Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out

Chin-wing Leung
University of Warwick
Coventry, United Kingdom
chin-wing.leung@warwick.ac.uk

Paolo Turrini
University of Warwick
Coventry, United Kingdom
p.turrini@warwick.ac.uk

ABSTRACT

We study populations of self-interested agents playing a 2-person repeated Prisoner's Dilemma game, with each player having the option of opting out of the interaction and choosing to be randomly assigned to another partner instead. The partner selection component makes these games akin to random matching, where defection is known to take over the entire population. Results in the literature have shown that, when forcing agents to obey a set partner selection rule known as Out-for-Tat, where defectors are systematically being broken ties with, cooperation can be sustained in the long run. In this paper, we remove this assumption and study agents that learn both action- and partner-selection strategies. Through multiagent reinforcement learning, we show that cooperation can be sustained without forcing agents to play predetermined strategies. Our simulations show that agents are capable of learning in-game strategies by themselves, such as Tit-for-Tat. What is more, they are also able to simultaneously discover cooperation-sustaining partner selection rules, notably Out-for-Tat, as well as other new rules that make cooperation prevail.

KEYWORDS

Social Dilemmas, Partner Selection, Emergence of Cooperation.

ACM Reference Format:
Chin-wing Leung and Paolo Turrini. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

Understanding the conditions under which cooperation emerges in social dilemmas is a great challenge for many disciplines, including economics, evolutionary psychology and biology [14]. Over the past 50 years, the computational study of social interaction revealed a deep connection between population dynamics and reinforcement learning [7], with simple learning algorithms shown to correspond to complex evolutionary dynamics [6].

In social dilemmas of cooperation, agents choose whether or not to pay a cost to contribute to a collective project and rip the rewards of each contribution given. In well-mixed populations, where individuals are paired uniformly at random, defectors quickly become the dominant type, and mechanisms need to be put in place to make sure this does not happen [14]. As neatly argued by Martin Nowak in his seminal paper on the mechanisms that sustain cooperation in a world that would otherwise be dominated by mass defection, partner selection is a key enabler [14]. When individuals are able to choose who to play with, signalling cooperative intentions is paramount for the emergence of cooperation clusters that resist defectors and promote socially desirable behaviour [4, 5, 23].

Understanding the equilibria of social dilemmas with partner selection is a non-trivial task, as different layers of strategies coevolve, potentially at different timescales. The level of complexity makes it unfeasible to use the standard tools of evolutionary game theory, such as ODE-based replicator dynamics. Luckily, multiagent reinforcement learning can come to our rescue, providing an alternative toolbox to study the convergence of population dynamics through agent-based simulations, handling large environments and complex interactions [6].

A recent contribution using multi-agent reinforcement learning has shown that individuals being able to select their partner for a Prisoner's Dilemma game, while only able to observe their potential partner's last played move, is a sufficient condition for the emergence of cooperation [2]. Surprisingly, agents were able to learn interaction patterns that behave as the cooperation-sustaining strategies in the infinitely repeated game, notably Tit-for-Tat, a strategy that copies the last move played by their partner [3]. The same contribution showed that when this form of "active" partner selection is replaced by random rewiring, mass defection rapidly takes over. While being able to actively choose one's partners is an important mechanism, it comes with various restrictive assumptions, notably the possibility of unilaterally and unrestrictedly choosing any partner (without mutual consent) and the possibility of observing the past decisions of the entire population. On the other hand, sacrificing these assumptions is likely to be detrimental to the evolution of cooperation.

In social dilemmas with the option of opting out (SDOOs), participants are matched in pairs and each pair plays a Prisoner's Dilemma game repeatedly. At each round, each player can unilaterally break ties with the current partner and be randomly paired with another available partner, with whom they play a Prisoner's Dilemma game in the next round. SDOOs are just one step away from full random rewiring, introducing perhaps the faintest form of partner selection one can think of, with agents accepting *any* partner in response to the observed behaviour of the current one. Previous experimental research has shown that humans do display cooperation levels in SDOOs [32]. Further analytical studies [33]

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

have provided a simplified form of replicator dynamics that explain that cooperation can emerge in SDOOs, but these are based on the assumption that players play according to a specific partner selection rule, Out-for-Tat, where agents who experience defection unavoidably break ties with their partners. It is not yet known what happens in SDOOs where partner selection rules are not forced, but emerge from interaction instead. In light of what is known in the literature for fully random matching, showing even minimal signs of cooperation emerging in such systems, without forcing individuals to adhere to a given rule, would be a significant and surprising achievement.

Contribution. In this paper, we study learning agents playing a repeated Prisoner's Dilemma, having only access to the last action of their assigned opponent, and with the option to decide, at each time point, to end the current interaction and be randomly rewired to some other available agent to play the same game.

We show, for the first time, that cooperation emerges and is sustained in such games *without requiring agents to follow a predefined partner selection rule.* Using Multiagent Q-learning with Boltzmann exploration, we show that agents are capable of learning cooperation-sustaining rules by themselves, including Out-for-Tat, where ties are broken upon experiencing defection, and zealot partner selection strategies, that keep interacting after experiencing defection even at the cost of losing payoff. Consistently with the literature, but requiring much fewer assumptions on agents' observation and decision-making capabilities, we witness the emergence of a largely prevailing percentage of cooperative strategies, including the notorious Tit-for-Tat. The co-evolution of partner selection and in-game decision strategies in SDOOs leads to the emergence and stabilisation of novel cooperation-inducing behaviours, discovering new surprising patterns.

Related Literature. Research in multi-agent systems has produced a significant number of contributions over the years studying how mechanisms could arise to promote socially desirable behaviour in agents' societies [31], starting from centralised solutions such as social laws [26] and cooperation structures [9] to decentralised ones such as trust and reputation [8]. The study of reputation in particular has been one of the pillars of multiagent systems research ever since its inception [18, 20].

The role of partner selection in sustaining cooperation is well-known and widely investigated in the wider literature [23] and the capacity of breaking and forming ties was identified as one of the five key rules to enable it [14]. Partner selection was studied in relation to the emergence of cooperation [10, 22] and coordination [25] and as a tool to enforce the ostracism of unreliable partners [16, 29]. Theoretical and experimental findings have often assumed the capacity to observe the potential partners' behaviour [19]. When this is not the case other mechanisms were put in place, such as communication, social image and indeed reputation [17, 21, 24], to provide noisy signals of agents' types. The capacity to actively observe and even choose one's partner is also assumed in several models [2, 5], but this is unrealistic in many real-world situations. In our paper, we study the emergence of cooperation without assuming active partner selection or reputation mechanisms.

Without any mechanism at all, i.e., random matching, mass defection takes over [2]. Voluntary participation in a social dilemma is known to be a simple and effective mechanism to promote cooperation [27] and is closer in spirit to the games we study. In these settings, costs are usually modified to make it so that no participation is at least as good as the worst possible game outcome. In our framework, this assumption is not required.

$$\begin{array}{c|c|c|} & C & D \\ \hline C & R,R & S,T \\ \hline D & T,S & P,P \\ \hline \end{array} \qquad \begin{array}{c|c|c|} & C & D \\ \hline C & 3,3 & 0,5 \\ \hline D & 5,0 & 1,1 \\ \hline \end{array}$$

Figure 1: General payoff matrix (on the left side) and a concrete instantiation (on the right side) of the Prisoner's Dilemma game. The payoffs need to be such that $T > R > P > S$ and $2R > T + S$.

In the context of the Prisoner's Dilemma with opting out [32] shows experimental evidence that cooperation can be sustained, backing the findings with a replicator dynamics analysis, fully laid out in [33]. The analysis is however based on the assumption that agents obey the Out-for-Tat rule, which we lift in this paper.

The use of multiagent learning to find equilibrium strategies in complex social dilemmas is widely established [6] and has covered the mitigation of free riding in common resource appropriation problems [15], with active partner selection [2] and agents with non-strictly utilitarian moral stances [1]. The study of population-based training methods and their application to mean-field games is an important related research area [13].

Paper Structure. Section 2 provides the needed game theoretic and learning background. Section 3 presents our experimental approach and design, while Section 4 discusses our results. We conclude with Section 5 discussing various potential directions.

2 PRELIMINARIES

In what follows we introduce social dilemma games and equip them with the option of opting out. Further, we present the Q-learning algorithm, discussing the exploration policy of use.

2.1 Social Dilemmas with Opting Out

We study societies where self-interested agents are paired with one another and play a Prisoner's Dilemma (PD) repeatedly. At each round of the game, players pick an action from the set $A = \{C, D\}$, where C denotes cooperation and D defection, with each player receiving a payoff at each combination of choices, also known as strategy profile or outcome. The general form and a concrete instantiation are depicted in Figure 1, where the payoff vector (r_i, r_j) assigns r_i to the row player and r_j to the column player at the corresponding choice combination.

It is well-known that in the one-shot version of the game defection is a strictly dominant strategy and, when agents adjust their one-shot policies based on the reward received, the replicator dynamics converge to defection as a unique evolutionarily stable strategy [7]. However, mutual cooperation (C, C) would indeed be a more desirable outcome as it maximises social welfare, the sum of rewards of both players. In other words, a cooperative population is significantly more desirable than a population of defectors.

In the repeated version of this game, many other strategies are possible, including the well-known Tit-for-Tat (TFT), which copies the partner's last choice, and simple stubborn strategies such as Always-Cooperate (All-C) and Always-Defect (All-D), which keep cooperating and defecting, respectively.

In our setting, we study social dilemmas with the option of opting out (SDOOs). These work as follows:

(1) At the start, players are randomly matched in pairs.
(2) Each player decides whether to continue playing with the current partner or break ties. If ties are broken both the player and their partner become available.
(3) The players who chose to not break ties continue to play together.
(4) Available players are randomly matched in pairs.
(5) Each player plays a Prisoner's Dilemma with the current partner receiving the associated reward. The process continues to the next round at step (2) and repeats.

It is worth noting that in SDOOs agents who decide to break ties have no say on the "type" of player they will be paired with and they could continuously be matched with the same partner over again if they two are the only ones available.

In SDOOs – and similar considerations can be made for games with a partner selection phase – long-term strategies such as Tit-for-Tat are still possible, but the effects of such strategies need not be felt in classic direct retaliatory fashion: if my partner defected and I decide to break ties, I will take revenge on my next partner, rather than the previous one! If ties are not broken, Tit-for-Tat will function as usual.

Another important dimension of SDOOs is that there are partner selection strategies, as well. For example, Out-for-Tat (OFT), which breaks ties with defectors, Always-Stay (Stay) and Always-Switch (Switch), which keep staying and switching, respectively.

TFT was shown to be learnable when agents can choose who to play with at each round [2] but mass defection was shown to emerge under random rewiring. On the other hand, cooperation was shown to emerge when forcing OFT in SDOOs [32, 33]. Whether OFT can be learnt from first principles and this still enables cooperation to be sustained, for instance through co-learning TFT, is the question we ask in this paper.

2.2 Q-Learning

Q-learning is a widely used reinforcement learning algorithm [30]. It runs on a Markov decision process (MDP), a tuple $G = \langle S, A, T, R \rangle$, where S is a set of states, A is a set of the available action, $T : S \times A \times S \to [0, 1]$ is a state transition probability function and $R : S \times A \to \mathbb{R}$ is an *immediate* reward function. Note that the state transition satisfies the Markov property, where $Pr(s_{t+1} \mid s_0, a_0, ...s_t, a_t) = Pr(s_{t+1} \mid s_t, a_t)$. A learning agent aims to find a policy $\pi(a|s)$ that maximises the expected discounted cumulative reward or *profit* $J = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h \rho_{h+1}]$ during repeated game plays, where $\gamma \in [0, 1]$ is the discount factor, and $\rho_{h+1} = R(s_h, a_h)$ is the immediate reward obtained by the agent when it enters state s_{h+1} from s_h after choosing action a_h, starting from state s_0.

The Q-learning algorithm is a model-free approach to evaluate the optimal profit. Specifically, a Q-learning agent maintains a Q-value for each state-action pair (s, a) to estimate the profit of using each action $a \in A$ under each state $s \in S$. Suppose that at a given time step t, the agent is in state s and selects an action a_i, we denote the corresponding Q-value as $Q_i(t, s) := Q(t, s, a_i)$. Let r_t be the immediate reward it receives as the new state becomes s'. The agent updates its Q-value for the state-action pair (s, a_i) as follows:

$$Q_i(t+1, s) = (1-\alpha)Q_i(t, s) + \alpha(r_t + \gamma \max_{a_j \in A} Q_j(t, s')) \tag{1}$$

where $\alpha \in (0, 1)$ is the learning rate, and $\max_{a_j \in A} Q_j(t, s')$ estimates the profit after state transition. The solution $\boldsymbol{Q}^*(s) = (Q_1^*(s), ..., Q_d^*(s))$ is the optimal action value function.

The convergence of Q-learning to the optimal action-value function has been proved, under the conditions that each state-action pair (s, a) is visited infinitely often during training and with a suitable learning rate [28, 30]. An exploration mechanism aims to strike a balance between exploitation and exploration such that the performance of the agent is maximized during learning while ensuring the converging condition is met. Boltzmann exploration is a commonly used exploration mechanism. Under Boltzmann exploration, the action selection probabilities $\boldsymbol{\pi} := \boldsymbol{\pi}(t, s) = (\pi_1, ..., \pi_d) \in \Delta$ is given by

$$\pi_i = \frac{e^{\kappa Q_i}}{\sum_{j=1}^{d} e^{\kappa Q_j}} \tag{2}$$

where κ is a parameter known as the inverse of the temperature. The agent is in pure exploration (randomly taking each action) when κ is 0, and in pure exploitation (taking the action with the highest Q-value) when $\kappa \to \infty$.

3 EXPERIMENTAL SETTING

Consider a population of $N = 20$[1] agents learning to play rounds of the Prisoner's Dilemma (PD) game on the right side of Figure 1, with players initially paired at random. At the beginning of each new round, agents are able to decide whether they will switch partners to play the game. The goal of the agents is to earn the highest payoff across $M = 20$ rounds of the game. Therefore the number of actions taken is $2M$ in one episode, half of them are about whether to stay or switch, and the other half are about whether to cooperate or the defect. The best outcome for the population (in terms of total rewards of all agents) is achieved when all agents cooperate every time. However, such an outcome is hard to achieve since the immediate reward for defection is higher, especially in a highly cooperative society. With the opting-out mechanism, an agent has to choose between the immediate reward of defecting against its cooperative partner without fear of immediate retaliation and the future rewards generated by stable cooperation.

The computational model of the game dynamics is described in Algorithm 1. Each round of the game is divided into two stages. In the first stage (partner selection stage: line 5 to 16), each agent can choose whether to switch its partner to play the PD game in the next stage. The outcome is an action profile (a^i_{PS}, a^j_{PS}), where $a_{PS} \in \{N, Y\}$ for every pair of agents, agent i and agent j, where Y stands for yes, thus breaking ties, and N for no, thus staying. If both decide to stay, they will play the game in the next stage. If anyone in the pair decides to switch partners, both of them will be put into the shuffling pool and await to be paired (line 10 to 13). At the

[1] We also experimented with $N = 40$ and $N = 60$ and results are practically unaffected.

Algorithm 1 Social dilemmas with the option of opting out

```
Input: $N, M, T, \alpha, \tau, \gamma$
1:  Initialize $Agents$ with $N, \alpha, \tau, \gamma$, $Pool = [\,]$
2:  Initialize $Pairs$, $LastActions$ randomly
3:  for $episode = 1$ to $T$ do
4:      for $round = 1$ to $M$ do
5:          for each $pair$ in $Pairs$ parallel do
6:              $i \leftarrow pair[0]$, $j \leftarrow pair[1]$
7:              $s^i_{PS} \leftarrow LastActions[j]$, $s^j_{PS} \leftarrow LastActions[i]$
8:              $a^i_{PS} \leftarrow Agents[i].getAction(s^i_{PS})$
9:              $a^j_{PS} \leftarrow Agents[j].getAction(s^j_{PS})$
10:             if $a^i_{PS}$ or $a^j_{PS}$ == “$Y$” then
11:                 $Pool.add(i, j)$
12:                 $Pairs.remove(pair)$
13:             end if
14:         end for
15:         $newPairs \leftarrow generatePairs(Pool)$, $Pool = [\,]$
16:         $Pairs.add(newPairs)$
17:         for each $pair$ in $Pairs$ parallel do
18:             $i \leftarrow pair[0]$, $j \leftarrow pair[1]$
19:             $s^i_{PD} \leftarrow LastActions[j]$, $s^j_{PD} \leftarrow LastActions[i]$
20:             $a^i_{PD} \leftarrow Agents[i].getAction(s^i_{PD})$
21:             $a^j_{PD} \leftarrow Agents[j].getAction(s^j_{PD})$
22:             $r^i_{PD}, r^j_{PD} \leftarrow playGame(a^i_{PD}, a^j_{PD})$
23:             $LastActions[i] \leftarrow a^i_{PD}$, $LastActions[j] \leftarrow a^j_{PD}$
24:             $Agents[i].updateReward(r^i_{PD})$
25:             $Agents[j].updateReward(r^j_{PD})$
26:         end for
27:     end for
28:     for each $agent$ in $Agents$ parallel do
29:         $agent.train()$ with equation (1)
30:     end for
31: end for
```

end of this stage, agents in the shuffling pool are randomly paired up and continue to the next stage (line 15 to 16). As previously observed, the chances of being paired with the same partner are not insignificant, especially with a small number of available players. In the second stage (PD game stage: line 17 to 26), every pair of agents will play a PD game and receive their rewards (r^i_{PD}, r^j_{PD}) based on the payoff in Figure 1. The outcome is of the form (a^i_{PD}, a^j_{PD}), where $a_{PD} \in \{C, D\}$, for every pair of agents.

At every stage of the game, the agents' decisions are made based on the opponents' previous actions in the PD game. That is, the agent observes only the last action of their opponent as in related studies [2, 32, 33]. Therefore, the MDP has 4 different states, $S = \{PS_C, PS_D, PD_C, PD_D\}$, 2 from the partner selection stage $S_{PS} = \{PS_C, PS_D\}$ and 2 from the PD game stage $S_{PD} = \{PD_C, PD_D\}$. We leave the generalisation to strategies with memory, more fine-grained but computationally more expensive, to future analysis. As we shall show, agents are able to learn cooperation-inducing behaviour using only this basic piece of information, which also allows for a direct comparison with existing results.

Each agent maintains two policies for action selection. The first policy π_{PS} determines the probability of switching a_{PS} given s_{PS} and the second policy π_{PD} determines the action probabilities in the PD game a_{PD} given s_{PD}. During each episode, a trajectory $\kappa = \{s_{PS}, a_{PS}, r_{PS} = 0, s_{PD}, a_{PD}, r_{PD}, ...\}$ is sampled, and the policy is updated based on the Q-learning algorithm (line 28 to 30). Every agent in the population adopts the standard Q-learning with Boltzmann exploration, with the learning rate $\alpha = 0.05$, which we optimised through a simple line search from 0.01 to 0.1. The temperature is set at $\tau = 1$ and the discount rate at $\gamma = 1$.

4 EXPERIMENTAL RESULTS

In this section, we present our results, showing how cooperation can emergence in SDOOs. We will also analyse the types of strategies that are learnt, both during the game stage and the partner selection stage. We will then zoom in at the single-agent level, analysing policy traces and reward distributions. Finally, we will show what happens in our setting when OFT is imposed and how this compares with our main findings.

4.1 Emergence of cooperation in SDOOs

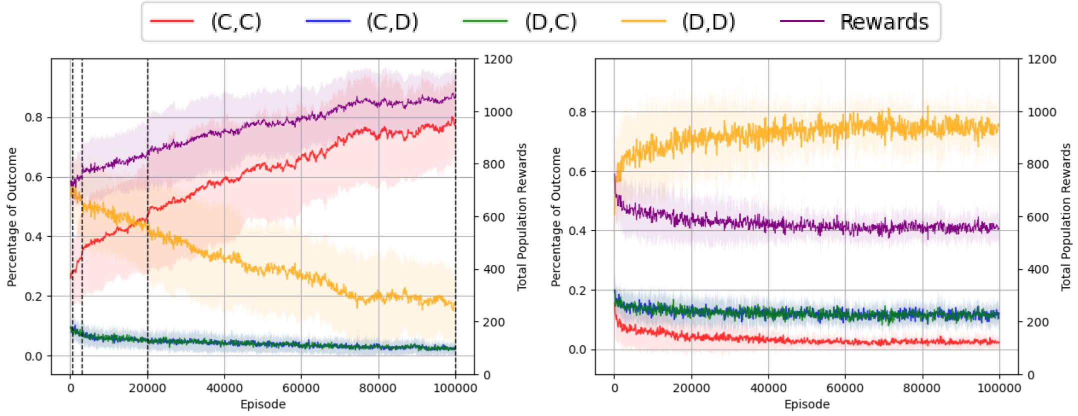
When agents play SDOOs and are allowed to form their decision on whether or not to stick to the current partner based on the reward received, cooperation prevails, with the population total reward increasing over time. Figure 2a displays the results, showing that cooperative choices dominate societies of self-interested agents. The figure shows the percentage of the strategy profiles of the PD game as well as the population rewards.

SDOOs can be seen as a game with almost random matchings. If the decision to break ties with the current partner were also to be decided without considering the obtained reward, we would be back to the case of well-mixed populations, where mass defection is known to emerge [2]. Indeed, as shown in Figure 2b, a population of agents playing a PD with random matching will see defection rapidly take over.

Zooming in on the learning curves in Figure 2a, we observe four different phases of learning which are characterized by different outcome distributions across the population in the PD game, and they are marked in Figure 2a accordingly. Each of these phases corresponds to the development of a population strategy as follows:

- Phase 1 (episodes 0 to 500): Agents learn to defect.
- Phase 2 (episodes 500 to 3, 000): Agents realise staying with the same partner (Always-Stay) and staying with a cooperator but not with a defector (Out-for-Tat) are good choices.
- Phase 3 (episodes 3, 000 to 20, 000): the Always-Stay and Out-for-Tat strategies become dominant in the partner selection stage, and agents start to learn to cooperate (Always-Cooperate) or cooperate with a cooperative partner but defect against a non-cooperative partner (Tit-for-Tat).
- Phase 4 (episodes 20, 000 to 100, 000): cooperation stabilises at the population level, Always-Cooperate and Tit-for-Tat strategies become dominant in the PD game.

In our analysis, we have limited the amount of information an agent can obtain to the last action performed by the opponent in the PD game (cooperate or defect), but agents trial strategy selection



(a) Matching in PDs with the option of opting out

(b) Random Matching

Figure 2: The mean and standard deviation of the percentage of the outcomes of the PD game and the total population rewards across episodes, summarized over 20 simulations. The introduction of the partner selection stage is key to the emergence of cooperation. The dashed lines are marked to indicate the ending of each phase. The percentage of cooperative outcome (C, C) starts to take over at the end of phase 3. In the case of random matching, the population learns to defect rapidly.

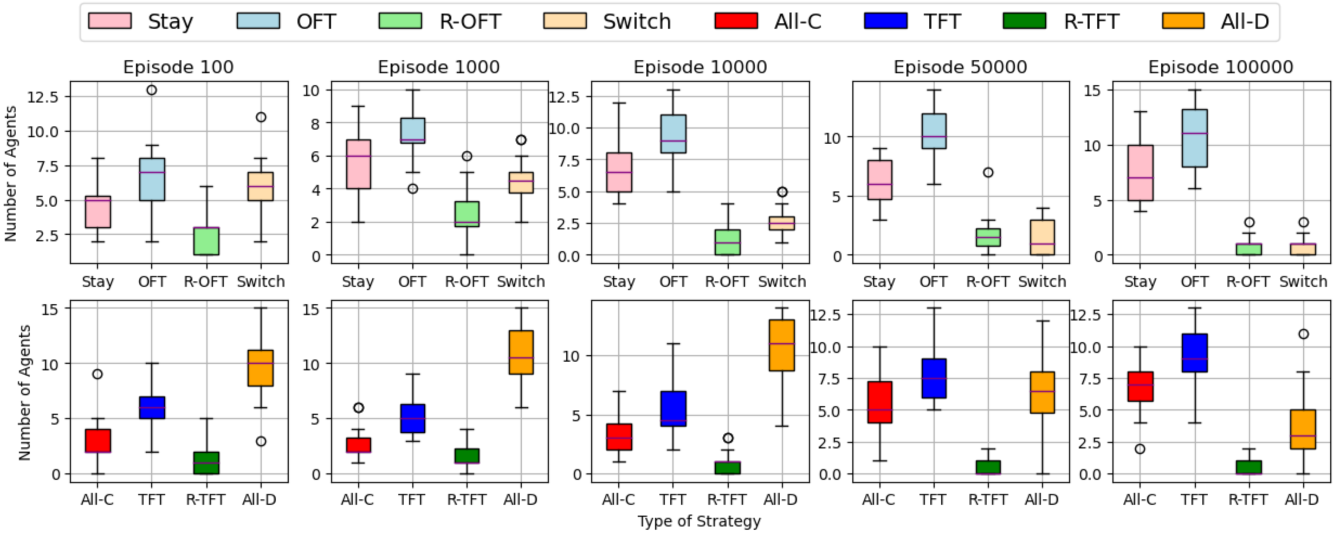


Figure 3: Box plots of the number of agents that use each strategy during different phase transitions. The OFT strategy in the partner selection stage is developed rapidly across the population. The successful development of OFT has led to the successful development of TFT and All-C, and thus a cooperative society.

over a timescale of 20 game plays which is, once again, in line with analogous studies with active partner selection [2].

By comparing the magnitude of Q-values at different states, we can classify the agent policy into different types of strategy. For example, in the partner selection stage, if the Q-value of action N is larger than that of action Y regardless of the opponent's last action ($Q(N|PS_C) > Q(Y|PS_C)$, $Q(N|PS_D) > Q(Y|PS_D)$), we classify the agent as adopting the Always-Stay strategy; If the Q-value of action N is larger than that of action Y when the opponent has cooperate in the last action, but reverse otherwise ($Q(N|PS_C) > Q(Y|PS_C)$, $Q(N|PS_D) < Q(Y|PS_D)$), we classify the agent as adopting the Out-for-Tat strategy; and so forth. We are therefore able to classify agents' strategies into four different types in the partner-selection stage, and four different types in the PD game stage. For the partner-selection stage, these are (1) Always-Stay (Stay), (2) Out-for-Tat (OFT) where the agent chooses to stay if its current partner cooperated and chooses to leave if its current partner has defected, (3) reverse Out-for-Tat (R-OFT) where the agent chooses to leave if its current partner cooperated and chooses to stay if its current partner defected, and (4) Always-Switch (Switch). Likewise, for the

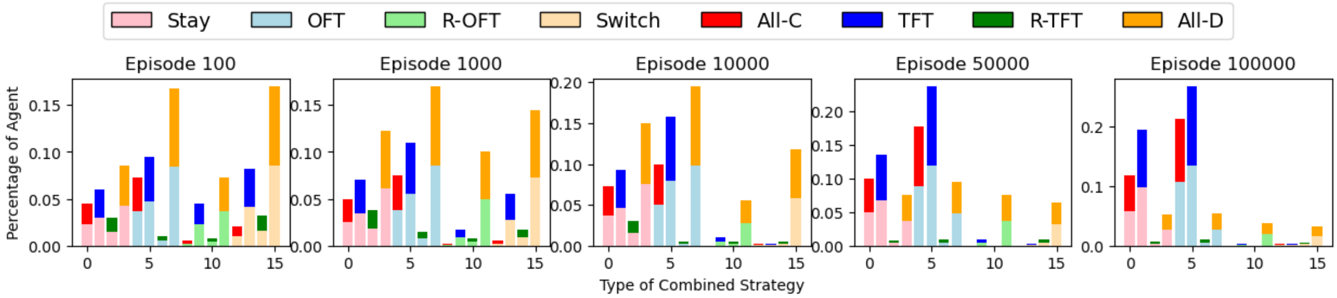


Figure 4: Percentage of agents adopting each combined strategy during different phase transitions. Nearly 50% of the agents have adopted the OFT, TFT and the OFT, All-C strategies by the end of the training phase.

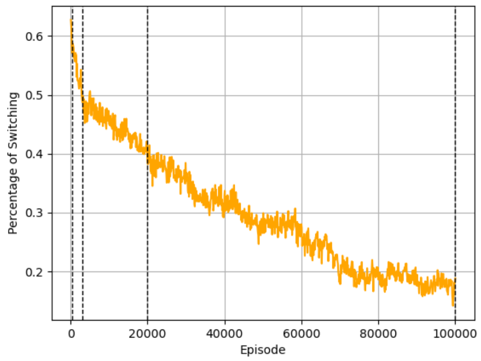


Figure 5: Percentage of agents who switch partners in the partner selection stage, by episode. An initial drop to roughly 45% is followed by a steady decrease to about 20%.

game stage, these are (1) Always-Cooperate (All-C), (2) Tit-for-Tat (TFT) where the agent chooses to cooperate next if its current partner cooperated and chooses to defect next if its current partner defected, (3) reverse Tit-for-Tat (R-TFT) where the agent chooses to defect next if its partner cooperated and chooses to cooperate next if its partner defected, and (4) Always-Defect (All-D) in the PD game stage. We note that, because of the random component of the exploration policy, actions that are actually taken by the agent may at times differ from their intended strategy type. In other words, our learning agents may attempt actions that are not considered currently best to avoid being stuck in local minima.

To understand how strategies emerge and stabilise in the population, we have further produced: (i) a box plot illustrating these trends at different phases (Figure 3) (ii) a bar chart showing the percentage of agents adopting the combined strategy ($16 = 4 \times 4$ in total) at different phases (Figure 4) (iii) a line plot outlining the proportion of agents changing partner across episodes (Figure 5).

In the first phase, agents learn to defect in the PD game. Whereas in the partner selection stage, agents generally learn to adopt the Out-for-Tat and the Always-Switch strategies. During this phase, agents have not yet come up with a good strategy for partner selection. As we can see in Figure 5, agents have switched their partners most of the time. This exploratory phase is comparable to agents learning in the random matching environment, where mutual defection is the best strategy for all.

In the second phase, some agents learn that staying with the same partner is better than switching to a new one, as stable connections facilitate cooperation. Besides the Out-for-Tat strategy, we can observe how the Always-Stay strategy has become more popular across the population, and we can see the percentage of partner switching has experienced a large decrease. On the other hand, defection in the PD game remains dominant. However, the cooperative outcome (C, C) in the PD game is increasing, as the cooperative and Tit-for-Tat agents have stuck to each other.

In the third phase, the Out-for-Tat strategy becomes dominant, followed by the Always-Stay strategy. As a consequence of this fact, the environment becomes more favourable to the emergence of Always-Cooperate and Tit-for-Tat. The use of Out-for-Tat has provided another way out to the defectors without hurting the rewards of the cooperative agent itself. This also lowers the expected reward for defective actions since they cannot take advantage of a cooperative agent continuously. At the end of this stage, the number of cooperative outcomes (C, C) starts to overtake the number of defection outcomes (D, D) in the PD gameplays.

In the fourth phase, the number of Tat-for-Tat agents becomes predominant. The increase of TFT agents has largely affected the payoff of the defecting agent, and this drives the defectors to learn to cooperate over time, causing the steady growth of cooperative outcomes (C, C) in the PD game. At the episode 100, 000, nearly 80% of agents have cooperated in the PD game, and we can foresee continuous growth in the future [2]. In the partner selection stage, we can see a stable combination of Stay agents and OFT agents. In the PD game, we can see a stable combination of All-C and TFT agents, while the All-D agents are now a minority.

We can clearly see the importance of the TFT strategy in developing a cooperative population, as TFT is the best way to limit defecting behaviour, confirming what is known from the literature [14]. We also note the role of partner selection in the development of the TFT strategy, even when retaliation is not necessarily direct, like in our case. Punishing defection is likely to harm the perpetrator as well, and finding resilient cooperators is key to converging

[2] We have conducted the simulations up to 200, 000 episodes and the percentage of cooperation stabilises around the 85% mark.

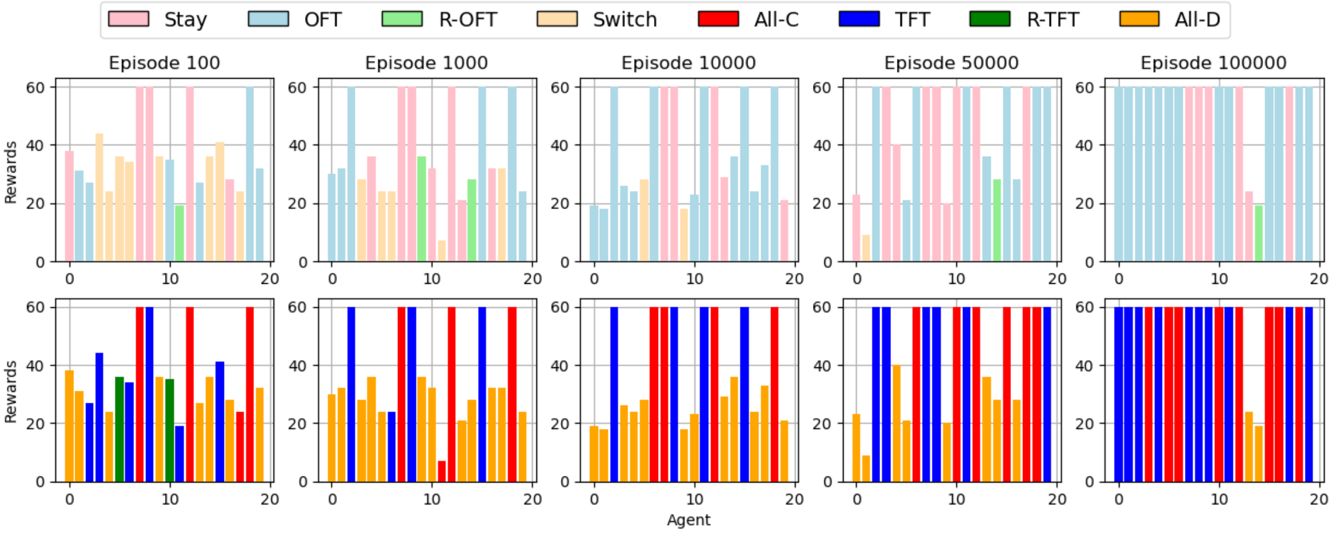


Figure 6: Episodic rewards by agents using different strategies during different phase transitions. The OFT strategy in the partner selection stage facilitates the TFT and All-C agents to find and stick to each other. This is important to the survival of cooperative agents, especially among the population of defectors.

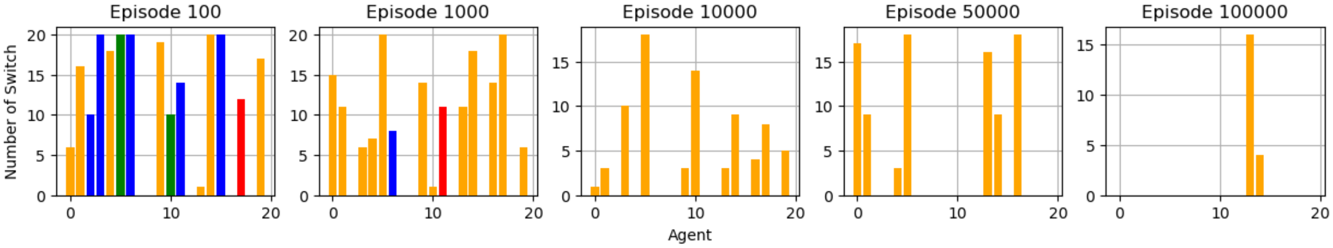


Figure 7: Number of partner switches decided by agents using different PD game strategies during different phase transitions. At the later phase of learning, most of the decisions to switch are made by the All-D agents.

towards more positive rewards in the long run. The opting-out mechanism allows agents for minimal freedom with partner selection. The agents can only decide to exit the current interaction given the last observed move of the current partner, and no other information about the other players. Compared to TFT, on the other hand, the OFT strategy is far easier to learn by the agents. This is because staying to play with a cooperator is always good in any case, and leaving a defector does not cause a loss to the reward of the agent. The OFT strategy enables the cooperators to meet and stick together, and this is favourable to the development of the TFT strategy, therefore the success of OFT has led to the successful development of TFT, and thus a cooperative society.

4.2 Analysis of an individual agent

The total reward and the total number of switches in the population provide insightful information about the evolution of the population, yet examining the performance at the individual level can help us realise how the agent's strategy is developed and sustained. We plot the episodic rewards for every agent from a single simulation marked by the type of strategy they adopt in Figure 6. We also plot the number of switches they decided to make in Figure 7.

In episode 100, various types of strategies have been developed, and we can see some cooperative and TFT agents have recognised themselves and formed some stable pairs, receiving the reward of $60 = 3 \times 20$. At this stage, switching partners in the partner selection phase is a popular choice among the agents. In episode $1,000$, the OFT strategy is getting more popular and there is an obvious decrease in switching partners. However, as the number of All-D agents becomes dominant in the population, we can see the rewards of All-C and TFT agents will be lower if they cannot find the right partner. In episode $10,000$, the OFT and Stay strategies dominate in the partner selection phase. This provides a good environment for the defectors to learn to cooperate. We can see the minority of All-C and TFT agents manage to survive by sticking together, and this is also a key to future success when the TFT agents manage to affect the All-D agents in the population due to exploration. The positive factors are reflected in episode $50,000$, where the All-C and TFT agents become dominant. This trend continues with only 2 defectors left in the entire population by episode $100,000$.

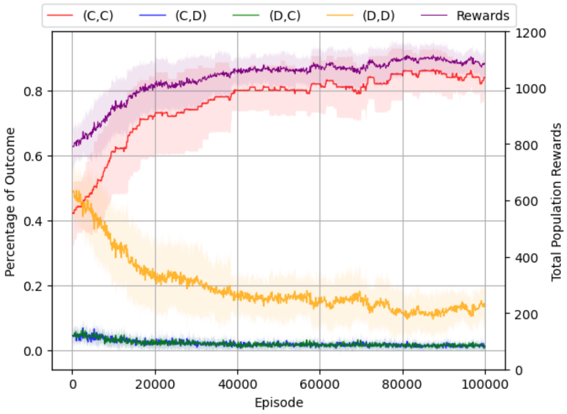


Figure 8: Percentage of the outcomes on the PD game and the total population rewards across episodes, forcing the Out-for-Tat strategy in the partner selection phase. The percentage of cooperative outcome (C, C) grows rapidly under this setting.

4.3 Forcing Out-for-Tat

Cooperation was found to emerge in games where agents follow OFT [32, 33]. In this section, we analyse what happens when these conditions are reproduced in our framework, which is why we conducted the same experiments in the previous section with OFT imposed. Notice that under these constraints it follows that if any agent has chosen defection then, at the next round, both they and their partner will be randomly rematched to available agents, including themselves. Otherwise, they will keep playing together. The strategy profile distribution in the PD game, as well as the total rewards across the episodes, are plotted in Figure 8. We can observe that defectors dominate at the beginning of training. However, the situation reverses quickly due to the steady growth of cooperators and TFT agents. By the end of the training, the percentage of cooperative outcome (C, C) has grown significantly and more than 80% of agents choose to cooperate. Similarly to what was observed in the previous sections, the use of the OFT strategy has helped All-C agents and TFT agents to identify each other and prevent the All-D agents from exploiting cooperators, leading the population as a whole to be fundamentally cooperative. Unlike the general case, though, where OFT is not imposed, we note a difference in the shape of the curves. With OFT, the (D, D) outcomes decrease more quickly and stabilise earlier, with similar considerations for the profiles where at least one of the players cooperates.

5 DISCUSSION

We carried out a multiagent reinforcement learning analysis on repeated social dilemma games with the option of opting out where, for the first time, we have seen the emergence of cooperation without forcing players to follow a given partner selection rule. We observed that agents learn such rules by themselves, in particular Out-for-Tat, which was known to promote significant levels of cooperation. Our agents were able to learn several other partner selection rules and co-evolve cooperative strategies. Simulations have shown an interesting effect of timescales in strategy adoption, where agents were quicker to learn partner selection rules than in-game decisions. Although the effects of timescales in promoting cooperation is a known phenomenon [5, 23], for this to happen with only the option of opting out is surprising and certainly deserves further investigation.

Although we have obtained predominantly cooperative behaviour, it is important to check how fragile this is, in other words how perturbations of the starting conditions will affect the equilibrium behaviour. The first variant that comes to mind is the introduction of trembling-hand behaviour, where mistakes are allowed, not only within the game but also during partner selection level. For example [33] study the robustness of OFT assuming players that choose to stay still break ties with probability ρ. In our framework, where partner selection rules are not fixed, but learnt, this would amount to allowing a fixed exploration rate rather than a Boltzmann one, which may have repercussions for convergence. While Boltzmann exploration was functional to the stabilisation of cooperative pairs, the same may not hold for other exploration strategies, notably ϵ-greedy. Together with exploration, other hyperparameters need to be accurately optimised, starting from the number of training rounds at the game stage, discounting factor and learning step, which may present non-linear behaviour. We are leaving this more extensive analysis of alternative configurations for future work.

An important research direction is studying what happens when going beyond agent pairs, e.g., allowing players to be chosen by multiple partners while keeping tie-breaking an individual decision. This may give rise to nodes that act as attractors, where more frequent cooperators tend to have a higher degree distribution. Following up on this idea, we could also allow for matchings that themselves depend on the degree distribution (agents are more likely to be paired to agents with a higher number of current partners) or agent types (agents are more likely to be paired with agents who behaved like them in the past). Using multiagent learning to learn partner selection rules on networks may shed light on the emergence of cooperator-cores and defector-peripheries [5].

When active partner selection is allowed, another interesting direction is that of going beyond unilateral attachment, with players forming connections only by mutual consent [11]. Another potential direction involves payoff transfers and endogenous games [12], with players willing to sacrifice payoff to make other players not leave the current game and achieve cooperation.

In our work, we have limited the amount of information an agent can obtain to the last action performed by the current partner in the Prisoner's Dilemma game. It is natural, although computationally more demanding, to allow players to remember more moves and react accordingly. This will allow for more complex strategies to emerge and likely transform the game equilibria, opening further avenues for future research. Finally, the experimental analysis of social dilemmas with opting out shows that humans display particular cooperation rates [32]. It is an exciting research challenge to understand the learning conditions needed to retrieve those results.

ACKNOWLEDGMENTS

The authors acknowledge the support of the Leverhulme Trust for the Research Grant RPG-2023-050 and the TAILOR Connectivity Fund (Agreement 29).

REFERENCES

[1] Nicolas Anastassacos, Julian García, Stephen Hailes, and Mirco Musolesi. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). ACM, 115–123. https://doi.org/10.5555/3463952.3463972

[2] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. 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, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.* AAAI Press, 7047–7054. https://doi.org/10.1609/aaai.v34i05.6190

[3] Robert Axelrod. 1980. Effective Choice in the Prisoner's Dilemma. *The Journal of Conflict Resolution* 24, 1 (1980), 3–25. http://www.jstor.org/stable/173932

[4] Jacques Bara, Fernando P. Santos, and Paolo Turrini. 2023. The Role of Space, Density and Migration in Social Dilemmas. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh (Eds.). ACM, 625–633. https://doi.org/10.5555/3545946.3598692

[5] Jacques Bara, Paolo Turrini, and Giulia Andrighetto. 2022. Enabling imitation-based cooperation in dynamic social networks. *Auton. Agents Multi Agent Syst.* 36, 2 (2022), 34. https://doi.org/10.1007/s10458-022-09562-w

[6] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *J. Artif. Intell. Res.* 53 (2015), 659–697. https://doi.org/10.1613/jair.4818

[7] Tilman Börgers and Rajiv Sarin. 1997. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* 77, 1 (1997), 1–14. https://doi.org/10.1006/jeth.1997.2319

[8] Cristiano Castelfranchi. 1998. Modelling Social Action for AI Agents. *Artif. Intell.* 103, 1-2 (1998), 157–182. https://doi.org/10.1016/S0004-3702(98)00056-3

[9] Mark d'Inverno, Michael Luck, and Michael J. Wooldridge. 1997. Cooperation Structures. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence, IJCAI 97, Nagoya, Japan, August 23-29, 1997, 2 Volumes.* Morgan Kaufmann, 600–605.

[10] Nigel Gilbert. 1995. Emergence in social simulation. In *Artificial Societies: The Computer Simulation Of Social Life*, Nigel Gilbert and Rosaria Conte (Eds.). Routledge. https://doi.org/10.4324/9780203993699

[11] Matthew O. Jackson and Alison Watts. 2002. The Evolution of Social and Economic Networks. *J. Econ. Theory* 106, 2 (2002), 265–295. https://doi.org/10.1006/jeth.2001.2903

[12] Matthew O. Jackson and Simon Wilkie. 2005. Endogenous Games and Mechanisms: Side Payments Among Players. *The Review of Economic Studies* 72, 2 (04 2005), 543–566. https://doi.org/10.1111/j.1467-937X.2005.00342.x arXiv:https://academic.oup.com/restud/article-pdf/72/2/543/18338314/72-2-543.pdf

[13] Paul Muller, Mark Rowland, Romuald Elie, Georgios Piliouras, Julien Pérolat, Mathieu Laurière, Raphaël Marinier, Olivier Pietquin, and Karl Tuyls. 2022. Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 926–934. https://doi.org/10.5555/3535850.3535954

[14] Martin A. Nowak. 2006. Five Rules for the Evolution of Cooperation. *Science* 314, 5805 (2006), 1560–1563. https://doi.org/10.1126/science.1133755 arXiv:https://www.science.org/doi/pdf/10.1126/science.1133755

[15] Julien Pérolat, Joel Z. Leibo, Vinícius Flores Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 3643–3652. https://proceedings.neurips.cc/paper/2017/hash/2b0f658cbffd284984fb11d90254081f-Abstract.html

[16] Adrian Perreau de Pinninck, Carles Sierra, and Marco Schorlemmer. 2010. A multiagent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems* 21, 3 (01 Nov 2010), 397–424. https://doi.org/10.1007/s10458-009-9107-8

[17] T. Pfeiffer, L. Tran, C. Krumme, and D.G. Rand. 2012. The value of reputation. *Journal of the Royal Society Interface* 9 (2012), 2791–2797.

[18] Josep M. Pujol, Ramon Sangüesa, and Jordi Delgado. 2002. Extracting Reputation in Multi Agent Systems by Means of Social Network Topology. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1* (Bologna, Italy) *(AAMAS '02)*. Association for Computing Machinery, New York, NY, USA, 467–474. https://doi.org/10.1145/544741.544853

[19] David G. Rand, Samuel Arbesman, and Nicholas A. Christakis. 2011. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences* 108, 48 (2011), 19193–19198. https://doi.org/10.1073/pnas.1108243108 arXiv:https://www.pnas.org/content/108/48/19193.full.pdf

[20] Jordi Sabater and Carles Sierra. 2002. Reputation and social network analysis in multi-agent systems. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02* (Bologna, Italy). ACM Press, New York, New York, USA.

[21] Jordi Sabater-Mir, Mario Paolucci, and Rosaria Conte. 2006. Repage: REPutation and ImAGE Among Limited Autonomous Partners. *Journal of Artificial Societies and Social Simulation* 9, 2 (2006), 3. https://www.jasss.org/9/2/3.html

[22] Norman Salazar, Juan A. Rodriguez-Aguilar, Josep Ll. Arcos, Ana Peleteiro, and Juan C. Burguillo-Rial. 2011. Emerging Cooperation on Complex Networks. In *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2* (Taipei, Taiwan) *(AAMAS '11)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 669–676.

[23] Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. 2006. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLoS Comput. Biol.* 2, 10 (2006). https://doi.org/10.1371/journal.pcbi.0020140

[24] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. 2018. Social Norms of Cooperation With Costly Reputation Building. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 4727–4734. https://doi.org/10.1609/aaai.v32i1.11582

[25] Sven Van Segbroeck, Steven de Jong, Ann Nowé, Francisco C. Santos, and Tom Lenaerts. 2010. Learning to coordinate in complex networks. *Adapt. Behav.* 18, 5 (2010), 416–427. https://doi.org/10.1177/1059712310384282

[26] Yoav Shoham and Moshe Tennenholtz. 1995. On Social Laws for Artificial Agent Societies: Off-Line Design. *Artif. Intell.* 73, 1-2 (1995), 231–252. https://doi.org/10.1016/0004-3702(94)00007-N

[27] Gyorgy Szabo and Christoph Hauert. 2003. Evolutionary prisoner's dilemma games with voluntary participation. *Physical review. E, Statistical, nonlinear, and soft matter physics* 66 (01 2003), 062903. https://doi.org/10.1103/PhysRevE.66.062903

[28] John N Tsitsiklis. 1994. Asynchronous stochastic approximation and Q-learning. *Machine learning* 16, 3 (1994), 185–202.

[29] Jing Wang, Siddharth Suri, and Duncan J. Watts. 2012. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences* 109, 36 (2012), 14363–14368. https://doi.org/10.1073/pnas.1120867109 arXiv:https://www.pnas.org/content/109/36/14363.full.pdf

[30] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

[31] Michael J. Wooldridge. 2009. *An Introduction to MultiAgent Systems, Second Edition.* Wiley.

[32] Bo-Yu Zhang, Song-Jia Fan, Cong Li, Xiu-Deng Zheng, Jian-Zhang Bao, Ross Cressman, and Yi Tao. 2016. Opting out against defection leads to stable coexistence with cooperation. *Scientific reports* 6 (October 2016), 35902. https://doi.org/10.1038/srep35902

[33] Xiu-Deng Zheng, Cong Li, Jie-Ru Yu, Shi-Chang Wang, Song-Jia Fan, Bo-Yu Zhang, and Yi Tao. 2017. A simple rule of direct reciprocity leads to the stable coexistence of cooperation and defection in the Prisoner's Dilemma game. *Journal of Theoretical Biology* 420 (2017), 12–17. https://doi.org/10.1016/j.jtbi.2017.02.036