To Promote Full Cooperation in Social Dilemmas, Agents Need to Unlearn Loyalty

Chin-wing Leung [1], **Tom Lenaerts** [2,3,4] and **Paolo Turrini** [1]
[1]Department of Computer Science, University of Warwick
[2]Machine Learning Group, Université Libre de Bruxelles
[3]Artificial Intelligence Lab, Vrije Universiteit Brussel
[4]Center for Human-Compatible AI, UC Berkeley
chin-wing.leung@warwick.ac.uk, tom.lenaerts@ulb.be, p.turrini@warwick.ac.uk

Abstract

If given the choice, what strategy should agents use to switch partners in strategic social interactions? While many analyses have been performed on specific switching heuristics, showing how and when these lead to more cooperation, no insights have been provided into which rule will actually be learnt by agents when given the freedom to do so. Starting from a baseline model that has demonstrated the potential of rewiring for cooperation, we provide answers to this question over the full spectrum of social dilemmas. Multi-agent Q-learning with Boltzmann exploration is used to learn when to sever or maintain an association. In both the Prisoner's Dilemma and the Stag Hunt games we observe that the Out-for-Tat rewiring rule, breaking ties with other agents choosing socially undesirable actions, becomes dominant, confirming at the same time that cooperation flourishes when rewiring is fast enough relative to imitation. Nonetheless, in the transitory region before full cooperation, a Stay strategy, keeping a connection at all costs, remains present, which shows that loyalty needs to be overcome for full cooperation to emerge. In conclusion, individuals learn cooperation-promoting rewiring rules but need to overcome a kind of loyalty to achieve full cooperation in the full spectrum of social dilemmas.

1 Introduction

In social dilemmas, individuals are faced with the choice of paying a cost to contribute to a common good or simply abstaining from doing so and reaping the benefits from the other contributors. The exact payoff structure of such dilemmas translates into well-known strategic games, such as the Prisoner's Dilemma, where the selfish action of defection is strictly dominant, but mass defection makes everyone worse off than they would have been by cooperating.

It is well known that in well-mixed populations mass defection is bound to take over under random matching without any added mechanism [Nowak, 2006], so the problem arises of how to modify the interaction so that individuals are driven towards socially desirable outcomes. The desire to form social relationships has long been recognised as a key enabler to promote cooperative behaviour [Eshel and Cavalli-Sforza, 1982] and "network reciprocity" was singled out as one of the five mechanisms for doing so [Nowak, 2006]. The possibility for partner selection, allowing agents the choice of who and how often to interact with, was shown to promote cooperation in numerous experimental studies [Rand *et al.*, 2011; Wang *et al.*, 2012; Zhang *et al.*, 2016] and computational models [Segbroeck *et al.*, 2009; Zheng *et al.*, 2017; Bara *et al.*, 2022; Pacheco *et al.*, 2006; Santos *et al.*, 2006a]. But while the emergence of cooperation-sustaining in-game strategies, such as the well-known Tit-for-Tat in the Prisoner's Dilemma game, is relatively well-understood, we do not have the same understanding of which partner selection rules will co-evolve with them and whether the learnt rules provide an advantage or not for cooperation.

In a seminal contribution [Santos *et al.*, 2006a] have shown that cooperation prevails when individuals adjust their social ties, proposing a partner selection rule that causes full cooperation to emerge. If individuals are tied with cooperators, they choose to stay connected, but if they are tied with defectors they switch to a random partner proportionally to a (Fermi) function of the fitness difference with the partners. This is a more general version of the Out-for-Tat (OFT) rule, which always breaks ties with defectors and keeps them with cooperators, shown to lead to (partial) cooperation in the Prisoner's Dilemma with the option of opting out when imposed on the agents [Zhang *et al.*, 2016]. But while such handcrafted rules have been shown to promote cooperation, it is not clear whether cooperation can be achieved when agents learn partner selection rules by themselves, and whether patterns emerge in learnt rules across different social dilemmas.

Contribution. In this paper, we study the emergence of partner selection rules and examine how they influence the strategic dynamics in social dilemmas played on networks, without imposing agents to follow pre-defined heuristics or even specifying that cooperation should be achieved. As such, our work reveals which assortment rules are found by evolution and whether they are bound to lead to cooperation (or not) [Eshel and Cavalli-Sforza, 1982]. We show that in both the Prisoner's Dilemma and the Stag Hunt games an OFT rewiring rule, which keeps ties cooperators and beaks them with defectors becomes dominant, allowing for coop-

eration to flourish provided rewiring is faster than imitation. The main bottleneck to achieving cooperation is a Stay strategy that tries to maintain the connection no matter what. Sufficiently fast rewiring removes this strategy, thus opening the door to full cooperation.

Related Research. Understanding why phenomena such as cooperation, reciprocity and altruism emerge from self-interested behaviour is one of the greatest challenges across many fields of science, such as biology, psychology and economics [Nowak, 2006]. The capacity of individuals to develop and maintain meaningful social ties is considered a key enabler in moving away from selfish behaviour, together with kin selection, direct reciprocity, indirect reciprocity and group selection [Nowak, 2006].

Partner selection introduces an extra dimension to the interaction. Not only do individuals think about how to behave in a social dilemma, but also who to play this social dilemma with. This naturally suggests the study of two-dimensional timescales, where the relative speed of structural and strategic update was shown to make a difference in the emergence of cooperation in models with interaction propensity [Pacheco *et al.*, 2006; Santos *et al.*, 2006b], the decision to leave from the current partner [Zhang *et al.*, 2016; Zheng *et al.*, 2017] or group [Santos *et al.*, 2006a], and unilateral attachment [Bara *et al.*, 2022].

The spatio-temporal aspects typical of games with partner selection add further complexity to the study of the evolutionary dynamics, as these are ignored in the replicator equation [Roca *et al.*, 2009], which makes standard Ordinary Differential Equation based analysis infeasible. Social simulation has often been employed to provide insights into the behaviour of complex societies assorted in networks [Roca *et al.*, 2009]. Using computer-aided analysis, partner selection was shown to be key to the emergence of cooperation in social dilemmas [Gilbert, 1995; Salazar *et al.*, 2011] and coordination games[Segbroeck *et al.*, 2010], and studied in combination with reputation [Sabater and Sierra, 2002; Pujol *et al.*, 2002; Sabater-Mir *et al.*, 2006] as a tool to ostracise unreliable partners [Perreau de Pinninck *et al.*, 2010; Wang *et al.*, 2012; Santos *et al.*, 2018]. In this paper, we are concerned with minimal enablers for cooperation, without assuming the ability of agents to communicate with one another or form evaluative meta-beliefs as in the reputation approaches.

Recently, Reinforcement Learning has emerged as a fundamental tool to investigate strategic interactions [Bloembergen *et al.*, 2015], building on the deep connection with replicator dynamics [Börgers and Sarin, 1997]. Rather than finding evolutionary stable strategies by "solving" a game, we can approximate population dynamics by having agents learn the distribution of types. The emergence of pro-social behaviour has received increased attention thanks to the development of deep reinforcement learning algorithms, for example in the context of common pool resource appropriation [Pérolat *et al.*, 2017]. Partner selection was also recently studied as an explicit strategy profile, modelling Q-learning agents engaged in a two-person Prisoner's Dilemma with direct and fully informed partner selection [Anastassacos *et al.*, 2020] where agents can unilaterally select who to play with, and have access to the past behaviour of every other agent in the game. In games on networks, [Fulker *et al.*, 2021] models the co-evolution of network weights, representing individuals' openness to interact, and in-game strategies while [Foley *et al.*, 2018] a co-evolutionary model is presented where strategy and structure evolve by reinforcement learning, but only able to account for the emergence of conventions, while social dilemmas require more complex learning approaches.

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

	C	D
C	$1, 1$	S, T
D	T, S	$0, 0$

Figure 1: Payoff bi-matrix for social dilemmas of cooperation. The relation between R, S, T, P will determine the exact nature of the game. Our analysis fixes $R = 1$ and $P = 0$.

Paper Structure. Section 2 presents the background on social dilemmas and Q-learning, Section 3 introduces our learning algorithm for partner selection and analyses the emergence of cooperation-sustaining partner selection, while Section 4 zooms in the specific social dilemmas.

2 Preliminaries

2.1 Social Dilemmas

Social dilemmas of cooperation are modelled as 2-player symmetric games $\langle A, M \rangle$, where players' action space $A = \{C, D\}$ denotes their action of cooperate or defect and M the payoff matrix for the row player (see Figure 1). If both players cooperate, they will receive a payoff of R (the reward). If both defect, they will receive P (the punishment). If one player cooperates and another one defects, the cooperator will receive a payoff of S (the disadvantage of being cheated on) and the defector will receive a payoff of T (the temptation to cheat). Depending on the values of the payoff, we can classify the game into three different social dilemmas, the Snow-Drift (SD) game, with $T > R > S > P$, the Stag-Hunt (SH) game, with $R > T > P > S$, and the Prisoner's Dilemma (PD), with $T > R > P > S$. They represent various degrees of conflict between players. For SD, unilateral defection is preferred to mutual cooperation while unilateral cooperation is better than mutual defection ($T > R, S > P$), therefore the pure Nash equilibria of the game are (C, D) and (D, C). On the other hand, for SH, mutual cooperation and mutual defection are preferred to unilateral defection and unilateral cooperation ($R > T, P > S$), and the pure Nash equilibria of the game are (C, C) and (D, D). The PD is the harder case for cooperation, as defection is the strictly dominant strategy ($T > R, P > S$), with the unique Nash equilibrium of the game being (D, D). We adopt the convention in [Santos *et al.*, 2006a] and normalize the payoff for mutual cooperation and mutual defection to $R = 1$ and $P = 0$, as shown in Figure 1. We are studying games played on networks, where agents play a social dilemma with their neighbours uniformly. For $\mathcal{N}(i)$ being the set of agents connected to i, the fitness (total

payoff) of agent i is given by

$$f(i) = \sum_{j=1}^{\mathcal{N}(i)} M(m_i, m_j) \tag{1}$$

where m_i and m_j are strategies by agent i and j, respectively.

2.2 Q-Learning

Our agents use Q-learning, a widely established Reinforcement Learning algorithm [Watkins and Dayan, 1992]. They act on a Markov decision process (MDP), represented as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} \rangle$, where $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of available actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a state transition probability function and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is an *immediate* reward function. A learning agent aims to find a policy $\pi(a|s)$ that maximises the expected discounted cumulative reward or *profit* $J = \mathbb{E}[\sum_{h=0}^{\infty} \gamma^h \rho_{h+1}]$ by repeated game plays, where $\gamma \in [0, 1]$ is the discount factor, and $\rho_{h+1} = \mathcal{R}(s_h, a_h)$ 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.

A Q-learning agent maintains a Q-value for each state-action pair (s, a) to estimate the profit of using each action $a \in \mathcal{A}$ under each state $s \in \mathcal{S}$. 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 $Q_i(t, s) := Q(t, s, a_i)$. Consider the case where the game ends once the action is performed. Let r_t be the immediate reward, The agent updates its Q-value for the state-action pair (s, a_i) as follows:

$$Q_i(t + 1, s) = (1 - \alpha)Q_i(t, s) + \alpha r_t \tag{2}$$

where $\alpha \in (0, 1)$ is the learning rate. An exploration mechanism aims to strike a balance between exploitation and exploration such that the performance of the agent is maximised during learning while ensuring the converging condition is met. Boltzmann exploration is a commonly used mechanism, where the action selection probability $\boldsymbol{\pi} := \boldsymbol{\pi}(t, s) = (\pi_1, ..., \pi_d) \in \Delta$ is given by

$$\pi_i = \frac{e^{\tau Q_i}}{\sum_{j=1}^{d} e^{\tau Q_j}} \tag{3}$$

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

3 Co-Evolution of Strategy and Structure

The Co-Evolutionary Model. Here we present our model for the co-evolution of strategy and structure in the spectrum of social dilemmas. Consider a network of agents where each agent is randomly connected with z neighbours and assigned an action for the underlying game, i.e., Cooperate or Defect, uniformly at random. During the simulation, agents are able to learn to update their social ties by maintaining the link with their neighbours or cutting the link and rewiring to a random neighbour in the population. They are furthermore able to update their strategy by imitating their neighbours according to their relative fitness. Algorithm 1 describes our model, which

Algorithm 1 The Co-evolutionary Model

Input: $N, z, W, S, T, H, \alpha, \tau, \beta$

1: Initialize $Agents$ with N, α, τ
2: Initialize $Strategies$ with N
3: Initialize $Network$ as a random z-regular graph
4: **for** $iteration = 1$ to H **do**
5: draw $x \in [0, 1)$ randomly
6: **if** $x \geq 1/(1 + W)$ **then**
7: draw i from $Network$, and j from $\mathcal{N}(i)$
8: $s^i \leftarrow Strategies[j]$, $s^j \leftarrow Strategies[i]$
9: $a^i \leftarrow Agents[i].getAction(s^i)$
10: $a^j \leftarrow Agents[j].getAction(s^j)$
11: **if** a^i or $a^j ==$ "Y" **then**
12: $Network.removeEdge(i, j)$
13: draw $n1$ from $\{i, j\}$, $n2$ from $Network$
14: $Network.addEdge(n1, n2)$
15: **if** $n1 == i$ **then**
16: $r^i \leftarrow M(Strategies[i], Strategies[n2])$
17: $r^j \leftarrow 0$
18: **else**
19: $r^i \leftarrow 0$
20: $r^j \leftarrow M(Strategies[j], Strategies[n2])$
21: **end if**
22: **else**
23: $r^i \leftarrow M(Strategies[i], Strategies[j])$
24: $r^j \leftarrow M(Strategies[j], Strategies[i])$
25: **end if**
26: $Agents[i].train(s^i, a^i, r^i)$
27: $Agents[j].train(s^j, a^j, r^j)$
28: **else**
29: draw i from $Network$, and j from $\mathcal{N}(i)$
30: evaluate $f(i), f(j)$ with (1)
31: draw $y \in [0, 1)$ randomly
32: **if** $y < 1/[1 + e^{-\beta(f(j)-f(i))}]$ **then**
33: $Strategies[i] \leftarrow Strategies[j]$
34: **end if**
35: **end if**
36: **end for**

is equivalent to [Santos *et al.*, 2006a], our baseline, when fixing the partner selection rule to follow a Fermi distribution.

Let us define the timescale ratio W between the strategy update and link update, to be such that for each iteration, the strategy update is performed with probability $p = 1/(1+W)$, and the link update with probability $1 - p$. $W \to 0$ corresponds to the situation where strategy evolution happens on a fixed network. As W increases, agents are given the chance to react more promptly to their neighbours' behaviour expecting, intuitively, a more prominent selection of cooperative partners at the expense of defectors.

In the case of a link update (line $7 - 27$), a link is selected between agent i and j, and the agents involved need to decide whether to keep or sever their connection. Both agents are informed of their opponent's game strategy (s^i, s^j), where $s \in \{C, D\}$ [1] and asked to come up with an action profile

[1] We considered agents with profile-dependent policies, rather

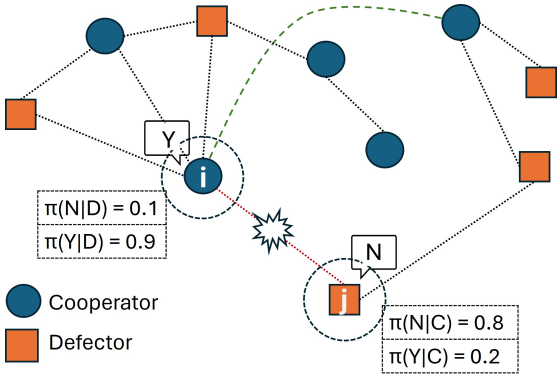

Figure 2: Illustration on a link update. A pair of linked agents are drawn (a Cooperator and a Defector in this example) and they can decide to keep or break the link with their partner according to their policy determined by their Q-values. Agent i chooses to break the tie and thus the link is removed, and one of them (agent i in this example) will form a new link with another agent chosen at random.

(a^i, a^j), where $a \in \{N, Y\}$, Y stands for yes, thus breaking ties, and N for no, thus staying. If both agents decide to stay, their link will be maintained. If either decides to cut the tie, their link will be removed and one of them will form a new link with another agent chosen at random, in such a way that the average neighbourhood size (the degree of the graph) is retained (see figure 2 for illustration). After the link is updated, both agents i and j will receive a payoff from their new connection based on their strategies, with their Q-values updated accordingly. If the link is severed, the agent that does not get re-wired will receive a payoff of 0 at that time point. To simplify the analysis, the minimum degree for an agent is set to be 1, so no agent can be fully disconnected at any point.

In the case of strategy update (line $29 - 34$), an agent is selected to update their strategy by imitation according to the pairwise comparison rule [Traulsen *et al.*, 2006] based on a Fermi function. Specifically, an agent i and one of its neighbours j are selected and their respective fitness $f(i)$ and $f(j)$ are evaluated. With probability $p = 1/[1 + e^{-\beta(f(j)-f(i))}]$, agent i will copy the strategy of agent j. In line with our baseline model, we conducted experiments over a population size of $N = 1000$ and a total number of iterations of $H = 1,000,000$. Unless otherwise specified, the average neighbourhood size is $z = 30$, the learning rate $\alpha = 0.05$, the inverse temperature for Q-learning $\tau = 5$, and the inverse temperature for imitation $\beta = 0.005$.

Emergent Partner Selection Rules Sustain Cooperation, when Fast Enough. We present our results, showing that (full) cooperation can emerge when agents adjust their social ties even when they still have to learn how to select their partners. The contour plots in Figure 3 demonstrate the percentage of cooperation in the population for different values of W on different dilemma games. Specifically, the upper right region of the plot corresponds to the Snow-Drift (SD) game, the lower left region to the Stag-Hunt (SH) game, and the lower right region to the Prisoner's Dilemma (PD) game.

than opponent-dependent only, but results are largely unaffected.

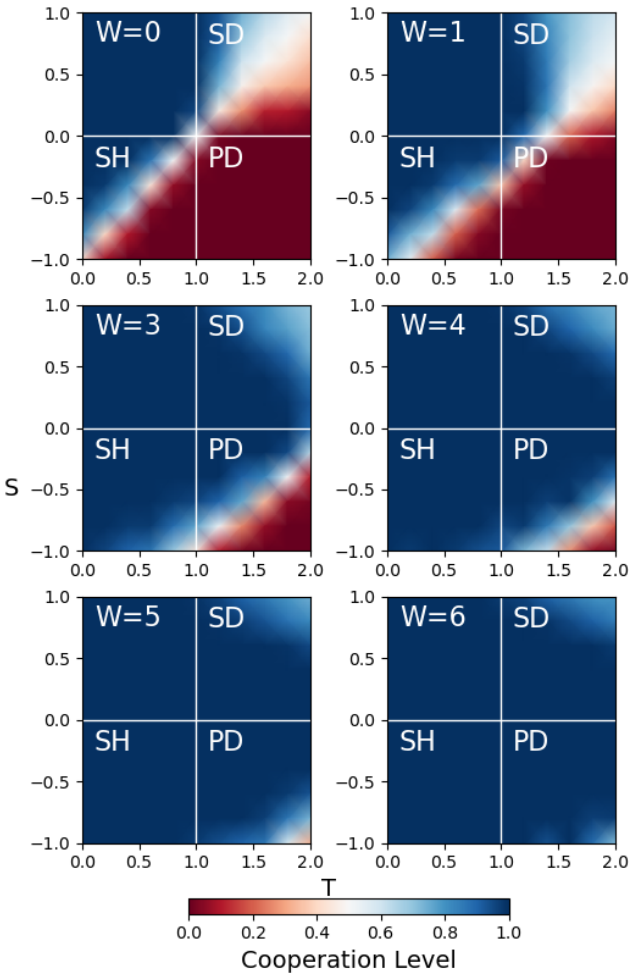

Figure 3: Cooperation levels for different timescales (W) on the full spectrum of social dilemma games, with the x-axis representing the value of T and the y-axis the value of S, encoding the Prisoner's Dilemma (PD), Snow-Drift (SD), and Stag-Hunt (SH) games, with the remaining quadrant being a fully cooperative game, $N = 1000, z = 30, \alpha = 0.05, \tau = 5, \beta = 0.005$. As W increases, so does cooperation. Note the SD corner case when cooperation is around 80%, which we explore in Section 4.2.

The upper left region completes the plot and corresponds to the degenerate case where cooperation is strictly dominant, i.e. the harmony game (HG). The levels of cooperation increase with the darkness of the blue colour, while dark red denotes a high level of defection. We can see from Figure 3 that, with suitable values of W, cooperation is sustained without the need to impose any partner selection rule. For $W = 0$ (fixed network), the result confirms the previous findings on well-mixed populations [Santos *et al.*, 2006b]. As W increases, the wave of cooperation pushes towards the lower right corner ($T = 2, S = -1$) of the plot, which presents the hardest challenge for cooperation to emerge, as defection is always strictly preferred. Leaving the upper right corner momentarily aside, we can see that with a sufficiently large value of W, full cooperation among the population is attain-

able. Intuitively, the conditions of the interaction provide sufficient opportunity for cooperators to learn to cut ties with defectors and cluster themselves in cooperator communities. While these results are equivalent to those obtained in [Santos *et al.*, 2006a], it needs to be understood that here, no specific partner selection heuristics were enforced. What makes these results thus innovative is that evolution appears to select rewiring rules that lead to more cooperation in all social dilemmas. Note nonetheless that in the upper right corner, the level of cooperation seems to stop increasing with W. In fact, for the extreme case of the SD game ($T = 2, S = 1$), cooperation rates settle around the 80% mark. As we shall see in the next section, the learnt partner selection rules are affected by the game structure and in particular their Nash equilibria. Although important similarities are shown between the different social dilemma games, their payoff structure determines the distribution and the speed at which rules are learnt.

4 Learnt Partner Selection Rules

We now analyse the emergence of partner selection rules in the spectrum of social dilemmas. After classifying the policy types based on Q-value comparison, we zoom in on the PD game, where cooperation is the harder to attain, and then discuss the SD/SH dimension.

Policy Types. By comparing the magnitude of Q-values for partner selection at different states, we can classify the agent's policy into different types. For example, if the Q-value of action N is larger than that of action Y regardless of the opponent's strategy in the social dilemma, i.e., $(Q(N|C) > Q(Y|C), Q(N|D) > Q(Y|D))$, 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 cooperates, but reverse otherwise $(Q(N|C) > Q(Y|C), Q(N|D) < Q(Y|D))$, we classify the agent as adopting the Out-for-Tat strategy; and so on. We are therefore able to classify agents' partner selection policies into four different types. They are (1) Always-Stay (Stay), where the "loyal" agent always maintains the tie regardless of the opponent's strategy, (2) Out-for-Tat (OFT), where the agent maintains the tie if the opponent cooperates and cuts the tie if the opponent defects, (3) reverse Out-for-Tat (R-OFT), which reverses the behaviour of OFT, and (4) Always-Switch (Switch), where the agent cuts ties regardless.

4.1 Zooming in the Prisoner's Dilemma

Higher Mean Degree Slows Down Cooperation. In Figure 4, we show the cooperation levels as well as the learnt partner selection policy of agents for the PD game at $T = 2, S = -1$, under the different values of mean neighbourhood size $z \in \{10, 20, 30, 40\}$. Looking at the solid lines in the upper plot, we can see the cooperation rate increases as a function of W. Full cooperation becomes more difficult to obtain when z increases. This is because, intuitively, cooperators require much more time to cut ties with the defectors.

Learning Slows Down Cooperation. We further conducted our experiment when the OFT policy is enforced, to classify the "cost" of learning this policy rather than having it imposed on the agents straight away. The dashed lines in the upper plot present the cooperation rate across W for different z. The gaps between dashed and solid lines illustrate the cost of having agents learn to adjust ties by themselves. As z increases, so does this cost. Notice that, with OFT enforced, cooperation drops as W becomes too large. This is because everyone has had enough time to cut ties with defectors, resulting in a disconnected graph with islands of defecting pairs. These agents are not able to cut ties with their only partner and imitate others' strategies, and are being forced into an unhealthy partnership with no alternatives, as a consequence of which we observe a drop in cooperation levels.

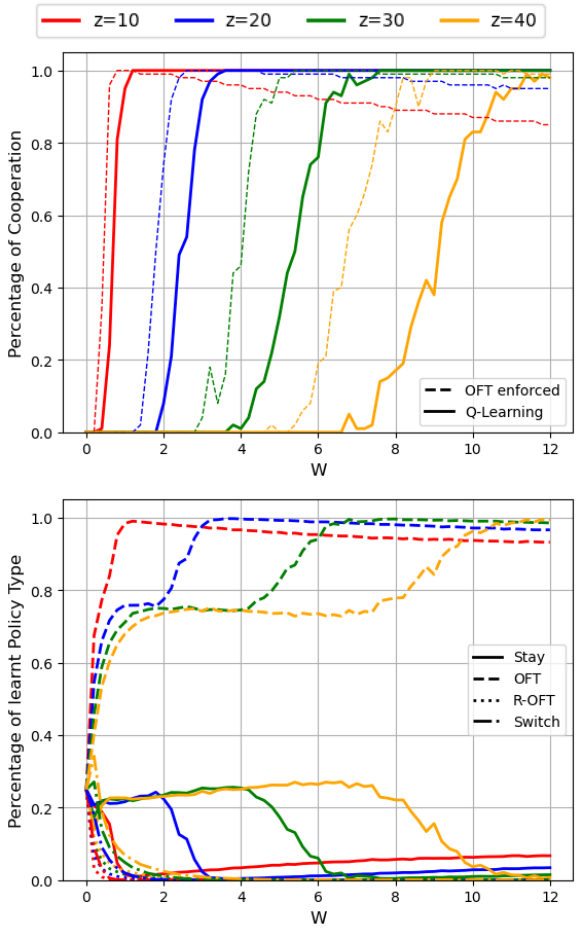

Figure 4: Interplay between cooperation rates and learnt policy types at different timescales (W) and average neighbourhood size z on the PD ($T = 2, S = -1$), $N = 1000, \alpha = 0.05, \tau = 5, \beta = 0.005$. (Top) Cooperation rates as a function of W and z show that higher z make cooperation slower to emerge. Solid lines represent the outcome of Q-learning partner selection, while the dashed shows what happens when OFT is imposed. As z increases cooperation emerges more slowly. (Bottom) Rate of learnt policy types for different W and z. As W increases, OFT agents dominate the other policy types.

Loyalty is not Enough to Outcast Defectors. The lower plot in Figure 4 presents the percentage of agents adopting different policy types across different timescales W and the average neighbourhood size z in the PD. We can see that the Always-Stay and OFT policies are the only ones that are

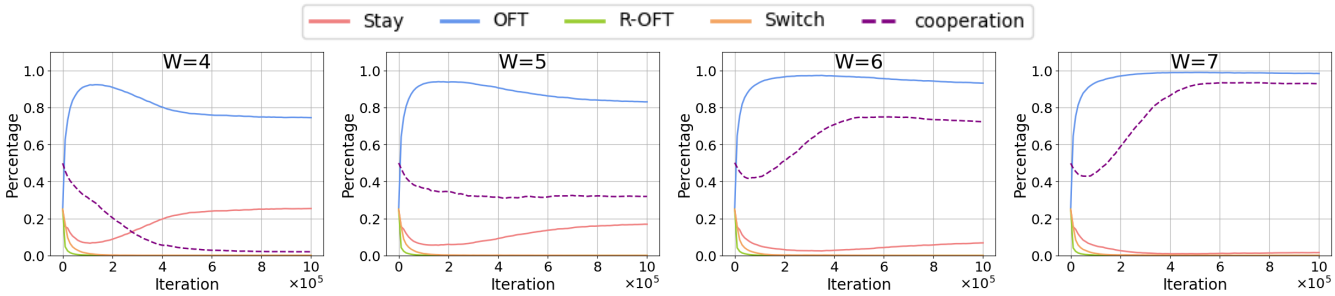

Figure 5: Learnt policy types over iteration for different timescales (W) for the PD game ($T = 2$, $S = -1$), $N = 1000, z = 30, \alpha = 0.05, \tau = 5, \beta = 0.005$. Over 90% of agents learn OFT before $200,000$ iterations. Cooperation and partner selection co-evolve based on W.

learnt by the agents. When comparing with the upper plot, we can also see that the proportion of policy types is directly related to the final cooperation rate in the PD. If defection emerges, the policy type of the population is a mixture of Stay and OFT agents. If cooperation emerges, almost all agents in the population are adopting OFT. The dominance of the OFT policy among agents makes intuitive sense. In the PD, an agent always receives a higher payoff when maintaining the link with a cooperator regardless of its strategy, and it always receives a higher payoff by cutting the link with a defector. The adoption of OFT helps the agents maintain links with cooperators and cut ties with defectors, contributing to the rise in the cooperation rate throughout the simulation. However, consider the case where the percentage of defectors is dominant in the population. Then the advantage of cutting the link with a defector is not immediately obvious, since the agent will likely be rewiring to another defector in the population. This is also the reason why a certain portion of the agents have adopted the "loyal" Stay policy if defection emerges, as the choice of random rewiring in response to a single defection may make the agent worse off. As the saying goes, "better the devil you know than the devil you don't know".

OFT Co-Evolves with Cooperation. Figure 5 shows the co-evolution of cooperation rate and learnt partner selection policy across iterations for different W. In all cases, the adoption rate of OFT has risen to over 90% at the beginning of the simulations, confirming our earlier observation on the role of the OFT policy. Depending on the value of W and the overall cooperation levels, the adoption of OFT may or may not be able to reverse the decreasing trend in cooperation rate, which will in turn affect the adoption rate of the OFT policy at the later stages of the simulation. If the cooperation rate increases, the percentage of OFT agents continues to rise; otherwise, some agents will move to adopt the Stay policy. All in all, it is evident that the partner selection policy alone does not "cause" the increased cooperation level, nor does the increased cooperation level alone "cause" the emergence of the partner selection policy, but the two co-evolve towards a fully cooperative society, if the right conditions are met.

On the Role of Inverse Temperatures. As a final point, we look at the role of the softmax function τ used in Q-learning and the Fermi function β used in the imitation with pairwise comparison rule, which are both inverse temperatures of the kind found in statistical physics. The intensity of these temperatures affects the greediness of action selection. When $\tau, \beta \to 0$, decisions are purely random. When $\tau \to \infty$, the decision to break the tie follows the agent's highest Q-value. When $\beta \to \infty$, the decision to copy the opponent's strategy is determined by whether the opponent's fitness is higher.

We now look at the effect of changing the intensity of inverse temperatures on the evolution of cooperation. Imposing $T = 1 - S$, we move along the diagonal of the PD region (see Figure 3) from $(T, S) = (1, 0)$ to $(T, S) = (2, -1)$, which corresponds to a PD with $\infty > b/c \geq 2$ [Ohtsuki *et al.*, 2006], while keeping the time scale value $W = 4$. In Figure 6, we plot the cooperation rate across $1 \leq T \leq 2$ for different intensities of the inverse temperatures. The results are shown for three different values of β for each plot, and within each panel, three different values of τ are examined. We can see a clear transition from full cooperation to full defection when T increases, as the "difficulty" of the game increases. For the effects of either inverse temperature, we can observe how τ has a positive effect on cooperation, while β has a negative one. Therefore, to improve the level of cooperation, it is preferable to adopt the learnt partner selection rule quickly, while copying the other's strategy slowly, which is in line with the timescales analysis of our baseline model [Santos *et al.*, 2006a]. In the previous section, we have shown how the adoption of the OFT policy is an effective rule for promoting cooperation and that it can quickly emerge in the right conditions. It is intuitive to see why increasing τ promotes cooperation, independently of the adopted game strategies. On the other hand, we know that defectors are more successful (in terms of fitness) at the earlier stage of the game, while cooperators could catch up after cutting the links with defectors. Thus, to promote cooperation, agents need to avoid copying early success. It is finally worth noting that the magnitudes of τ and β are not directly comparable. This is because Q-learning estimates the payoffs between the actions of a single agent, without requiring interpersonal comparison, whereas imitation compares the fitness (total payoff) between agents. As a rule of thumb, for an average neighbourhood size of $z = 30$, we expect a 30 times difference in parameter value.

4.2 The Snow-Drift/Stag-Hunt Dimension

In the previous section, we have shown that the level of cooperation in certain SD games is capped at around 80% for the extreme payoff distributions, even though we consider the SD

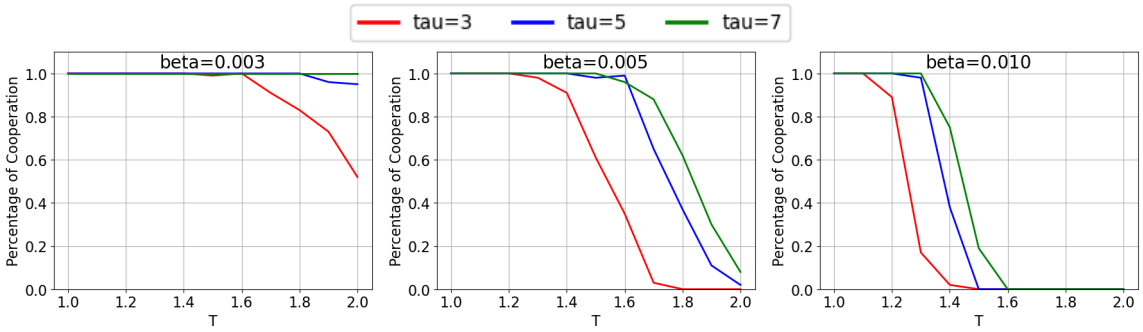

Figure 6: The effect of inverse temperatures τ (positive) and β (negative) on the cooperation levels in the PD game. The x-axis represents the diagonal region of the PD in Figure 3 from $(T, S) = (1, 0)$ to $(T, S) = (2, -1)$, $N = 1000, z = 30, W = 4, \alpha = 0.05$.

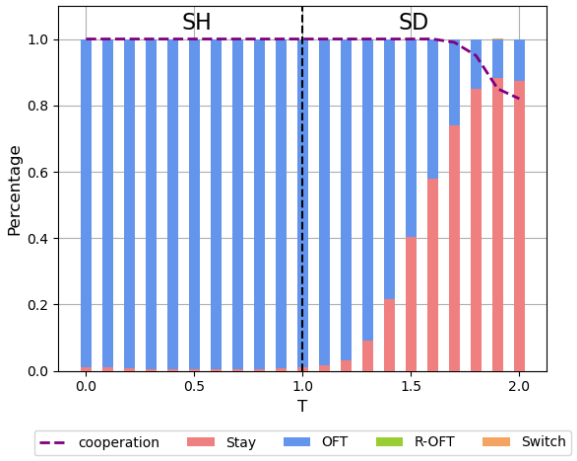

Figure 7: The effect of learnt policy types and cooperation levels on the SH and SD games, corresponding to the induced diagonal in Figure 3 from $(T, S) = (0, -1)$ to $(T, S) = (2, 1)$, $N = 1000, z = 30, W = 8, \alpha = 0.05, \tau = 5, \beta = 0.005$. In SH, full cooperation co-evolves with the OFT policy. In SD, Stay becomes prevalent with more extreme payoffs and cooperation settles around 80%.

game to be a less difficult scenario compared to the PD for the emergence of cooperation. This is due to the SD payoff structure, requiring matching opposing behaviours, which affects the emergent policy types, which in turn affects the overall cooperation rate in the population. We thus now look at the effect that the payoffs in the SH and SD games have on the learnt policy types and cooperation rate. Imposing $T = S-1$, we move along the diagonal in the SH and SD game region (see Figure 3) from $(S, T) = (-1, 0)$ to $(S, T) = (1, 2)$, while keeping the time scale value $W = 8$. Figure 7 plots the cooperation rate and distribution of agents' policy type across $0 \leq T \leq 2$. We can see in the range of SH games that the population achieves full cooperation and most agents adopt the OFT policy. On the other hand, in the range of SD games, Stay agents start to take over and the cooperation rate experiences a (mild) drop. The SH game is a common interest game with symmetric pure Nash equilibria. When agents are given the option to switch their ties and update their strategies, cooperation is not hard to emerge as it maximises the payoff of the individuals. On the other hand, in the SD game, the relative advantage of a cooperator leaving a defector is less obvious as the difference between the payoff of being cheated on and mutual operation is not significant. In the extreme case of $S = 1, T = 2$, the difference is effectively zero, therefore cooperators are indifferent between maintaining or cutting ties with defectors. Regarding the strategy, as the SD game contains asymmetric pure Nash equilibria, cooperators have less motivation to copy the strategy from the defectors, therefore the level of cooperation is kept at a higher rate.

5 Discussion

We studied the emergence of cooperation in social dilemmas played on networks, with individuals learning partner selection rules by themselves. We showed that the learnt strategies support the levels of cooperation observed in the literature using hard-wired heuristics, confirming that cooperation flourishes when rewiring is fast enough relative to imitation [Santos *et al.*, 2006a]. We demonstrated the role of OFT, a simple rule that keeps ties with cooperators and breaks them with defectors, across the spectrum of social dilemmas. The Stay strategy is also helpful in supporting cooperation, but its loyal nature is not sufficient to outcast defectors.

Our results open several avenues for future research. While our focus was on learning partner selection rules, game decisions were still based on imitation-learning. A natural next step would be the study of learnt in-game strategies as well, in the spirit of [Leung and Turrini, 2024], although we believe cooperation will be harder to maintain when defectors can exploit a large number of neighbours at the same time. Whether an interpersonal comparison rule, as in [Santos *et al.*, 2006a], can emerge to sustain cooperation is far from clear and will likely require more complex strategy spaces. Furthermore, alternative exploration algorithms warrant investigation, e.g., epsilon-greedy or learning automata [Segbroeck *et al.*, 2010].

Despite the challenges of a replicator dynamic analysis, we can still explore extreme or simplified cases, in the spirit of [Zheng *et al.*, 2017], or, for example, generalising the two-dimensional timescale dynamics for extreme cooperator popularity in [Bara *et al.*, 2022], to account for non-deterministic partner selection rules. A final important follow-up research direction concerns the relation with human behaviour, looking at which partner selection rules are selected in practice.

Acknowledgments

CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG-2023-050 and the TAILOR Connectivity Fund (Agreement 29). TL acknowledges the support of a F.R.S.-FNRS PDR (40007793), the EU Horizon 2020 project TAILOR (952215) and a Service Public de Wallonie Recherche project (2010235–ariac) by digitalwallonia4.ai.

References

[Anastassacos *et al.*, 2020] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner selection for the emergence of cooperation in multi-agent systems using reinforcement learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, 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*, pages 7047–7054. AAAI Press, 2020.

[Bara *et al.*, 2022] Jacques Bara, Paolo Turrini, and Giulia Andrighetto. Enabling imitation-based cooperation in dynamic social networks. *Auton. Agents Multi Agent Syst.*, 36(2):34, 2022.

[Bloembergen *et al.*, 2015] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *J. Artif. Intell. Res.*, 53:659–697, 2015.

[Börgers and Sarin, 1997] Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

[Eshel and Cavalli-Sforza, 1982] Ilan Eshel and L. L. Cavalli-Sforza. Assortment of encounters and evolution of cooperativeness. *Proceedings of the National Academy of Sciences*, 79(4):1331–1335, 1982.

[Foley *et al.*, 2018] Michael Foley, Patrick Forber, Rory Smead, and Christoph Riedl. Conflict and convention in dynamic networks. *Journal of The Royal Society Interface*, 15(140):20170835, 2018.

[Fulker *et al.*, 2021] Zachary Fulker, Patrick Forber, Rory Smead, and Christoph Riedl. Spite is contagious in dynamic networks. *Nature Communications*, 12(1):260, 2021.

[Gilbert, 1995] Nigel Gilbert. Emergence in social simulation. In Nigel Gilbert and Rosaria Conte, editors, *Artificial Societies: The Computer Simulation Of Social Life*. Routledge, 1995.

[Leung and Turrini, 2024] Chin-wing Leung and Paolo Turrini. Learning partner selection rules that sustain cooperation in social dilemmas with the option of opting out. In Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum, editors, *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, pages 1110–1118. ACM, 2024.

[Nowak, 2006] Martin A. Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

[Ohtsuki *et al.*, 2006] Hisashi Ohtsuki, Christoph Hauert, Erez Lieberman, and Martin A Nowak. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092):502–505, 2006.

[Pacheco *et al.*, 2006] Jorge M. Pacheco, Arne Traulsen, and Martin A. Nowak. Coevolution of strategy and structure in complex networks with dynamical linking. *Phys. Rev. Lett.*, 97:258103, Dec 2006.

[Pérolat *et al.*, 2017] Julien Pérolat, Joel Z. Leibo, Vinícius Flores Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3643–3652, 2017.

[Perreau de Pinninck *et al.*, 2010] Adrian Perreau de Pinninck, Carles Sierra, and Marco Schorlemmer. A multi-agent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems*, 21(3):397–424, Nov 2010.

[Pujol *et al.*, 2002] Josep M. Pujol, Ramon Sangüesa, and Jordi Delgado. 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*, AAMAS '02, page 467–474, New York, NY, USA, 2002. Association for Computing Machinery.

[Rand *et al.*, 2011] David G. Rand, Samuel Arbesman, and Nicholas A. Christakis. Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences*, 108(48):19193–19198, 2011.

[Roca *et al.*, 2009] Carlos P. Roca, José A. Cuesta, and Angel Sánchez. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of Life Reviews*, 6(4):208–249, 2009.

[Sabater and Sierra, 2002] Jordi Sabater and Carles Sierra. 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*, New York, New York, USA, 2002. ACM Press.

[Sabater-Mir *et al.*, 2006] Jordi Sabater-Mir, Mario Paolucci, and Rosaria Conte. Repage: Reputation and image among limited autonomous partners. *Journal of Artificial Societies and Social Simulation*, 9(2):3, 2006.

[Salazar *et al.*, 2011] Norman Salazar, Juan A. Rodriguez-Aguilar, Josep Ll. Arcos, Ana Peleteiro, and Juan C. Burguillo-Rial. Emerging cooperation on complex networks. In *The 10th International Conference on Au-*

tonomous Agents and Multiagent Systems - Volume 2, AAMAS '11, page 669–676, Richland, SC, 2011. International Foundation for Autonomous Agents and Multiagent Systems.

[Santos *et al.*, 2006a] Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. Cooperation prevails when individuals adjust their social ties. *PLoS Comput. Biol.*, 2(10), 2006.

[Santos *et al.*, 2006b] Francisco C. Santos, Jorge M. Pacheco, and Tom Lenaerts. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *Proceedings of the National Academy of Sciences*, 103(9):3490–3494, 2006.

[Santos *et al.*, 2018] Fernando P. Santos, Jorge M. Pacheco, and Francisco C. Santos. Social norms of cooperation with costly reputation building. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *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*, pages 4727–4734. AAAI Press, 2018.

[Segbroeck *et al.*, 2009] Sven Van Segbroeck, Francisco C. Santos, Ann Nowé, Jorge M. Pacheco, and Tom Lenaerts. The coevolution of loyalty and cooperation. In *Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009*, pages 500–505. IEEE, 2009.

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

[Traulsen *et al.*, 2006] Arne Traulsen, Martin A Nowak, and Jorge M Pacheco. Stochastic dynamics of invasion and fixation. *Physical Review E*, 74(1):011909, 2006.

[Wang *et al.*, 2012] Jing Wang, Siddharth Suri, and Duncan J. Watts. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36):14363–14368, 2012.

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

[Zhang *et al.*, 2016] Bo-Yu Zhang, Song-Jia Fan, Cong Li, Xiu-Deng Zheng, Jian-Zhang Bao, Ross Cressman, and Yi Tao. Opting out against defection leads to stable coexistence with cooperation. *Scientific reports*, 6:35902, October 2016.

[Zheng *et al.*, 2017] Xiu-Deng Zheng, Cong Li, Jie-Ru Yu, Shi-Chang Wang, Song-Jia Fan, Bo-Yu Zhang, and Yi Tao. 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*, 420:12–17, 2017.