Defection at First Sight: Learning Partner Selection in Optional Social Dilemmas without Prior Information

Benedict Russell
University of Warwick
Coventry, United Kingdom
benedict.i.russell@warwick.ac.uk

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 the option of opting out of the interaction and instead being randomly assigned to a new partner in the population. In contrast to previous work, we remove the assumption that agents know the previous move of every other agent in the game, even when not directly interacting with them. Instead, agents adopt interaction-length dependent policies: in the first round they act without any information about their opponent, whilst observed behaviour informs the choice of action in subsequent rounds. Using multi-agent reinforcement learning, we show that cooperation can emerge and be sustained without any hard-wired partner selection mechanisms. In the initial interactions, agents learn to defect before adopting reciprocal strategies such as Tit-for-Tat, what is known in the literature as the "hazing period". Interestingly, agents learn to unconditionally stay in initial interactions, before adopting known cooperation-promoting partner selection rules like Out-for-Tat, leaving defectors and staying with cooperators, in subsequent rounds. Finally, we show how this scales when agents adopt longer interaction-length dependent policies.

KEYWORDS

Social Dilemmas, Partner Selection, Emergence of Cooperation

ACM Reference Format:
Benedict Russell, Chin-wing Leung, and Paolo Turrini. 2026. Defection at First Sight: Learning Partner Selection in Optional Social Dilemmas without Prior Information. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/IBSZ1473

1 INTRODUCTION

In social dilemmas, a population of individuals are deciding between a cooperative choice, e.g., contributing to a common project, and a non-cooperative one, e.g., exploiting the contributions of others. While cooperating is socially desirable, the temptation not to do so is often stronger. Understanding the conditions under which self-interested individuals make cooperative choices is a question that fascinated scientists across many disciplines, from economics to evolutionary psychology and biology [15]. In well-mixed populations, where individuals are paired uniformly at random, it is well known that the population choices converge to defection unless some external mechanism is in place [6, 15]. In particular, the capacity of individuals to choose their partner was identified as a key mechanism to avoid population-wide defection [15]. Partner selection enables the ability to maintain profitable and stable relationships, and walk away from exploiters.

The longstanding interest in cooperation and coordination, coupled with the recent advances in reinforcement learning, have made the problem increasingly central for the multi-agent systems community [9]. Can self-interested learning agents learn to go beyond individual gains and act for a greater good? Extensive computational work has linked population-level outcomes to simple adaptive rules, showing that reinforcement learning agents can effectively reproduce evolutionary dynamics [5, 6]. However, this simply means that independent reinforcement learners playing social dilemmas will end up exploiting each other.

A number of contributions in multi-agent reinforcement learning have shown that giving agents partner selection capabilities - where they track the value of potential partners and choose who to play with - induces them to learn cooperative behaviour [1, 12, 13]. Giving agents as little as the capacity to opt out of the interaction and be randomly reassigned to someone else instead is sufficient for cooperation to prevail [13]. In these "optional" social dilemmas, learning agents reciprocate at the game level - learning the Tit-for-Tat strategy, which copies the partner's last action - and at the partner selection level - learning the Out-for-Tat strategy, which breaks ties with defectors and keeps them with cooperators.

These results are however based on the strong assumption that agents know the past behaviour of their currently assigned partners, even at the start of the interaction [12, 13]. In many real-world scenarios, though, such informational freedom does not exist. In anonymous settings with many players, an agent can defect before switching partners and avoid retaliation. To prevent this, recent work in sequential games with trigger-restarts has suggested treating earlier moves differently [4]. Optimal strategies where actions depend on the interaction-length start with a 'hazing period', where defections early on discourage later exploitation. Whilst this avoids the pitfall of pure defection, the theoretical guarantees are limited to two-agent scenarios and rely on a rule-based restart mechanism.

In this paper, we remove the restriction to two agents and the fixed restart mechanism. Instead, agents learn to play the sequential game with an additional partner selection strategy at each step. In tandem, we do not assume any informational transfer when new pairs are formed. Under this interaction model, where agents adopt interaction-length dependent policies, we show how cooperation can emerge and be sustained.

Contribution. We study a population of learning agents playing the repeated Prisoner's Dilemma, where new interactions begin

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/IBSZ1473

with no prior information on the opponent. As the interaction progresses, agents can respond to the previous action of the opponent and, at the same time, decide whether to end their current interaction and be randomly rematched to another agent in the population. We allow agents to adopt an interaction-length-dependent policy, where they make decisions based on both the number of games they have played repeatedly with their partner and their partner's last action. Using multi-agent Q-learning, we show that cooperation can emerge and be sustained without any hard-wired partner selection or trigger-restart mechanisms. In the initial interactions, agents learn to defect before adopting reciprocal strategies such as Tit-for-Tat, what is known in the literature as the "hazing period". Interestingly, agents learn to unconditionally stay in initial interactions, before adopting known cooperation-promoting partner selection rules like Out-for-Tat, leaving defectors and staying with cooperators, in subsequent rounds. Finally, we show how this scales up to agents with longer interaction-length dependent policies.

Related Literature. Mechanisms which promote socially desirable behaviour in multi-agent systems have long been examined, with approaches ranging from centralised ones such as social laws [24] to decentralised ones such as trust [7] and reputation [17, 19, 20, 22]. Without such signals, agents are left in the dark about their opponents' past behaviour and defection prevails [8].

Even with such behavioural cues, random matching can prevent the emergence of cooperation [1]. As such, research has focused on how partner selection can be used as a mechanism to promote cooperation, yielding results both empirically [3, 32] and theoretically [2, 10, 23, 26, 31, 33]. This ability to break and form ties has been identified as one of five canonical ways to enable the emergence of cooperation [15]; and has been used to examine cooperation on networks [21], coordination problems [14], and ostracism [16]. Human experiments show that these dynamic networks and partner updates facilitate a higher level of cooperation, creating assortative interaction patterns and exclusion of defectors [18, 29].

One recent study showed that self-interested reinforcement learners were able to co-learn cooperation-inducing partner selection whilst playing a repeated social dilemma [13]. Having access to their opponent's previous move, agents predominantly learn the Tit-for-Tat strategy coupled with Out-for-Tat: responding to defectors by breaking ties and finding a new partner. The ability to observe a new partner's actions against previous players is often assumed [32, 33], but this can be unrealistic in many real-world scenarios. In this paper, we study the emergence of cooperation where agents do not have any prior information on their opponent.

Formation of equilibria under longer memory-length has been studied for two-player symmetric games [28], and under unanimous agreement over entire histories [11]. The action sequences induced by sequential strategies have been formalised, showing optimality and stability under a trigger-restart mechanism [4] in the two-player game. In particular, this work highlights that agents need to treat early interactions differently from sustained ones. The optimal sequence of actions includes a 'hazing period', where agents defect in initial interactions before cooperating. In this paper, we utilise this interaction-length dependence whilst lifting the restrictions to 2 agents and the fixed trigger-restart mechanism. Instead, agents extend the interaction-length dependent policy for the Prisoner's Dilemma to also include partner selection. The study of evolving

	C	D
C	R, R	S, T
D	T, S	P, P

	C	D
C	3, 3	0, 5
D	5, 0	1, 1

Figure 1: General payoff matrix (left) and instantiation (right) for the Prisoner's Dilemma game. The payoff is structured such that $T > R > P > S$ and $2R > T + S$.

populations where strategies include partner selection opens a richer class of behaviour, and is important to understanding the emergence of cooperation where environmental information is limited and without external structure. Additionally, this opens an avenue of research in how learning populations change behaviour depending on the interaction-length.

Paper Structure. Section 2 provides the necessary game-theoretic and Q-learning background. Section 3 presents the experimental setup and analysis while Section 4 delves into the emergent co-evolutionary structure; we discuss future directions in Section 5.

2 PRELIMINARIES

We first introduce sequential social dilemma games with the option of opting out and then outline the chosen Q-learning algorithm.

2.1 Sequential Social Dilemmas with Opting Out

We consider populations of self-interested agents paired to play a repeated Prisoner's Dilemma (PD). In each round, agents choose an action from $A = \{C, D\}$, where C and D denote cooperation and defection respectively. The players receive a payoff for the joint action profile of the pair. The general form and an example are shown in Figure 1, where the first entries denote the row player's payoff, and the second the column player's.

In the one-shot game, defection is the only Nash equilibrium. As such, populations updating their policies with the replicator equation quickly converge to this unique evolutionarily stable strategy [6]. This outcome is in spite of mutual cooperation, (C, C), being the social-optimal strategy, maximising the total reward for all players. In the repeated game, the strategy space induces a richer set of strategies which enable a wider range of outcomes. The most notable of these are Tit-for-Tat (TFT), which replicates the action of their opponent's previous action, and simple strategies such as Always-Cooperate (ALL-C) and Always-Defect (ALL-D). In the context of this paper, social dilemmas with the option of opting out (SDOOs) [13] follow:

(1) Players are randomly matched in pairs.
(2) Agents play a Prisoner's Dilemma with their partner, and receive a reward.
(3) Each player can then choose to continue playing with their current partner or break ties. If ties are broken, both agents become available.
(4) Available players are randomly matched in pairs. The process repeats from step (2).

Once an agent is paired with a new partner, they will play at least one Prisoner's Dilemma before they can choose to keep or break ties. If ties are broken and the pair is the only two agents available, they will be re-paired with the same partner but without information on

the previous action. This preserves the agent anonymity structure associated with this setup.

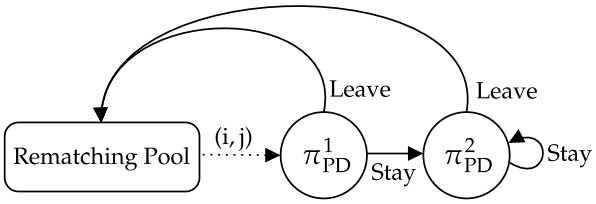


Figure 2: Illustration of the sequential game where agents choose actions in the Prisoners Dilemma (PD) and Partner Selection after each game. Initial interactions (π^1_{PD}) are made without any information on the opponent, whereas subsequent rounds (π^2_{PD}) benefit from experience. The final strategy is used for interactions which last more than 2 games.

In the sequential version of this repeated game, agents adopt an interaction-length dependent strategy. This builds upon theoretical analysis where agents in a 2-player game choose a designated sequence of actions and restart if there is any deviation between the two [4]. In this paper, we use the idea of interaction-length dependent actions but remove the assumptions of a fixed restart mechanism, allowing a population of agents to adopt both in-game and partner selection strategies. As a consequence, agents do not necessarily face the same partner in their next interaction. We also adopt the notion of a finite sequence, where the final policy is used for all interactions above a certain length. This repeated behaviour is shown to emerge theoretically in the sequential game, and allows for tractability in learning populations.

Figure 2 illustrates the Markov process for the agents. Of note, the first interaction being information-free prevents network retaliation. Strategies such as Tit-for-Tat will be felt in the classical sense against their current opponent; when a new partner is selected, any prior interactions with the previous opponent are forgotten.

In tandem with these responsive action strategies, partner selection plays a pivotal role. Shown to induce cooperation when enforced [32, 33], Out-for-Tat (OFT) denotes leaving when the partner defects and staying otherwise. Always-Stay (Stay) and Always-Defect (Switch) unconditionally keep and break ties respectively.

2.2 Q-learning

Q-learning is a common reinforcement learning algorithm, used in the context of a Markov Decision Process (MDP) [30]. Consider the tuple $G = \langle S, A, T, R \rangle$, where S is the set of states, A is the set of actions, T is the state transition function, and R is the reward function. Each agent stores a Q-value for every state-action pair (s, a), which is updated at each step of the episode according to Equation (1). The update rule aims to approach the optimal value, Q^*, which maximises the expected discounted future rewards. Let H be the episode length, then for $h \in [1, H]$ the discounted future reward is given by

$$G^h = \sum_{j=0}^{H-h} \gamma^j r_{h+j},$$

where $\gamma \in (0, 1)$ is the discount rate (how much an agent values the future state). We consider the constant-step size Monte Carlo control update rule [25], given by

$$Q(s_h, a_h) \leftarrow Q(s_h, a_h) + \alpha[G^h - Q(s_h, a_h)] \tag{1}$$

where $\alpha \in (0, 1)$ is the learning rate. Q-learning has been shown to converge to the optimal policy, provided infinitely many visits to each (s, a) tuple and certain learning rates [27, 30]. The Q-values do not themselves state which action to choose at a given time step; instead, we can use a mechanism which favours actions with higher values whilst still allowing for exploration. Boltzmann exploration is commonly used for this purpose, with the probability of choosing action a in state s at time t given by

$$\pi^i_t(s, a) = \frac{e^{\tau Q^i_t(s,a)}}{\sum_{b \in A} e^{\tau Q^i_t(s,b)}}, \tag{2}$$

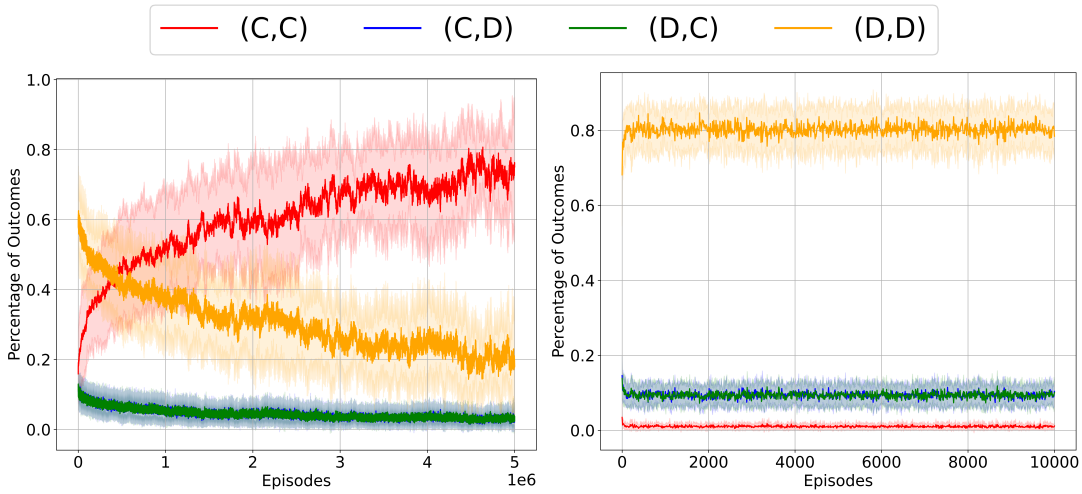
where the parameter τ is known as the inverse of the temperature. The agent is in pure exploration when τ is 0, and in pure exploitation when $\tau \to \infty$.

3 SEQUENTIAL CO-EVOLUTIONARY MODEL

3.1 Population and Game Structure

Consider a population of $N = 20$ agents learning to play the Prisoner's Dilemma shown on the right of Figure 1, with agents are randomly paired initially. In each round, agents choose an action for the Prisoner's Dilemma game. They are then able to respond to their opponent's action through partner selection, with the option to leave their current opponent and be randomly re-paired to an available agent. Therefore across an episode of length H, there are exactly $2H$ actions for each agent, with half deciding to cooperate or defect and half deciding to stay or switch. The social optimal solution is achieved in all agents cooperate with each other over these H rounds. For this to happen, agents must set aside the immediate reward associated with defecting against a cooperative opponent. The possible retaliation of their opponent severing the connection acts as a potential deterrent, with the agent not knowing who they will be paired with in the next round.

Playing without Prior Information. The informational restriction for agents in a new pairing plays a pivotal role in the actions taken. In their first interaction, agents play with Prisoner's Dilemma with no prior information on their opponent. In effect, the agent is in a stateless environment, with only the option to cooperate or defect. Players must balance indicating cooperative behaviours with the possibility of exploitation, without any signal of their partners' behaviour. Subsequent rounds store distinct Q-values for the length of interactions and allow for the direct reciprocity mechanism to evolve. Explicitly, the state-action value $Q(s^m, a)$ for interaction length m parametrises the action policy for all interactions after the initial game. This enables agents to behave differently for longer interactions, with the threat of restarting ensuring that earlier policies will be repeated. To maintain tractability, m is bounded by some constant M (the decision space). Any interactions which extend beyond this will utilise this final policy until a new interaction is formed. This generalises the current restriction to $M = 1$ strategies such as TFT, and removes the informational requirement in the first round of interaction. Partner selection benefits from experienced



(a) Option of opting out **(b) Random Matching**

Figure 3: Mean and standard deviation of outcome percentage over 20 simulations. The introduction of the partner selection stage is key to the emergence of cooperation. The percentage of cooperative outcome (C, C) starts to take over as the level of defection decreases. In the case of random matching, agents quickly learn defection.

information: agents first play a Prisoner's Dilemma, and therefore are always able to gauge some behavioural information of their opponent. This policy follows the same structural setup as the game actions, except all interaction-length-dependent policies are stated.

Formalising the Markov Process. The corresponding Markov decision process (MDP) has $4M-1$ states, $\mathcal{S}$, separated into

$$\mathcal{S}_{PD} := \{PD^1\} \cup \{PD_C^m, PD_D^m : m \in 2, ..., M\},$$
$$\mathcal{S}_{PS} := \{PS_C^m, PS_D^m : m \in 1, ..., M\}.$$

The first $2M-1$ states denote the information states used for the Prisoner's Dilemma, with PD^1 denoting the information-less first interaction. Likewise, the final $2M$ states are used for the partner selection phase of each round. In practice, agents adopt two separate policies for action selection. The first, π_{PD}, gives the action probabilities for cooperating and defecting given a state s_{PD} which embeds the interaction length. The second, π_{PS}, gives the probability of switching partners given the state s_{PS}. To find the minimal requirements for cooperation to emerge, the experiments will consider $M = 2$ with the effect of larger values discussed later. Each round these probabilities are used to select the two actions a_{PD} and a_{PS}, which for each episode generates a trajectory: $\tau = \{s_{PD}, a_{PD}, r_{PD}, s_{PS}, a_{PS}, r_{PS} = 0, ...\}$. At the end of the episode, agents update their policy according to (1), with the learning rate $\alpha = 0.05$ and the discount rate $\gamma = 1$. All agents adopt Boltzmann action selection (2) with $\tau = 1$. The algorithm pseudocode is provided in the supplementary materials document.

3.2 Emergence of Cooperation

Despite agents having no prior information on initial interactions, cooperation can still emerge in sequential SDOOs when partners are able to break ties based on the action of their opponent. As Figure 3 shows, mutual cooperation becomes the dominant outcome in the population of self-interested agents. The strategies in the population can be analysed to reveal how this cooperation emerges. In comparison with the literature where a new opponent's previous action is observable [13], the complexity of strategy space induced by M increases the number of phase transitions in population behaviour. These transitions are as follows (episode number denoted in brackets):

- Phase 1 (0-100): Agents quickly learn to defect in their first interaction ($m = 1$). Agents realise leaving their partner will result in their new partner playing defection, and adopt the Stay and R-OFT strategies for $m = 1$.
- Phase 2 (100-10,000): Agents' second interaction ($m = 2$) starts to shift towards Defection, with the corresponding partner selection policy following OFT and Switch.
- Phase 3 (10,000-100,000): OFT in interactions with $m = 2$ becomes the dominant partner selection rule. In the first round agents are predominantly using Stay and R-OFT.
- Phase 4 (100,000-1mil): With this partner selection in place, All-C and TFT become the presiding strategies for $m = 2$, taking over All-D.
- Phase 5 (1mil-): Cooperation stabilises, with the population largely using defection-stay in initial interactions, and a mix of All-C and TFT in subsequent rounds. Partner selection for these interactions mostly follows OFT and Stay.

Emergent Policy Types. Following a similar analysis to that in the literature [13], we analyse the strategy types of individuals in the population by comparing the difference in their Q-values for each state and action pair. Explicitly, for a given state s and possible actions a and b, if $Q^i(s,a) > Q^i(s,b)$ we classify this agent i as playing a in state s. For example, given the option to cooperate (C) or defect (D), we would label agent i as TFT agent if $Q(PD_C, C) > Q(PD_C, D)$ and $Q(PD_D, C) < Q(PD_D, D)$. This enables a full classification for

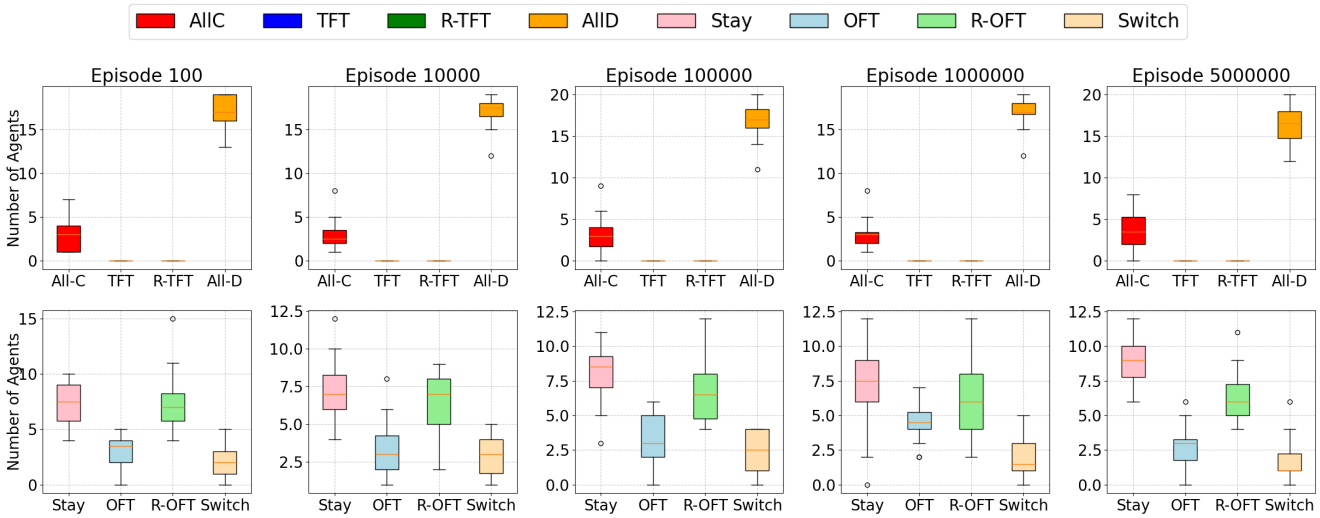


Figure 4: Box plots of the number of agents that use each strategy for the initial interaction ($m = 1$) during different phase transitions. Agents quickly adopt defection in their first interaction with new partners. For partner selection, agents predominantly learn Stay and R-OFT, indicating a preference for similar partners.

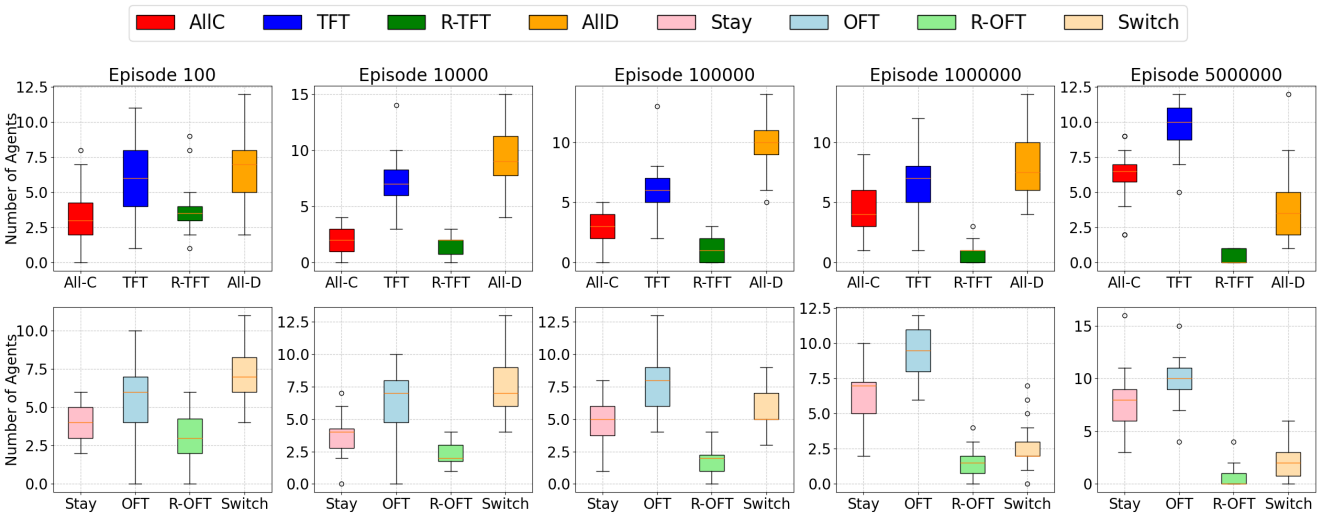


Figure 5: Box plots of the number of agents that use each strategy for the subsequent interactions ($m = 2$) during different phase transitions. Agents adopt the OFT strategy, which leads to the emergence of TFT and All-C.

all Q-value combinations. For the Prisoner's Dilemma, the types are: (1) Always Cooperate (All-C), (2) Tit-for-Tat (TFT): copy the previous action of your opponent, (3) Reverse Tit-for-Tat (R-TFT): play the opposite of your opponent's previous action, and (4) Always Defect (All-D). Likewise for the partner selection phase, the types are given by: (1) Always Stay (Stay), (2) Out-for-Tat (OFT): if they defect, break ties, (3) Reverse Out-for-Tat (R-OFT): if they cooperate, break ties, and (4) Always Leave (Switch). This deterministic interpretation does not include the stochastic effects of action selection, meaning simulations will demonstrate differences. This does, however, enable a deeper understanding of how the population is evolving. To this end, we plot: i) a box plot indicating the number of agents playing each strategy for both interactions (Figures 4, 5), ii) a bar chart illustrating the combined strategy frequency for each interaction (Figure 6), and iii) a line chart showing the number of partner switches over time (Figure 7).

Phase Transition Analysis. In the initial phase, agents quickly learn to defect when they have no access to the previous action of their opponent ($m = 1$). Subsequent interactions ($m = 2$) show a similar yet smaller shift towards All-D. Agents learn to avoid the

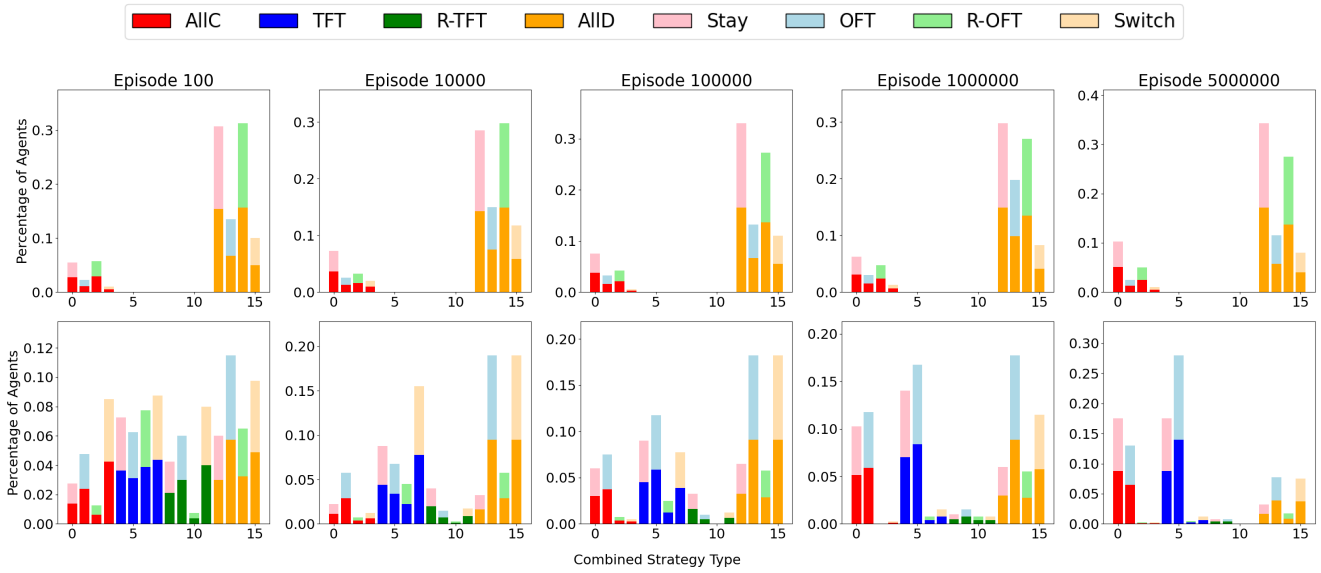


Figure 6: Percentage of agents adopting each combined strategy during different phase transitions. The top diagram illustrates the initial interaction policy ($m = 1$), and the bottom all subsequent interactions ($m = 2$). Nearly 70% adopt a (All-D, Stay) or (All-D, R-OFT). For future rounds, agents learn a mix of All-C and TFT, combined with Stay and TFT.

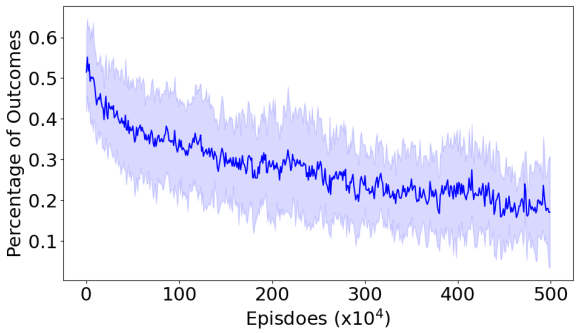


Figure 7: Mean and standard deviation of the percentage of agents who switch partners in the partner selection stage, by episode. A sharp drop at the beginning is followed by a steady decline, stabilising around 25%.

almost guaranteed defection faced when meeting a new partner. The partner selection policies for $m = 1$ shift towards Stay and R-OFT. The former allows for their next interaction to be conditional on their partner's information and play against a less defective opposition. The early adoption of R-OFT presents a fascinating idea: if the opponent cooperates, you defect and gain the selfish reward before re-pairing in the population. The benefits of this are two-fold: firstly, the agent does not suffer from possible relation in future rounds, and secondly, there is a chance to be re-paired with a cooperator who will not observe this defection and exploitation can be repeated. As a result, there is a sharp fall in the number of partner switches in this phase, shown in Figure 7. For longer interactions ($m = 2$), TFT and All-D begin to emerge, with Switch acting as the main partner selection choice.

In phase two, the threat of defection with new partners and the informational gain of staying increase the frequency of TFT and All-D agents for interactions with $m = 2$. The agents begin to co-learn OFT and Switch for these interactions, the first sign of stable cooperative behaviours. In phase three, the OFT strategy is predominant for $m = 2$, marking a change from the switching preference in phase three. The corresponding game policy begins to show some increase in adoption on All-C. Phase four demonstrates the last major shift, with the PD policy shift away from defection for interactions past the initial pairing ($m = 2$). The level of All-C and TFT grows to become the two dominant strategies in the population. In the final phase, cooperation levels reach 75% at episode 5mil, stabilising around this value for future episodes. The PD population composition is largely defection in the first interaction followed by All-C and TFT. For partner selection, Stay is used in initial interactions and a mix of Stay and OFT is adopted for subsequent rounds. The behaviour in these continuing interactions replicates what found to emerge when information is always present [13].

Experience Drives Cooperation. The interaction-length dependent policy mechanism is paramount for cooperation to emerge in these SDOOs. The majority of agents learn to defect against new partners and later choose a cooperation-inducing strategy such as All-C and TFT. This game strategy goes hand-in-hand with the partner selection choice; the likely defection an agent will face if they are re-paired pushes agents to learn to Stay or R-OFT after initial interactions. It is not surprising that agents adopt these strategies which allow for future information. The OFT mechanism in subsequent interactions becomes dominant, which stays with cooperators and leaves defectors. Clearly, this is beneficial

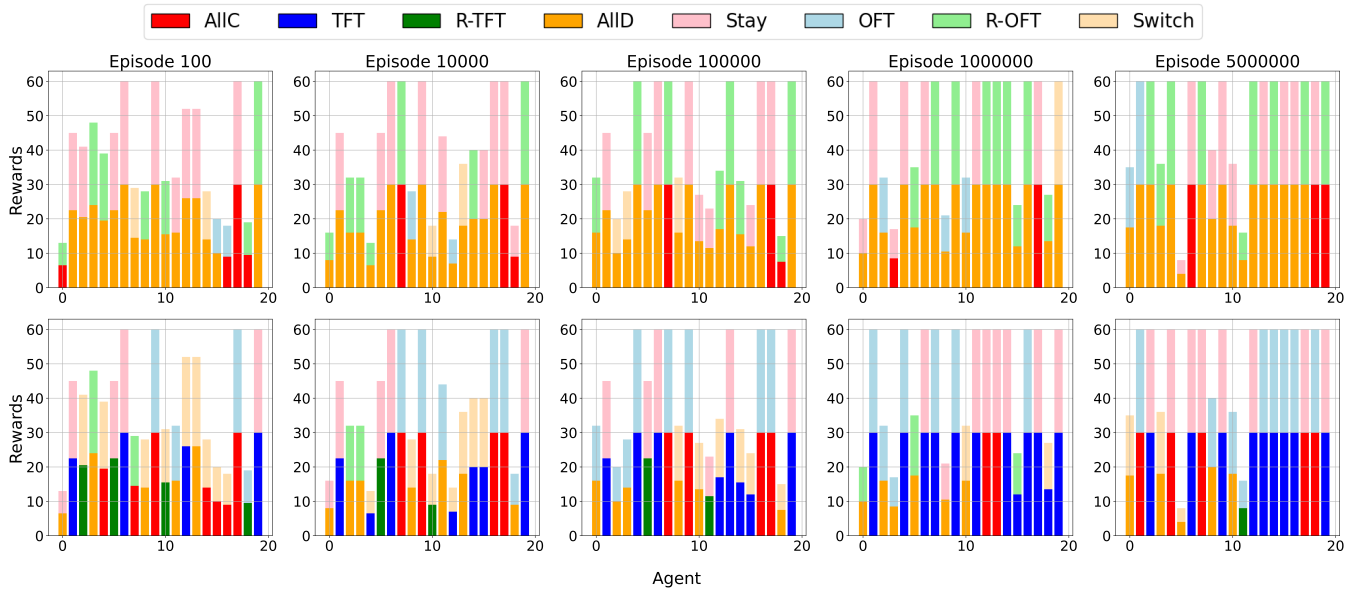


Figure 8: The episodic reward and strategy of each agent over different phase transitions; $m = 1$ and $m = 2$ are shown in the top and bottom rows respectively. Agents almost unanimously learn to defect in the first round and stay with their partner. The OFT and Stay strategies in the subsequent partner selection stage allow for the TFT and All-C agents to find and stick to each other, enabling sustained cooperation in spite of a few defectors.

to any strategy type as mutual cooperation or exploitation can be maximised. Coupled with TFT, agents become matched in a mixture of All-C and TFT pairs, with the OFT mechanism ensuring a cooperative partner is found.

4 CO-EVOLUTIONARY STRUCTURES

We turn our attention to look at how the specific combined type of an agent can evolve in these environments. Analysing this evolution and corresponding rewards provides a more insightful depiction of the process in which cooperation emerges. We plot the episodic rewards for each agent across different episodes of a representative simulation in Figure 8, where the colours show the combined PD and PS policy types, the top row shows policies for the first encounter, and the bottom subsequent interactions.

Cooperators Pairs Ostracise Defectors. By episode 100, there is an emergence of the (All-D, Stay) and (All-D, R-OFT) types for initial interactions. These initially defecting partners will stay as a pair and use their subsequent strategies. The majority of these strategies have not yet matured, with partner selection actions causing switching to be prevalent. Even at this early stage, four agents have achieved an episodic reward of 60, which indicates fully cooperative actions across the 20-round episode. These agents have adopted the (TFT, Stay) and (All-C, OFT) strategies for $m = 2$, enabling the generation of stable cooperative pairings. By episode 10000, the strategy type for initial interactions is largely unchanged. However, the rewards are higher with more agents learning the above two cooperative strategies in longer interactions. By episode 100,000, this trend is strengthened with R-OFT eliminated from the population for $m = 2$. This is also the first instance in which (TFT, OFT) emerges as a potential strategy. By episode 5 million, all defecting strategies are heavily punished: all cooperative agents are paired together, and the defectors can only exploit each other.

Hazing on a Network. The sequential nature of the game induces a 'hazing period', whereby agents must spend valuable time facing defection whilst finding a new cooperative agent to either stay with or exploit. This extends the traditional notion of 'hazing', in which two agents are faced with repeating the initial actions if they deviate [4]. Instead, the population variance and responsive partner selection strategies mean leaving a partner has a greater consequence than just resetting the current interaction. In fact, in the repeated PD which we are considering, no cooperative equilibria exist for sequences of length 2 under the restart mechanism. The addition of partner selection strategies in a multi-agent environment allows agents to learn and sustain cooperative outcomes which cannot be guaranteed in the rule-based restart setting.

Increasing Strategic Freedom Does not Increase Cooperation. Until now we have focused a repeated PD where initial interactions are information-free, whilst subsequent rounds adopt conditional strategies. This secondary policy is used for all subsequent rounds where the interaction continues. The analysis above shows cooperation can emerge when the decision space M is restricted to 2. Naturally, this raises the question of how increasing M will affect the learning dynamics and outcome.

We let $M = 3$, allowing for three interaction-length dependent strategies. The first is used for new interactions, and preserves the information-less state. The subsequent policies allow agents to respond not only to their opponents' actions here, but also to their behaviour in future rounds. Figure 9 shows how cooperation can still become the dominant outcome with (C, C) reaching around 60%

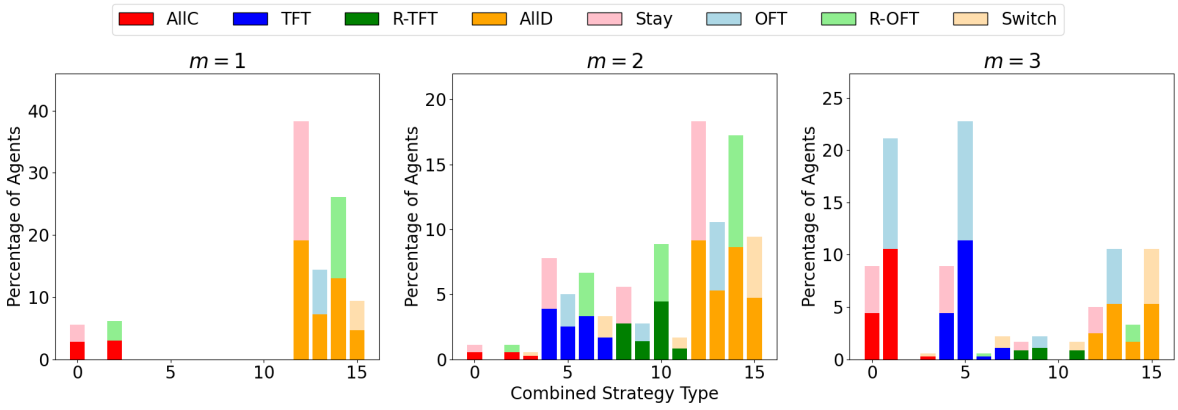


Figure 9: Percentage of agents adopting each combined strategy for each interaction length, m, after 10 million episodes. The left diagram illustrates the initial interaction policy ($m = 1$), where defection is prevalent. Agents learn a similar strategy for interactions which make it to length two ($m = 2$). Interactions which last 3 or more rounds demonstrate more cooperative behaviour with the majority employing TFT or All-C ($m = 3$).

after 10 million episodes. This is notably lower than when strategic freedom is restricted to $M = 2$. The answer lies in how increasing complexity can be detrimental to learning social-optimal behaviour. In the context of SDOOs, the increase in strategy variance and frequent switching prevents the population from distinguishing between strategies which induce stable long-term reward and those which are exploitable. For example, whilst agents learn to defect in initial interactions ($m = 1$), their policy does not replicate the previous behaviour for $m = 2$. Instead, players mostly learn that defection is the best outcome option. This does serve as a deterrent: any interaction which gets to length $m = 3$ must have sufficiently high reward to merit not restarting and facing repeated defection. This shows that agents are learning a form of hazing, where mutual defection is followed by a period of cooperation.

5 DISCUSSION

We have conducted analysis on learning agents in repeated social dilemmas with the option of opting out. This extends theoretical guarantees in two-player sequential games, equipped with the trigger restart mechanism. We relax the constraints of a fixed partner selection rule, extending the interaction-length dependence to include a partner selection policy at each interaction length. In initial interactions, agents must act without any information about their opponent. Despite this restriction, we observe the emergence and persistence of cooperation in multi-agent populations. Agents learn to defect when they first meet an opponent, but choose to stick with them to build a cooperative relationship. In these longer relations, agents forgive the actions of initial defection, playing a mixture of All-C and TFT which can be paired for stable cooperation. Agents also endogenously acquire the Out-for-Tat partner selection rule, which is notably equipped when the strategy All-C is attempted, preventing permanent exploitation. These insights show that cooperation does not hinge on pre-existing knowledge of an opponent: it can be learn from scratch through first-move actions and subsequent conditional responses. Of particular note is how partner selection is used to find and maintain pro-social connections. Expecting defection if ties are broken, agents learn two strategies: Stay and R-OFT. The choice to stay is fundamental to the development of future cooperative strategies, whilst R-OFT exploits initial cooperators in the population and ensures any lasting partnership starts on equal footing. In the following interactions, OFT is dominant amongst cooperative agents, which is particularly difficult to emerge given that any defection by the opponent leads to random re-pairing and likely defection in the next round. The evidence suggests that these secondary strategies form once the initial interactions have been solidified, allowing consistent value approximations for actions in this state. The fact that this can arise - without starting new relations with knowledge - suggests that this is a particularly robust route for cooperation to emerge.

Whilst we have shown the emergence of a predominantly cooperative populous, there remain questions as to the stability and level achieved. One main avenue concerns emergence in longer sequential games. In Section 4, we extended the policy space to include 3-step policies, revealing that under the same conditions as $M = 2$, a lower level of cooperation is achieved. This begs the question of whether analytical studies are possible of repeated games with restarts - in the spirit of [4] -where agents learn partner selection strategies by themselves. In the same vein, deeper sensitivity analysis is required to see if a similar cooperation level could be achieved, and if longer policies could be more robust to errors. Finally, this framework poses new and interesting questions around human behaviour in a similar environment. Faced with SDOOs, humans have already been shown to develop specific cooperation rates [32]; a similar study could extend this to mixed observability. There are three key outcomes which would replicate the behaviours we found in Q-learning populations: i) initial defection, ii) forgiveness and responsive strategies in future rounds, and iii) length-dependent partner switching. Finding and understanding the necessary conditions under which human behaviour is replicated would provide a fascinating stream of research.

ACKNOWLEDGMENTS

BR was supported by the Engineering and Physical Sciences Research Council through the Mathematics of Systems II Centre for Doctoral Training at the university of Warwick (reference EP/S022244/1). CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG 2023-050.

REFERENCES

[1] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20).* https://aaai.org/Library/conferences-library.php Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20).

[2] 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

[3] Pat Barclay and Robb Willer. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences* 274, 1610 (2007), 749–753.

[4] Ratip Emin Berker and Vincent Conitzer. 2024. Computing optimal equilibria in repeated games with restarts. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI '24).* Jeju, Korea, 2669–2677. https://doi.org/10.24963/ijcai.2024/295

[5] 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://api.semanticscholar.org/CorpusID:13171013

[6] 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

[7] Cristiano Castelfranchi. 1998. Modelling social action for AI agents. *Artificial Intelligence* 103, 1 (1998), 157–182. https://doi.org/10.1016/S0004-3702(98)00056-3 Artificial Intelligence 40 years later.

[8] Ross Cressman. 1996. Evolutionary Stability in the Finitely Repeated Prisoner 's Dilemma Game. *Journal of Economic Theory* 68, 1 (1996), 234–248. https://doi.org/10.1006/jeth.1996.0012

[9] Shaheen Fatima, Nicholas R. Jennings, and Michael J. Wooldridge. 2024. Learning to Resolve Social Dilemmas: A Survey. *J. Artif. Intell. Res.* 79 (2024), 895–969. https://doi.org/10.1613/JAIR.1.15167

[10] Christopher Graser, Takako Fujiwara-Greve, Julián García, and Matthijs Van Veelen. 2025. Repeated games with partner choice. *PLOS Computational Biology* 21, 2 (2025), e1012810.

[11] Christian Hilbe, Luis A. Martinez-Vaquero, Krishnendu Chatterjee, and Martin A. Nowak. 2017. Memory-n strategies of direct reciprocity. *Proceedings of the National Academy of Sciences* 114, 18 (2017), 4715–4720. https://doi.org/10.1073/pnas.1621239114

[12] Chin-wing Leung, Tom Lenaerts, and Paolo Turrini. 2024. To Promote Full Cooperation in Social Dilemmas, Agents Need to Unlearn Loyalty. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024.* ijcai.org, 111–119. https://www.ijcai.org/proceedings/2024/13

[13] Chin-wing Leung and Paolo Turrini. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out *(AAMAS '24).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1110–1118.

[14] Chin-wing Leung, Paolo Turrini, and Ann Nowé. 2025. Curiosity-Driven Partner Selection Accelerates Convention Emergence in Language Games. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems* (Detroit, MI, USA) *(AAMAS '25).* International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1282–1290.

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

[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] 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 1st International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02).*

[18] 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

[19] Tianyu Ren, Xuan Yao, Yang Li, and Xiao-Jun Zeng. 2025. Bottom-Up Reputation Promotes Cooperation with Multi-Agent Reinforcement Learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'25).*

[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] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. 2006. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLOS Computational Biology* 2, 10 (10 2006), 1–8. https://doi.org/10.1371/journal.pcbi.0020140

[22] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. 2018. Social Norms of Cooperation With Costly Reputation Building. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18).*

[23] Sven Van Segbroeck, Francisco C. Santos, Ann Nowé, Jorge M. Pacheco, and Tom Lenaerts. 2009. The coevolution of loyalty and cooperation. In *Proceedings of the 11th IEEE Congress on Evolutionary Computation (CEC'09).*

[24] Yoav Shoham and Moshe Tennenholtz. 1995. On social laws for artificial agent societies: off-line design. *Artificial Intelligence* 73, 1 (1995), 231–252. https://doi.org/10.1016/0004-3702(94)00007-N Computational Research on Interaction and Agency, Part 2.

[25] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement learning: An Introduction.* MIT press.

[26] Karolina Sylwester and Gilbert Roberts. 2010. Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters* 6, 5 (2010), 659–662.

[27] John N. Tsitsiklis. 1994. Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning* 16 (1994), 185–202. https://api.semanticscholar.org/CorpusID:4449439

[28] Masahiko Ueda. 2023. Memory-two strategies forming symmetric mutual reinforcement learning equilibrium in repeated prisoners' dilemma game. *Appl. Math. Comput.* 444 (May 2023), 127819. https://doi.org/10.1016/j.amc.2022.127819

[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 J. C. H. Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* 8, 3 (may 1992), 279–292. https://doi.org/10.1007/BF00992698

[31] Toshio Yamagishi, Nahoko Hayashi, and Nobuhito Jin. 1994. Prisoner's dilemma networks: Selection strategy versus action strategy. In *Social Dilemmas and Cooperation*, Ulrich Schulz, Wulf Albers, and Ulrich Mueller (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 233–250.

[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