Resolving social dilemmas through reward transfer commitments

Richard Willis
King's College London
London, United Kingdom
richard.willis@kcl.ac.uk

Michael Luck
King's College London
London, United Kingdom
michael.luck@kcl.ac.uk

ABSTRACT

In file sharing networks, users can either act for personal gain by downloading files, or help the network by uploading files. Similar scenarios are important in many diverse situations where it does not pay to be nice to others. As a result, self-interested agents shirk collective behaviour, leading to poor outcomes for everyone. In response, in this paper we introduce a new metric for social dilemmas that quantifies the discrepancy between what is rational for individual agents and what is rational for the group, which we call the *exchange threshold*. It is the smallest fraction of individual rewards that need to be shared to make the unique Nash equilibrium a social welfare optimum. The exchange threshold differs to notions of altruism or prosocial behaviours because agents are transferring their rewards in order to induce others to care about their welfare. We investigate how the exchange threshold of a strategic game representing a *tragedy of the commons* varies with the abundance of resources to provide a deeper understanding of the underlying incentives.

KEYWORDS

Game Theory, Social Dilemma, Multiagent Reinforcement Learning

1 INTRODUCTION

Social dilemmas are situations in which individuals can either act selfishly for their personal gain, or act in a prosocial manner to help the collective, which delivers more benefits overall. The problem with social dilemmas is that it does not pay to be nice to others; individuals have incentives to act in ways that undermine socially optimal outcomes. As a result, self-interested agents will shirk collective behaviour, leading to poor outcomes for everyone. Although every agent prefers the outcome of mutual collective behaviour compared to mutual selfish behaviour, individually they are powerless to bring this about. This tension between collective and individual rationality is characterised by agents who engage in selfish behaviour outperforming agents engaged in collective, prosocial behaviour within a group, but prosocial groups outperforming selfish groups.

For example, suppose a group of farmers has access to an area of common land, and each farmer benefits from grazing their sheep on that land. The number of sheep that can be supported by this area is limited, however, and the land will degrade if overgrazed, permanently reducing the number of sheep it can support in the future. The collective does best when the land is used sustainably, but the individual farmers profit from grazing larger numbers of sheep. This particular social dilemma represents a common pool resource, and is known as a *tragedy of the commons*.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 9-10, 2023, Online, https://ala2023.github.io/. 2023.

Social dilemmas are important because they occur in many diverse situations, such as in file sharing networks where users can either act for personal gain by downloading files, or help the network by uploading files. Consequently, methods to resolve such dilemmas, achieved by causing all agents to prefer to take collective action, are important. Using an approach from game theory, we typically represent social dilemmas abstractly as games. We say that an agent *defects* when it acts selfishly, and that it *cooperates* when it acts to improve social outcomes. Many solutions to social dilemmas have been developed [1, 10, 11, 22], but they typically have specific requirements to be applicable, such as requiring agents to have preferences beyond the game rewards. In addition, solutions should require only a single agreement between participants, be able to scale to an arbitrary number of agents, and not require agents to have private knowledge of, or previous experience with, others. To our knowledge, only a potential application of trading in shares of future rewards [24] could meet these criteria.

In this paper, therefore, we propose a solution that resolves social dilemmas while addressing the issues highlighted above. The key principle underpinning our approach is that if we sufficiently align the individual and group incentives, then rational agents maximising their own rewards will also maximise the group rewards. To align individual incentives, we introduce the concept of an agent committing to share a proportion of its future game rewards with the other agents, a technique that we label *reward transfer*. By engaging in reward transfer, an agent incentivises the recipients to help it prosper which, paradoxically, can lead to a net profit for the transferring agent if it leads to a beneficial behavioural change in the recipients. We investigate two specific cases of reward transfer, firstly when a single agent gifts a proportion of its reward, and secondly when all agents exchange the same proportion of their reward with each other. In the former case, we are interested in determining the proportion of reward an agent should commit to sharing to maximise its resulting game reward. For the latter, we show that a sufficiently large amount of reward transfer from all agents is guaranteed to resolve a social dilemma, and we determine the minimum proportion that must be exchanged to do so. Concretely, we make the following contributions: we introduce a metric denoting the optimal proportion of future rewards an agent should unilaterally commit to sharing; we introduce a second metric quantifying the minimum proportion of future rewards that must be shared by all agents to resolve a dilemma; and we develop a technique to find these limiting values in games that are computationally intractable. The effectiveness of our solution is demonstrated with experiments and results using reinforcement learning agents in a stochastic game.

The paper is structured as follows. In Section 2 we review existing work on metrics and solutions applying to social dilemmas, and we formalise the notion of a social dilemma in Section 3. In Section 4 we

introduce our reward transfer concept, demonstrate it on normal-form games, and suggest a method to derive the limiting values in intractable stochastic games in Section /5. We provide experiments using our concept with an example social dilemma in Section 6 and present our results in Section 7. Finally, we conclude with a discussion in Section 8.

2 RELATED WORK

2.1 Metrics and Contracts

Some metrics, such as the price of anarchy or Pareto optimality [6], can be used to assess the quality of certain game outcomes, but say little about how such outcomes are achieved. Other metrics specify how difficult it is to achieve a certain outcome. For example, the *selfishness level* of a game measures the willingness of the players to cooperate [1]; it is the smallest fraction of the social welfare that needs to be offered to each player so that a social optimum is realised in a pure Nash equilibrium. The selfishness level both describes some aspect of a game and suggests a solution to achieving a social optimum by providing players with sufficient additional incentives.

In social dilemmas, the key question is how to ensure mutual cooperation. One option is to use *binding agreements*, or contracts, between players. Instead of a commitment to playing certain actions, Deng and Conitzer propose that a player could alternatively commit to avoid playing certain (possibly mixed) strategies, a method called *disarmament* [3, 4]. By removing all non-prosocial actions, or at least reducing the probability that an agent may play them, more socially beneficial outcomes can be achieved. While some social dilemmas cannot be resolved using disarmament, Deng and Conitzer introduce a negotiation protocol that can lead to improvements in the expected welfare of the participants.

Hughes et al. [10] demonstrate how extending a stochastic game to include a joint action contract protocol improves outcomes in social dilemmas. If contracts can include a *side payment*, then an agent can be paid to take an action that benefits another. Building on this, Christoffersten et al. [2] formalised voluntary contracts for stochastic games to include payments. While it isn't clear how much should be transferred, a fixed point about which players of a two-player game could negotiate the fair value of a joint action, called the Coco value, has been proposed [27].

An alternative to specifying payments for each particular action is to take a stake in the future outcomes of an agent through the trading of reward shares, as suggested by Schmid et al. [24]. Becoming invested in the future rewards of others promotes collective behaviour, and Schmid et al. demonstrate that this approach can lead to cooperation in an iterated game and conduct experiments on its applicability in a stochastic game.

2.2 Reward shaping

By default, an agent only cares about the rewards it receives in a game, but agent designers can include additional factors beyond the game reward to influence agent behaviour. For example, in reinforcement learning, *reward shaping* modifies a player's reward function during training to include auxiliary intrinsic rewards that encourage certain behaviours.

When an intrinsic reward is proportional to the mean collective reward [21], it is regarded as a *prosocial reward*, and has been shown to help reinforcement learning agents converge to a better equilibrium in the Stag Hunt social dilemma. McKee et al. [20] generate different degrees of prosocial rewards in mixed-motive games; they train populations with homogeneous and heterogeneous preferences and find some evidence that heterogeneous populations achieve more equal payoffs. Furthermore, they experimentally find good average prosocial reward coefficients for different environments. Alternatively, using reward shaping so that agents receive a negative intrinsic reward when they achieve lower than average rewards creates an incentive for the group to prefer egalitarian rewards [11]. This can be an effective approach in social dilemmas where agents have access to a punishment mechanism: when an agent benefits from selfish behaviour, the other agents receive lower than average rewards, and their intrinsic reward motivates them to punish the selfish agent. In this way, punishment helps to prevent agents from learning selfish behaviour.

According to Haeri [9], an agent can be considered to have a relationship with another specific agent if it has a personal reward term that depends upon the reward of the other agent. In the general case, a weighted graph can be used to specify the relationships between all agents; Haeri investigates how different relationship networks lead to differences in performance. Rather than specifying the form of the reward shaping by hand, they can be determined via an optimisation process, for example by using a genetic algorithm to optimise a reward network [31]. Gemp et al. develop an algorithm to reduce the Price of Anarchy through reward shaping (here called *loss-sharing*), [8]. During training, the players learn a loss sharing matrix that increases the social welfare of the worst case Nash equilibrium solution, leading to more cooperative strategies.

While reward shaping has been effective at promoting collective behaviour in social dilemmas, it requires agents to have intrinsic motivations. By adjusting behaviour to include factors beyond the game reward, intrinsic motivations may not be rational for a self-interested agent to include.

2.3 Strategic play

Repeated games and games involving a temporal component can permit cooperative strategies if they are enforceable via punishment [12]. In the famous example, *Tit-For-Tat* plays the action its partner chose in the previous round, and thus reciprocates cooperation while resisting exploitation from a defecting partner. Several researchers have approached cooperation with strategic play for multiagent reinforcement learning agents, in two main approaches. The first trains an agent with two policies, one for punishment, (defect) and the other for cooperation, and selects between them to punish defection and reciprocate cooperation during game play [15, 23]. The second permits a range of cooperative behaviour, such as mirroring the degree of *niceness* of an opponent, as determined by the change in the state value attributed to the actions of the opponent [5], or reciprocating the degree of cooperation displayed by the opponent as estimated by a classifier [32]. Similarly, if agents have the ability to detect violations of social norms, they are able to respond with sanctions [29].

While strategic play can add cooperative equilibria, because it only increases the range of strategies available to the players, it does not remove equilibria that lead to poor outcomes. Additionally,

it can be difficult to detect when an opponent is defecting, and other players still have an incentive to defect if they can avoid punishment. Strategic play is therefore difficult in settings with hidden information.

2.4 Opponent shaping

When training against fixed agents using learning algorithms, several methods have been developed that take into account the impact that an agent's actions have on the learning of their opponents. These algorithms guide their opponents towards beneficial behaviours, in an approach known as *opponent shaping* [7, 12, 16, 17, 22]. Because it considers the full lifetime deployment of the learning algorithms, rather than focusing on the performance in a single episode, it deals with the performance of the policies that learning algorithms converge to, in a similar manner to multi-armed bandit algorithms.

Instead of shaping opponent behaviour by choosing certain actions, if agents are able to transfer reward and observe the actions of their opponents, then they may opt to reward opponents when they take certain actions to encourage them to learn good behaviour. Explicitly adding an action that allows a player to gift reward to their peers has been investigated in a tragedy of the commons scenario [18] and in coordination games [33]. This is potentially more powerful than opponent shaping, because it can incentivise opponents to take actions for which they would not normally be able to receive reward. Yang et al. [36] explicitly optimise the transferred reward to shape the behaviour of opponents to improve the overall reward to the gifting agent.

2.5 Discussion

There are several approaches that have been taken to try to ensure cooperative behaviour, but there remain weaknesses and limitations. For example, while contracts provide assurance of behaviour, they can be burdensome to specify: Not only must a policy able to specify the contracts it would enter for all possible states, if it is costly to write a contract, then agents will incur a cost multiple times. While reward shaping has proven effective at learning cooperative policies in multiagent games, it often does not explain where the additional rewards come from, relying upon an innate disposition of the agents themselves. Such techniques may consequently be of limited value with a group of self-interested, independent agents. Strategic play and opponent shaping can be effective in social dilemmas, but because they construct additional equilibria, they do not exclude the original deficient equilibria, and thus they do not resolve the social dilemma. Additionally, shaping opponent behaviour by giving rewards dependent upon the joint action of all opponents scales poorly as the number of opponents increases. Additionally, some of these approaches are specific to either two-player games, or are for only one opponent that is not part of the team; indeed, it is an open question as to how they can be scaled to an arbitrary number of players.

3 SOCIAL DILEMMAS

We begin by formalising what it means for a game to be a social dilemma so that we are then able to demonstrate how our proposed solution directly addresses their core challenges. We believe that our formalisation, which focuses on prosocial choices, represents the underlying dynamics of these dilemmas. While others have defined social dilemmas previously, there remain limitations and implications. For example, Macy and Flache [19] consider them to be normal-form games with payoffs satisfying certain conditions, but some of their conditions are justified only if players hold particular beliefs. In turn, Hughes et al. [11] generalises the Macy and Flache definition to apply to Markov games, but we believe that their formulation is loose and includes games that could alternatively be characterised as coordination games.

In short, a social dilemma is a game in which all players have the option to take prosocial actions (to cooperate) at a (potential) cost to themselves. That is to say that while the benefit to their group is higher when players cooperate, the benefit to an individual player from cooperating is lower than when it defects. The players therefore face a choice between acting for the benefit of the group or in their own interests.

Now, in order to define a social dilemma, we need to introduce the notion of collective good, or *social welfare*, a metric indicating the benefit overall to a group. A social dilemma can therefore be defined as a situation in which, for all agents: (i) the social welfare of the group is strictly greater when an agent chooses to cooperate than when it chooses to defect, regardless of the actions of the other agents; and (ii) each agent does better individually when it defects than when it cooperates.

Consider a general-sum normal-form game played by $n \geq 2$ agents, where each agent faces a single choice to either cooperate, C, or defect, D. The reward function $R(\vec{a})$ returns a tuple of individual rewards $\vec{r} = (r_1, ..., r_n)$, where r_i is the reward received by player i, and depends only upon the action tuple $\vec{a} = (a_1, ..., a_n)$, where $a_i \in \{C, D\}$ is the action chosen by player i. For convenience, we also use R_i to represent the function that returns the reward for player i, and we write $\overrightarrow{a_{-i}}$ as the tuple of actions of all players other than player i. If we then denote our given social welfare metric as $SW(\vec{r})$, then we can define a social dilemma as follows.

$$\forall i \quad SW(R(\vec{a}|a_i = C)) > SW(R(\vec{a}|a_i = D)) \tag{1}$$

$$\forall i \quad R_i(\vec{a}|a_i = D) > R_i(\vec{a}|a_i = C) \tag{2}$$

A *partial* social dilemma can be said to occur when each agent might benefit from defecting, depending on what the other agents play. In this respect, there exists at least one combination of opponent actions that would lead to an agent preferring to defect. Thus, while there is always at least some personal cost to cooperation in a strict social dilemma, there is not always a personal cost to cooperation in a partial social dilemma. For a partial social dilemma, the second inequality above can therefore be softened as follows.

$$\forall i \;\; \exists \overrightarrow{a_{-i}} : R_i(\vec{a}|a_i = D, \overrightarrow{a_{-i}}) > R_i(\vec{a}|a_i = C, \overrightarrow{a_{-i}})$$

We also introduce the qualifier of a *weak* social dilemma when cooperation does not *strictly* increase the social welfare in all outcomes. Here it is sufficient that, for all agents, cooperation only sometimes leads to greater social welfare, and never reduces it. For a weak social dilemma, therefore, the first inequality can be replaced with the following.

$$\forall i \quad SW(R(\vec{a}|a_i = C)) \geq SW(R(\vec{a}|a_i = D))$$

$$\forall i \; \exists \overrightarrow{a_{-i}} : SW(R(\vec{a}|a_i = C, \overrightarrow{a_{-i}})) > SW(R(\vec{a}|a_i = D, \overrightarrow{a_{-i}}))$$

	C	D
C	3, 3	0, 4
D	4, 0	1, 1

(a) Prisoner's Dilemma

	C	D
C	3, 3	1, 4
D	4, 1	0, 0

(b) Chicken

Table 1: Social Dilemmas

	C	D
C	2, 4	0, 4
D	$\frac{8}{3}, \frac{4}{3}$	$\frac{2}{3}, \frac{4}{3}$

(a) Prisoner's Dilemma with $g_r = \frac{1}{3}$

	C	D
C	$\frac{3}{2}, \frac{9}{2}$	$\frac{1}{2}, \frac{9}{2}$
D	2, 3	0, 0

(b) Chicken with $g_r = \frac{1}{2}$

Table 2: Limiting values for reward gifting

In this paper for social welfare, we use the utilitarian metric, U, which measures the unweighted sum of rewards obtained by all players, as follows.

$$U(\vec{r}) = \sum_{i=1}^{n} r_i \tag{3}$$

Prisoner's Dilemma (Table 1a) is an example of a strict social dilemma, while Chicken (Table 1b) is an example of a partial social dilemma. We say that a social dilemma is *resolved* if all the Nash equilibria of the game maximise the social welfare.

A strict social dilemma is characterised by a unique, pure Nash equilibrium that is Pareto inferior. This follows from the fact that defect is a dominant strategy for every agent, and the social welfare strictly decreases with each defection.

4 REWARD TRANSFER

As discussed above, the core difficulty of social dilemmas is that prosocial actions are costly, unless agents are sufficiently motivated to care about others, when collective action becomes more attractive than selfish behaviours. To address this, an agent needs to provide some kind of incentive for others to care about its own well-being, and it can do so by sharing the reward it gains from their combined actions. Thus, in this section, we introduce a mechanism by which an agent can transfer a proportion of its reward to other agents.

4.1 Reward gifting

Consider a builder who has been awarded a construction contract, but lacks the ability to complete the project alone, so offers to share their income with another builder. Even though the first builder must transfer some of their payments to the second, both builders benefit. We refer to this possibility for an agent to donate a proportion of its rewards to others in a group as *reward gifting* and, for simplicity, assume that such a gift is split equally among the members of the group.

Formally, if an agent i gifts a proportion $g_i \in [0, 1]$ of its own reward to others, with the gifts of all agents denoted as $\vec{g} = (g_1, ..., g_n)$, then the *post-transfer reward* to agent i, r'_i, is the remaining part of the individual reward plus any gifts received from others, and is given by:

$$r'_i(\vec{r}, \vec{g}) \quad = \quad (1 - g_i)r_i + \frac{1}{n-1}\sum_{j \neq i} g_j r_j \tag{4}$$

This formulation is similar to the concept of *prosocial* rewards of Peysakhovich and Lerer [21], a form of reward shaping in which an agent modifies only its own reward by including an extra intrinsic reward term. However, our notion of reward gifting, involves only a redistribution of the *extrinsic* game rewards, and allows an agent to impact the rewards of other agents (in addition to its own reward).

Consider again the Prisoner's Dilemma of Table 1a, where game theory tells us that rational players will choose to defect, leading to a reward of 1 to both players. Suppose now that before playing the game, the row player commits to sharing a third of its reward with the column player, so $g_r = \frac{1}{3}$, as shown in Table 2a. The row player's individual reward is now $\frac{2}{3}$ of the original game reward, which does not change their preferred outcome, so defect remains a dominant strategy. However, because the column player now receives a proportion of row's reward, they are incentivised to cooperate to increase row's reward, which increases the gifted reward they receive.

When row gifts $\frac{1}{3}$ of its reward, column is ambivalent between cooperate and defect, but for any greater value, column strictly prefers to cooperate because the received gifted reward outweighs the benefit of defecting. As long as row gifts less than $\frac{3}{4}$, they receive more reward from the resulting outcome of (D, C) compared to (D, D) without reward gifting.

Note that not all social dilemmas afford the opportunity for an agent to benefit from unilateral reward gifting. For example, with Chicken, in order to make column weakly prefer to cooperate, row must gift $g = \frac{1}{2}$ (Table 2b). It isn't clear whether row benefits from this, because although row can now expect column to cooperate, and consequently row can defect and earn a reward of 2 in the outcome (D, C), there is no dominant strategy in Chicken for either player and therefore no particular expected outcome. It is also possible to construct a social dilemma without opportunities to benefit from reward gifting.

We have seen that a player can sometimes increase its overall reward by engaging in unilateral reward gifting. We define the *gifting optimum*, denoted $g^* \in [0, 1]$, for symmetric, non-negative social dilemmas, as the amount of unilateral reward gifting that leads to the highest reward for the gifting player. If there is no opportunity to benefit, then $g^* = 0$.

4.2 Reward exchange

Prisoner's Dilemma cannot be resolved with only one player gifting reward, because defecting remains dominant for the gifting agent. In order to change the situation so that mutual cooperation is dominant, both agents must engage in reward gifting. When all players gift the same proportion of their rewards, $e \in [0, 1]$, so that $\forall i \; g_i = e$, we call it *reward exchange*. Equation 4 then reduces to:

$$r'_i(\vec{r}, e) \quad = \quad (1 - e)r_i + \frac{e}{n-1}\sum_{j \neq i} r_j \tag{5}$$

Consider again Prisoner's Dilemma, in which the players agree to exchange $e = \frac{1}{4}$ of their payoffs, as shown in Table 3a. Cooperation is weakly dominant for both players, so the social dilemma is resolved. However, for any values of $e < \frac{1}{4}$, the dilemma would

	C	D
C	3, 3	1, 3
D	3, 1	1, 1

(a) Prisoner's Dilemma with $e = \frac{1}{4}$

	C	D
C	3, 3	2, 3
D	3, 2	0, 0

(b) Chicken with $e = \frac{1}{3}$

Table 3: Exchange threshold for PD and Chicken

not be resolved as both players would prefer to defect; we call this limiting value the *exchange threshold* of a game, denoted e^*. In the context of the tragedy of the commons, this is the smallest proportion of reward that the farmers must share such that no farmer has an incentive to graze more sheep than the land can support, and consequently the land is used sustainably. The exchange threshold for Chicken is $e^* = \frac{1}{3}$ and is shown in Table 3b.

Formally, the exchange threshold, of a game is the minimum reward exchange proportion e that resolves the social dilemma. If $\mathbb{NE}$ denotes the set of Nash equilibria:

$$e^* = \min_e : \mathbb{NE} = \{\forall i \ a_i = C\} \tag{6}$$

The exchange threshold of a game is invariant to scalar multiplication of the game rewards and exists for all social dilemmas using the utilitarian metric, which follows from the fact that when every player exchanges $\frac{n-1}{n}$ of their rewards, Equation 5 reduces to:

$$r'_i(\vec{r}, \frac{n-1}{n}) = \frac{1}{n}\sum_{j=0}^{n} r_j = \frac{1}{n}U(\vec{r}) \tag{7}$$

Consequently, each player wants to maximise the utilitarian metric, and cooperation increases the social welfare, so there is a unique social welfare maximum of all cooperate. We therefore expect that, everything else being equal, increasing the number of players of a game will increase its exchange threshold.

The exchange threshold of a game can be viewed as a measure of the willingness of the players to cooperate; it quantifies the disparity between the individual and group incentives. A low exchange threshold indicates that the players need to care little about each other to achieve a socially optimal outcome, whereas a high exchange threshold suggests that the players have strong incentives to shirk prosocial behaviour. Importantly, if we can determine the exchange threshold, we may also be able to find a way to resolve social dilemmas, through agents exchanging the proportion of their rewards specified in this way.

5 COMPLEX GAMES AS SOCIAL DILEMMAS

While normal-form games are effective for modelling some kinds of dilemmas, others are more complex, as our running example shows. First, the farmers are not restricted to either cooperate or defect, and can take a range of actions, such as adding or removing sheep from the land. Second, the impact of their actions plays out over time rather than being immediately apparent, and depends on circumstances, such that adding one sheep to an empty field has different consequences to adding an extra sheep to an overcrowded field. We therefore generalise our representation of social dilemmas to include games with: (i) a temporal component, (ii) a larger action space and (iii) a larger state space. We do this by extending our definition of a social dilemma to apply to Markov games. Then, in order to determine the gifting optimum and exchange threshold of a Markov game, we describe a technique that uses multiagent reinforcement learning.

5.1 Markov social dilemmas

A Markov game is one with a finite set of states $\mathbb{S}$, played by n players. The game is parameterised by sets of available actions for each player $\mathbb{A} = \{\mathbb{A}_1, ..., \mathbb{A}_n\}$, and a stochastic transition function $T : \mathbb{S} \times \mathbb{A} \rightarrow \Delta(\mathbb{S})$, mapping from joint actions at each state to a discrete probability distribution over states. We say that a Markov game is symmetric if, for any joint policy, the rewards to each player are unchanged by permutation of their indices.

Now, to extend the definition of social dilemmas to Markov games, we use *empirical game theoretic analysis* [28, 30, 34, 35] to reduce a Markov game to normal form. Here, the players are restricted to a choice between a policy representing cooperate or a policy representing defect. The players receive the total reward obtained by their chosen policy averaged over a large number of game roll-outs. A Markov game can then be defined as a social dilemma if the expected player rewards obtained by the combination of policies satisfies the inequalities in Section 3. Formally, an n-player Markov social dilemma is a tuple of a Markov game M and two disjoint sets of policies, that implement cooperate and defect respectively, with expected rewards satisfying inequalities 1 and 2.

5.2 Gifting optima and exchange thresholds of Markov social dilemmas

Recall that a social dilemma is resolved if all the Nash equilibria of the game maximise the social welfare. However, it is generally not tractable to compute the equilibria of a Markov game, so we modify our notion and say that a Markov social dilemma is resolved when the equilibrium joint policy found by an optimisation technique maximises the social welfare.

Learning algorithms such as multiagent reinforcement learning (MARL) can be used to find equilibrium policies. However, independent reinforcement learning algorithms in MARL treat opponents as static and do not take account of their ability to respond via their own policy updates. This typically results in so-called *naive learners* falling into suboptimal equilibria [17]. In response, recent work has addressed the challenge of learning collective action policies in social dilemmas [11, 14, 22, 23, 32], with many applying techniques from game theory to increase cooperation between independent learners, leading to increases in performance.

In our solution, each agent i independently learns a policy $\pi_i(a_i|s)$, with state $s \in \mathbb{S}$, to maximise the long-term γ-discounted sum of their individual reward given by their reward function $R_i(\vec{a}, s)$, which depends upon the joint action and state:

$$V(s_0) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t R_i(\vec{a_t}, s_t | \vec{a_t} \approx \vec{\pi})]$$

where $\vec{\pi}$ is the joint policy of all players. When training converges, the policies are an approximate best response to their opponents. When using MARL with reward transfer, we find policies to maximise the post-transfer reward r', given by Equations 4 or 5.

Despite the issue above of falling into suboptimal equilibria, for a symmetrical game MARL should learn policies to maximise the

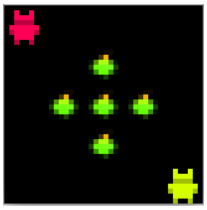


Figure 1: Commons Harvest

utilitarian metric (see Equation 7) when we set $e = \frac{n-1}{n}$. However, it may be possible to use smaller values of e and still find equilibria that achieve the same utilitarian metric. We estimate the *exchange threshold* (Equation 6), e^*, of a Markov social dilemma by finding the smallest value of reward exchange, e, such that the equilibrium joint policy given e, $\overrightarrow{\pi^*_e}$, found by independent learning algorithms achieves the same utilitarian metric value as the equilibrium joint policy explicitly trained to maximise the utilitarian reward.

$$e^* = \min_{0 \leq e \leq \frac{n-1}{n}} : \sum_i r'_i(R_i(\cdot|\overrightarrow{\pi^*_e}), e) = \sum_i r'_i(R_i(\cdot|\overrightarrow{\pi^*_{\frac{n-1}{n}}}), \frac{n-1}{n})$$

Our approach is to train policies with $e = \frac{n-1}{n}$, so that each player maximises the social welfare, and then we use an optimisation method to find the smallest value of e that leads to joint policies that achieve this same social welfare value. For the gifting optimum, we are only interested in the post-transfer reward obtained by the gifting agent, and the task is to estimate the value of $0 \leq g \leq 1$ that maximises this. Note that while finding optimal joint policies likely increases in complexity as the number of players increases, because we only need to sweep over the possible values of g or e, the task of finding these limiting reward transfer values does not increase in complexity. This contrasts with methods that transfer reward conditional upon opponent actions.

6 EXPERIMENTS

In this section, we demonstrate the techniques of Subsection 5.2 with MARL to find the gifting optimum and exchange threshold of a complex game, and show that sufficient reward exchange leads independent learners to converge to a social welfare optimum.

6.1 Commons Harvest

Commons Harvest is a sequential social dilemma [14] describing a group of agents harvesting a supply of apples. It represents a tragedy of the commons, in which the agents should harvest sustainably to avoid depleting the environment. We have configured a simplified version, illustrated in Figure 1 (with 5×5 squares and two players), using the Melting Pot library [13].

Two agents move around the grid-world environment and receive rewards for harvesting apples. At each timestep the agents can move one square (north, south, east, or west), or they may remain stationary. If an agent moves into a square containing an apple, it collects the apple and receives a reward of 1, but cannot receive further reward in the square until the apple regrows. Harvested apples regrow with a fixed probability if they are horizontally or vertically adjacent to an uncollected apple. In this way, as long as at least one apple remains, all apples will regrow given enough time. An apple may regrow if an agent is on its square, in which case the agent automatically collects the apple. Commons Harvest terminates after 500 timesteps.

In order to harvest the largest number of apples, the agents need to maximise the number of apples that regrow. If the central apple is harvested, then only the central apple can regrow in the next timestep, assuming at least one peripheral apple remains. If the central apple is present, however, all the peripheral apples are eligible to regrow. Therefore, to maximise the number of harvested apples, the central apple should be left, and peripheral apples collected as quickly as possible.

6.2 Training

Deep reinforcement learning, in particular *proximal policy optimisation* [26] with entropy regularisation and *generalised advantage estimation* [25], was used to train policy pairs to play Commons Harvest with different values of reward gifting and reward exchange. In order to experimentally determine the gifting optimum and exchange threshold, we used a grid search over unilateral reward gifting $g \in [0, 0.1, ..., 1]$ and reward exchange values $e \in [0, 0.05, ..., 0.5]$. Note that for reward exchange it is sufficient to incrementally decrease e from the maximum value until the social welfare falls, but we trained over the full range to better understand the dynamics of the environment. For each value of e and g, pairs of agents were trained for 4000 episodes, and evaluated over 30 episodes, for five different seeds. During evaluation, each agent took its maximum likelihood action, rather than sampling from its action probability distribution as in training.

To gain an understanding of the characteristics influencing the cooperative incentives, we varied the features of the game and analysed how the gifting optimum and exchange threshold changed in response: we repeated each experiment with different probabilities of apples regrowing, $G \in \{0.05, 0.15, 0.45\}$, to investigate how the abundance of resources impacts these values. For high reward exchange values, we encountered behaviour where one of the agents would learn to sit in a corner instead of harvesting apples. This is known as the *lazy agent* problem, often occurring in cooperative games. With $e = 0.5$, it is clear that a joint policy including a lazy agent is not optimal because both agents could achieve higher rewards if both harvested.

It is important that we find optimal joint strategies, otherwise we are instead measuring the deficiencies of the deployed learning algorithms rather than an intrinsic property of the environment. To avoid the lazy agent problem, we pre-trained the agents with *self-play* for 2000 episodes in the reward exchange experiments, so that each agent separately learns to harvest apples. However, with reward gifting, each agent prefers to collect the apples themselves, so no pre-training was needed.

The code for the training can be found at https://github.com/Muff2n/meltingpot/.

7 RESULTS

We observe that the policies in a two-player game of Commons Harvest typically converge to one of the following distinct behaviours.

Table 4: Approximate social welfare of behaviours for different apple regrowth probabilities

Behaviour	0.05	0.15	0.45
Unsustainable harvesting	5	N/A	N/A
Centre camping	25	65	155
Cooperative harvesting	90	220	425

Unsustainable harvesting Both agents harvest all the apples immediately.

Centre camping Both agents rush to the central apple. One agent reaches the centre first (with probability 0.5, due to the mechanics of the environment) and then *camps* at the centre by not moving. The other agent continually tries to move onto the centre square, but fails because it is occupied.[1] This behaviour leads to unequal episode rewards, because the agent that gains the central square harvests the central apple when it periodically regrows, while the other agent receives no rewards.

Cooperative harvesting The agents harvest the peripheral apples together.

The values of the utilitarian metric associated with these behaviours for different probabilities of apple regrowth is detailed in Table 4.

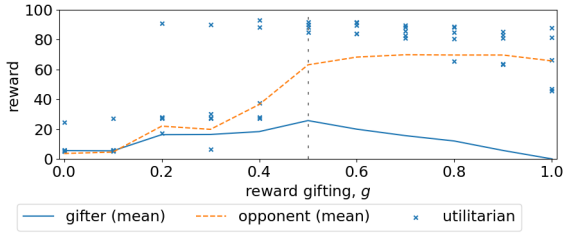
7.1 Reward gifting

The results of the reward gifting experiments are shown in Figure 2. The dashed vertical line is the *gifting optimum*, chosen to be the argmax of reward to the gifting player. For regrowth probability $G = 0.05$, we observe that as the gifting proportion increases, the overall behaviour transitions from unsustainable harvesting, through centre camping, to cooperative harvesting. Once cooperative harvesting is induced, there is no reason to gift a greater fraction of reward. At the gifting optimum, the gifting player has improved their rewards from ≈ 3 to ≈ 25, while the receiving agent achieves ≈ 65. This happens because the gifting agent harvests around 50 apples, but must share $g^* = 0.5$ of them with the receiver, who harvests around 40 apples. $G = 0.15$ starts with centre camping and transitions to cooperative harvesting. $G = 0.45$ never learns cooperative harvesting behaviour, however, and the gifting player does not increase its reward at any value attempted, so the gifting optimum is $g^* = 0$. Although the gifting player received larger average rewards when $g = 0.1$, this results from the stochasticity of training and evaluation.[2]

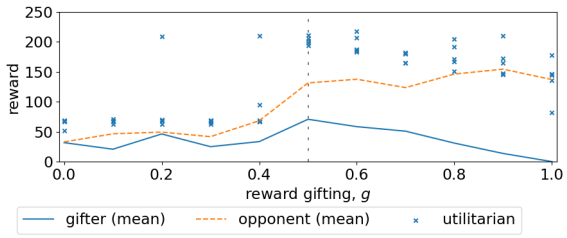
7.2 Reward exchange

Reward exchange results are shown in Figure 3 in which the dashed vertical line is the *exchange threshold*, chosen to be the smallest value of e that produces the same social welfare as $e = \frac{1}{2}$. To assess the harvesting distribution, we also plot the mean Gini value of the pre-transfer rewards. For $G = 0.05$ and $G = 0.15$ we observe the same sharp transition between centre camping and cooperative harvesting as with reward gifting. For $G = 0.45$, we now observe cooperative harvesting, but there is a range of cooperative behaviours, as shown by a number of policies returning a social welfare between centre camping and cooperative harvesting.

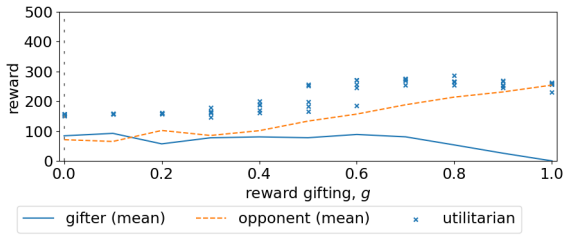
We expected increasing the scarcity of resources by lowering the apple regrowth probability to lead to higher reward gifting and exchange thresholds, because a scarcity of resources typically leads to greater conflict. The exchange threshold indeed decreases when the regrowth probability increases from 0.05 to 0.15, but is highest for $G = 0.45$. For the lower regrowth probabilities, the learned policy pairs jump from centre camping (a selfish behaviour) to cooperative harvesting (a fully cooperative behaviour), with few policy pairs returning social welfare values between these two equilibria (Table 4). For $G = 0.45$, the impact of harvesting the central apple is smaller, because it regrows faster. It may be that for environments where apple regrowth is low, cooperative policies avoid harvesting the central apple due to the high cost, but when apples regrow sufficiently quickly, a range of cooperative behaviours are possible, with the propensity for an agent to harvest the apple decreasing as e increases.



(a) Apple regrowth probability of 0.05



(b) Apple regrowth probability of 0.15

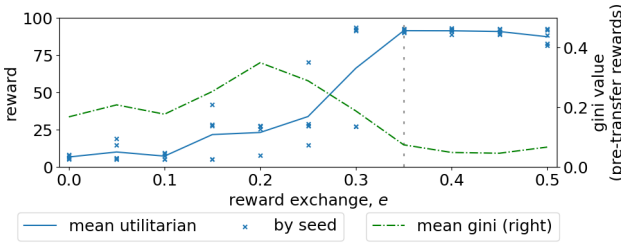


(c) Apple regrowth probability of 0.45

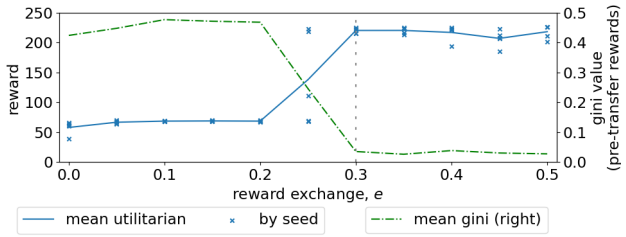
Figure 2: Reward gifting

[1] This behaviour is learned because agents sample actions stochastically during training. The agent at the centre has some probability of moving away, in which case the second agent will gain the centre, and the process continues with reversed roles.

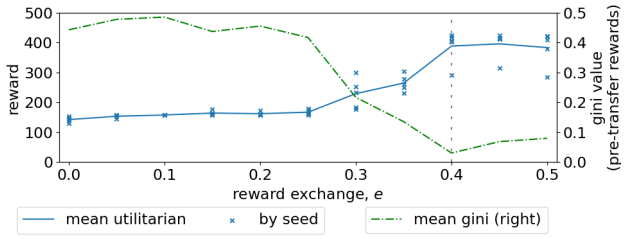
[2] For a two-player non-negative game, the recipient should never obtain lower rewards when receiving a higher proportion of gifted reward.



(a) Apple regrowth probability of 0.05



(b) Apple regrowth probability of 0.15



(c) Apple regrowth probability of 0.45

Figure 3: Reward exchange

7.3 Discussion

As mentioned above, pre-training the agents with self-play was needed to alleviate the lazy agent problem, as illustrated in Figure 4, where high values of e produced lower social welfare and a more unequal reward distribution, because one of the agents avoided harvesting. However, we note that at the exchange threshold of the game, the lazy agent problem does not manifest, because each agent has retained sufficient incentive to harvest apples, and the social welfare is maximised without requiring pre-training. Therefore, an additional benefit of the training with reward exchange equal to the exchange threshold may be that it is easier to train optimal agents.

A key question here is whether our configuration of Commons Harvest satisfies the Markov social dilemma constraints introduced in (Sections 3 and 5.1). In order to reduce Commons Harvest to a normal-form game, the players are restricted to a binary choice between deploying a learning algorithm without reward gifting, or using one that gifts a proportion of its reward equal to the gifting optimum of the game. Choosing the non-gifting policy represents defect and choosing the gifting policy represents cooperation. If we take $G = 0.05$, where $g^* = 0.5$, then we get the payoff matrix shown in Table 5. This game is a weak partial social dilemma (the social welfare does not strictly increase as the number of cooperators increases and cooperation is not always costly).

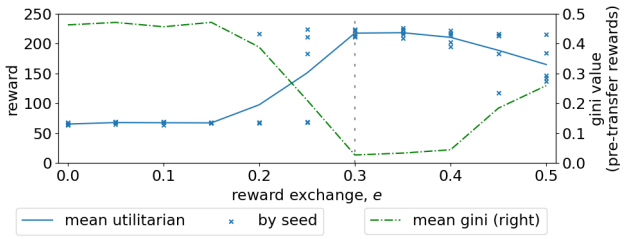


Figure 4: Training from scratch, $G = 0.15$

Table 5: Approximate empirical game of Commons Harvest

	$g_c = 0.5$	$g_c = 0$
$g_r = 0.5$	45,45	25,65
$g_r = 0$	65,25	3,3

8 CONCLUSION

In this paper we have introduced: the notion of an exchange threshold of a social dilemma, a metric to measure the alignment between individual and group incentives; and the gifting optimum, which measures the optimal proportion of reward an agent should offer to unilaterally transfer to other agents. We described a method to determine these metrics for normal-form and Markov games, and demonstrated it experimentally, analysing how the metrics vary with respect to resource scarcity. If the players of a social dilemma can mutually commit to transfer at least the exchange threshold of a game, then they can resolve any social dilemma.

While this work is a similar mechanism to that of Schmid et al. [24], our approach involves players choosing to donate their rewards rather than acquiring an interest in the rewards of the other players. Perhaps more importantly, we additionally provide a consideration of the limiting values of reward transfer that are required to resolve a social dilemma. This provides our work with a descriptive element; if reward transfer is expensive, it minimises costs by identifying the smallest amount to be transferred. We also consider the unilateral case, where it may be individually rational (not requiring collective action) to donate a share of reward. From a descriptive perspective, the work of Apt and Schafer [1] is closest, because the selfishness level of a game is an equivalent metric for describing the alignment between individual and group rewards, but as a normative method it does not explain where the additional reward that incentivises collective action originates.

ACKNOWLEDGMENTS

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org) and a BT/EPSRC funded iCASE Studentship [grant number EP/T517380/1].

REFERENCES

[1] K. R. Apt and G. Schaefer. 2014. Selfishness Level of Strategic Games. *Journal of Artificial Intelligence Research* 49 (Feb. 2014), 207–240. https://doi.org/10.1613/jair.4164

[2] Phillip J. K. Christoffersen, Andreas A. Haupt, and Dylan Hadfield-Menell. 2023. Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL. arXiv:arXiv:2208.10469

[3] Yuan Deng and Vincent Conitzer. 2017. Disarmament Games. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI Press, 473–479.

[4] Yuan Deng and Vincent Conitzer. 2018. Disarmament Games with Resources. In *Proceedings of the Thirty-Second {AAAI} Conference on Artificial Intelligence*. AAAI Press, 981–988.

[5] Tom Eccles, Edward Hughes, János Kramár, Steven Wheelwright, and Joel Z Leibo. 2019. The Imitation Game: Learned Reciprocity in Markov Games. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1934–1936.

[6] Edith Elkind, Angelo Fanelli, and Michele Flammini. 2020. Price of Pareto Optimality in Hedonic Games. *Artificial Intelligence* 288 (Nov. 2020), 103357. https://doi.org/10.1016/j.artint.2020.103357

[7] Jakob Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. 2018. Learning with Opponent-Learning Awareness. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 122–130.

[8] Ian Gemp, Kevin R. McKee, Richard Everett, Edgar A. Duéñez-Guzmán, Yoram Bachrach, David Balduzzi, and Andrea Tacchetti. 2022. D3C: Reducing the Price of Anarchy in Multi-Agent Learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 498–506. arXiv:2010.00575

[9] Hossein Haeri. 2021. Reward-Sharing Relational Networks in Multi-Agent Reinforcement Learning as a Framework for Emergent Behavior. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*. ACM, 1808–1810.

[10] Edward Hughes, Thomas W Anthony, Tom Eccles, Joel Z Leibo, David Balduzzi, and Yoram Bachrach. 2020. Learning to Resolve Alliance Dilemmas in Many-Player Zero-Sum Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 538–547.

[11] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, Heather Roff, and Thore Graepel. 2018. Inequity Aversion Improves Cooperation in Intertemporal Social Dilemmas. In *32nd Conference on Neural Information Processing Systems*. Curran Associates, Inc., 3330–3340.

[12] Alexis Jacq, Julien Perolat, Matthieu Geist, and Olivier Pietquin. 2020. Foolproof Cooperative Learning. arXiv:arXiv:1906.09831

[13] Joel Z. Leibo, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, John P. Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles 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 Machine Learning*, Vol. 139. PMLR, 6187–6199. arXiv:2107.06857

[14] Joel Z Leibo, Vinicius Zambaldi, and Marc Lanctot. 2017. Multi-Agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems*. ACM, 464–473.

[15] Adam Lerer and Alexander Peysakhovich. 2018. Maintaining Cooperation in Complex Social Dilemmas Using Deep Reinforcement Learning. arXiv:arXiv:1707.01068

[16] Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. 2019. Stable Opponent Shaping in Differentiable Games. In *ICLR*. OpenReview.net. arXiv:1811.08469

[17] Chris Lu, Timon Willi, Christian Schroeder de Witt, and Jakob Foerster. 2022. Model-Free Opponent Shaping. In *Proceedings of the 39 Th International Conference on Machine Learning*. PMLR, 14398–14411. arXiv:2205.01447 [cs]

[18] Andrei Lupu and Doina Precup. 2020. Gifting in Multi-Agent Reinforcement Learning. In *New Zealand*. 9.

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

[20] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social Diversity and Social Preferences in Mixed-Motive Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 869–877.

[21] Alexander Peysakhovich and Adam Lerer. 2017. Prosocial Learning Agents Solve Generalized Stag Hunts Better than Selfish Ones. arXiv:arXiv:1709.02865

[22] Alexander Peysakhovich and Adam Lerer. 2018. Consequentialist Conditional Cooperation in Social Dilemmas with Imperfect Information. In *6th International Conference on Learning Representations*. OpenReview.net. arXiv:1710.06975 [cs]

[23] Alexander Peysakhovich and Adam Lerer. 2018. Towards AI That Can Solve Social Dilemmas. *Association for the Advancement of Artificial Intelligence* (2018), 7.

[24] Kyrill Schmid, Michael Kölle, and Tim Matheis. 2022. Learning to Participate through Trading of Reward Shares. In *Proceedings of the Adaptive and Learning Agents Workshop*.

[25] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2016. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *Proceedings of the 4th International Conference on Learning Representations*. arXiv:1506.02438 [cs]

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. arXiv:arXiv:1707.06347

[27] Eric Sodomka, Elizabeth M Hilliard, Michael L Littman, and Amy Greenwald. 2013. Coco-Q: Learning in Stochastic Games with Side Payments. In *Proceedings of the 30th International Conference on Machine Learning*. JMLR.org, 1471–1479.

[28] Karl Tuyls, Julien Perolat, Marc Lanctot, Edward Hughes, Richard Everett, Joel Z. Leibo, Csaba Szepesvári, and Thore Graepel. 2020. Bounds and Dynamics for Empirical Game Theoretic Analysis. *Autonomous Agents and Multi-Agent Systems* 34, 1 (April 2020), 7. https://doi.org/10.1007/s10458-019-09432-y

[29] Eugene Vinitsky, Raphael Köster, John P. Agapiou, Edgar Duéñez-Guzmán, Alexander Sasha Vezhnevets, and Joel Z. Leibo. 2022. A Learning Agent That Acquires Social Norms from Public Sanctions in Decentralized Multi-Agent Settings. arXiv:arXiv:2106.09012

[30] William E Walsh, Rajarshi Das, Gerald Tesauro, and Jeffrey O Kephart. 2002. Analyzing Complex Strategic Interactions in Multi-Agent Systems. *AAAI Technical Report WS-02-06* (June 2002).

[31] Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. 2019. Evolving Intrinsic Motivations for Altruistic Behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 683–692.

[32] Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. 2019. Achieving Cooperation through Deep Multiagent Reinforcement Learning in Sequential Prisoner's Dilemmas. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*. ACM, Beijing China, 11:1–11:7. https://doi.org/10.1145/3356464.3357712

[33] Woodrow Z. Wang, Mark Beliaev, Erdem Bıyık, Daniel A. Lazar, Ramtin Pedarsani, and Dorsa Sadigh. 2021. Emergent Prosociality in Multi-Agent Games Through Gifting. arXiv:arXiv:2105.06593

[34] Yongzhao Wang, Qiurui Ma, and Michael P Wellman. 2022. Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis. In *21st International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Auckland, 1346–1354.

[35] Michael P Wellman. 2006. Methods for Empirical Game-Theoretic Analysis. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*. AAAI Press, 1552–1556.

[36] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to Incentivize Other Learning Agents. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 15208–15219.