deduct agents’ rewards but brings time costs. Agents must move away from the field for a while to clean the aquifer, and the apples generated by their labor will be soon consumed by others. There’s a dilemma between the temptation of higher individual payoffs and the risk of depleting the apples since selfish individuals tend to take free rides by collecting apples in the field and avoid being exploited. In our experiment, we set the episode length to 500 steps and extend the number of agents from 5 to 10 compared with previous works.

5.2 Results

The learning curves for collective rewards are shown in Fig.8(b). In general, our method is superior to all baselines in Cleanup, and the collective reward of F4PGD agents reaches around 900, about 4 times the rewards of A2C and IA agents. Actually, A2C and IA agents consume apples rapidly at the beginning of an episode. However, as waste increases to the depletion threshold, apples stop spawning and these agents start to wander aimlessly on the map, hardly generating any apples in the rest of the episode. Despite LIO’s excellent performance in a much smaller map with only two agents

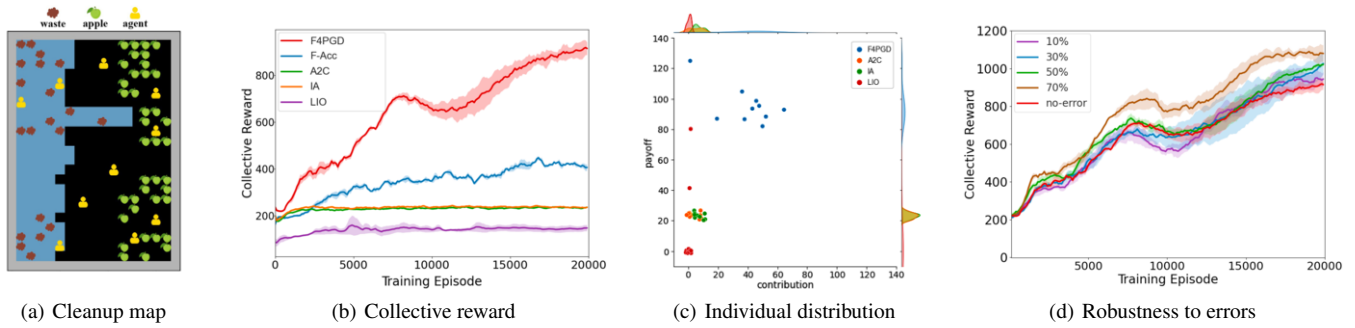


Figure 8. Cleanup map and the experimental results in Cleanup.

in previous work, it fails to learn effective strategies in the larger map with more agents. On the one hand, the incentive function of LIO takes agents' own observation and all other agents' actions as input of the network, thus the complexity of training increases with the number of agents. On the other hand, incentivizing others incurs individual costs, turning the incentive into a form of public good and causing the second-order social dilemma. The comparison between F4PGD and the ablation variant F-Acc demonstrates that a buffer with appropriate length is essential for the efficiency of agent systems. The result in Fig.8(c) indicates that A2C and IA agents collect the initial apples fairly, but fail to clean the aquifer to generate more apples. LIO agents also fail to improve collective reward by cleaning the aquifer. Worse still, the initial apples are allocated unfairly by LIO agents. In contrast, the majority of F4PGD agents acquire the behavior of cleaning the aquifer. The distribution of data points is less centralized than that in Donation Game and Escape Room, for the large state space and sparse rewards introduce more stochastic factors into the training process. Fig.8(d) shows the results of error experiment in Cleanup. The results show that F4PGD is robust to various ranges of estimation errors. It can be observed that a larger error leads to a slight increase in collective rewards, which is essentially consistent with the pattern we observed in Fig.6, i.e. the estimation error makes the group more selfless in general. Since agents are more likely to clean the aquifer, apples can spawn at a faster rate.

6 Related Work

Social dilemmas in MARL Reinforcement learning (RL) provides a way to solve sequential decision-making problems by training agents to optimize their policies for high rewards [20, 26]. In MARL, however, separately trained and rewarded agents fail to cooperate in social dilemmas, thus achieving low collective utility. Many works have studied this problem in matrix games, the widely used models to study social dilemmas in many fields [2, 6, 29]. However, real-world problems are much more complex owing to greater action spaces, larger group scales, and temporal extension. To capture these points, Leibo et al. [19] propose sequential social dilemmas (SSDs), general-sum Markov games where cooperation and defection are attributes of agents' strategies rather than elementary actions. SSDs prove to be more challenging and realistic, and recent works on this problem can be categorized into two main classes - *reward gifting* and *intrinsic reward*. The former empowers agents to give part of their own rewards to others so that cooperation behaviors are motivated by the gifts [22, 32, 8]. However, training agents to gift others brings extra expenses which are considerable in large-scale agent systems. In addition, such mechanism will result in second-order social dilemmas as agent systems scale up [9], where the gifts become

a form of public goods. The latter is inspired by the theories of other fields and takes inequity aversion [14], social influence [16], or social value orientation [24] as intrinsic rewards. However, they neglect system fairness, and the agents' strategies are diverse and specialized, which reduces their robustness in large-scale systems.

Fairness in MARL Fairness is essential for many agent systems such as traffic control [3] and cloud computing [25]. It helps to improve the efficiency and robustness of agent systems and thus has been widely studied [33, 28, 27]. In MARL, training agents to give consideration to efficiency and fairness simultaneously is a complex multi-objective, joint-policy optimization task. To solve this problem, Jiang and Lu [17] design a hierarchy consisting of several sub-policies and a controller which learns to switch between the sub-policies by optimizing a fair-efficient reward. Zimmer et al. [35] propose SOTO, a novel neural network architecture composed of self-oriented and team-oriented networks ensuring efficiency and fairness respectively. While these studies strike a good balance between efficiency and fairness, they only consider general competitive tasks where fairness is defined as an equal allocation of resources among agents. Grupen et al. [13] study cooperative multi-agent learning and propose a group-based fairness measure. However, the definition of team fairness is not applicable to mixed-motive scenarios. Adams' equity theory [1] believes that human individuals compare the relative inputs and outcomes to judge fairness. In comparison with reward-equity, fairness in matching contributions with payoffs is more promising to induce cooperation in tasks with PGDs.

7 Conclusions

In this work, we study the incentives for agents to cooperate in MARL tasks with public goods dilemmas. Inspired by Adam's equity theory and human social norms, we introduce fairness into the rewards of agents and propose F4PGD to induce cooperation. Instead of equality of agents' rewards, we define fairness as matching an agent's payoff with its contribution in this work. Experimental results show that our method eliminates public goods dilemmas effectively in tasks with different kinds of PGDs. Compared to previous works, F4PGD improves the overall efficiency of agent systems and ensures fairness among individuals. It is worth mentioning that agents trained by F4PGD tend to cooperate by providing public goods in turn, while it could be more efficient to train specialized agents that cooperate with a due division of labor in some tasks. However, systems with specialized agents neglect fairness and are less robust, because failure in any part of the task could have a heavy impact on the efficiency of the whole system. Our work provides a good trade-off between efficiency and fairness in MARL tasks with PGDs.

Acknowledgements

This work was supported in part by the National Science Foundation of China under Grants 62172154, and 62372473, in part by the Hunan Province Natural Science Foundation of China under Grant 2023JJ70016, in part by the Key Scientific and Technological Research and Development Plan of Hunan Province under Grant 2022NK2046, in part by General Research Fund under Grants 152169/22E and 152228/23E.

References

- [1] J. S. Adams. Inequity in social exchange. volume 2 of *Advances in Experimental Social Psychology*, pages 267–299. Academic Press, 1965.
- [2] N. Anastassacos, S. Hailes, and M. Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7047–7054, 2020.
- [3] B. Bakker, S. Whiteson, L. J. H. M. Kester, and F. C. A. Groen. Traffic light control by multiagent reinforcement learning systems. In *Interactive Collaborative Information Systems*, volume 281, pages 475–510. 2010.
- [4] C. B. Cadsby and E. Maynes. Voluntary provision of threshold public goods with continuous contributions: experimental evidence. *Journal of public economics*, 71(1):53–73, 1999.
- [5] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. A. Seshia, P. Abbeel, and A. D. Dragan. On the utility of learning about humans for human-ai coordination. In *Advances in Neural Information Processing Systems, NeurIPS 2019*, pages 5175–5186, 2019.
- [6] J. Chen and C. Wang. Reaching cooperation using emerging empathy and counter-empathy. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19*, pages 746–753, 2019.
- [7] X. Chen, X. Liu, S. Zhang, B. Ding, and K. Li. Goal consistency: An effective multi-agent cooperative method for multistage tasks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 172–178, 2022.
- [8] P. J. K. Christoffersen, A. A. Haupt, and D. Hadfield-Menell. 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, AAMAS 2023*, pages 448–456, 2023.
- [9] H. Dong, T. Wang, J. Liu, and C. Zhang. Birds of a feather flock together: A close look at cooperation emergence via multi-agent RL. *CoRR*, abs/2104.11455, 2021.
- [10] X. Du, Y. Ye, P. Zhang, Y. Yang, M. Chen, and T. Wang. Situation-dependent causal influence-based cooperative multi-agent reinforcement learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 17362–17370, 2024.
- [11] I. Durugkar, E. Liebman, and P. Stone. Balancing individual preferences and shared objectives in multiagent reinforcement learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2505–2511, 2020.
- [12] T. Eccles, E. Hughes, J. Kramár, S. Wheelwright, and J. Z. Leibo. Learning reciprocity in complex sequential social dilemmas. *CoRR*, abs/1903.08082, 2019.
- [13] N. A. Grupen, B. Selman, and D. D. Lee. Cooperative multi-agent fairness and equivariant policies. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*, pages 9350–9359, 2022.
- [14] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. A. Duñez-Guzmán, A. G. Castañeda, I. Dunning, T. Zhu, K. R. McKee, R. Koster, H. Roff, and T. Graepel. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in Neural Information Processing Systems, NeurIPS 2018*, pages 3330–3340, 2018.
- [15] S. Iqbal and F. Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 2961–2970, 2019.
- [16] N. Jaques, A. Lazaridou, E. Hughes, Ç. Gülçehre, P. A. Ortega, D. Strouse, J. Z. Leibo, and N. de Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, pages 3040–3049, 2019.
- [17] J. Jiang and Z. Lu. Learning fairness in multi-agent systems. In *Advances in Neural Information Processing Systems, NeurIPS 2019*, pages 13854–13865, 2019.
- [18] P. Kollock. Social dilemmas: The anatomy of cooperation. *Annual review of sociology*, pages 183–214, 1998.
- [19] J. Z. Leibo, V. F. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and Multi-Agent Systems, AAMAS 2017*, pages 464–473, 2017.
- [20] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.
- [21] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems, NeurIPS 2017*, pages 6379–6390, 2017.
- [22] A. Lupu and D. Precup. Gifting in multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20*, pages 789–797, 2020.
- [23] H. Mao, W. Liu, J. Hao, J. Luo, D. Li, Z. Zhang, J. Wang, and Z. Xiao. Neighborhood cognition consistent multi-agent reinforcement learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 7219–7226, 2020.
- [24] K. R. McKee, I. Gemp, B. McWilliams, E. A. Duñez-Guzmán, E. Hughes, and J. Z. Leibo. Social diversity and social preferences in mixed-motive reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20*, pages 869–877, 2020.
- [25] D. Minarolli and B. Freisleben. Virtual machine resource allocation in cloud computing via multi-agent fuzzy control. In *2013 International Conference on Cloud and Green Computing*, pages 188–194, 2013.
- [26] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, pages 1928–1937, 2016.
- [27] A. Pozanco and D. Borrajo. Fairness in multi-agent planning. *CoRR*, abs/2212.00506, 2022.
- [28] U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multi-objective (deep) reinforcement learning with average and discounted rewards. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pages 8905–8915, 2020.
- [29] E. Tennant, S. Hailes, and M. Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023*, pages 317–325, 2023.
- [30] J. X. Wang, E. Hughes, C. Fernando, W. M. Czarnecki, E. A. Duñez-Guzmán, and J. Z. Leibo. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19*, pages 683–692, 2019.
- [31] W. Z. Wang, M. Beliaev, E. Biyik, D. A. Lazar, R. Pedarsani, and D. Sadigh. Emergent prosociality in multi-agent games through gifting. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021*, pages 434–442, 2021.
- [32] J. Yang, A. Li, M. Farajtabar, P. Sunehag, E. Hughes, and H. Zha. Learning to incentivize other learning agents. In *Advances in Neural Information Processing Systems, NeurIPS 2020*, 2020.
- [33] C. Zhang and J. A. Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems, NeurIPS 2014*, pages 2636–2644, 2014.
- [34] J. Zhang, Y. Zhang, X. S. Zhang, Y. Zang, and J. Cheng. Intrinsic action tendency consistency for cooperative multi-agent reinforcement learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 17600–17608, 2024.
- [35] M. Zimmer, C. Glanois, U. Siddique, and P. Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pages 12967–12978, 2021.