Learning to Cooperate with Minimal Observability

Chin-wing Leung[1], **Paolo Turrini**[1], **Fernando P. Santos**[2], **Mirco Musolesi**[3,4]
[1]Department of Computer Science, University of Warwick, Coventry, United Kingdom
[2]Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
[3]Department of Computer Science, University College London, London, United Kingdom
[4]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
chin-wing.leung@warwick.ac.uk, p.turrini@warwick.ac.uk, f.p.santos@uva.nl, m.musolesi@ucl.ac.uk

Abstract

Cooperation among independent learning agents is desirable as it enables reaching collectively rewarding states. Recent work has shown that artificial agents can learn to act prosocially without the need for predefined cooperative preferences or behavioural heuristics, provided that they can observe others' actions or policies and select them as partners accordingly. This paper relaxes this constraint, studying reinforcement learning (RL) agents operating with only minimal information about others' behaviour. We propose a novel 'Observer Model', where agents gain insights from direct experience and limited, indirect observations. We show that direct experience alone cannot sustain cooperation, particularly in large societies. However, even minimal observations of third-party interactions, allowing as few as one observer per gameplay, lead to significant improvements, enabling the population to achieve and sustain robust cooperation across varying population sizes. Through numerical analysis, we show the co-evolution of strategy and interaction structure and disentangle how learning happens under various settings. Analysing the partner selection graph, we identify the reasons for cooperation to emerge, and we explore how different learning and exploration rates affect the outcome of social dilemmas played among RL agents.

Introduction

Over the past decade, there has been a significant and growing effort to design AI agents capable of cooperative behaviour, particularly in the context of social dilemmas (SD) (Fatima, Jennings, and Wooldridge 2024). This has motivated a variety of learning heuristics to avoid the uncooperative suboptimal outcomes that typically arise in such settings (Peysakhovich and Lerer 2018; Bazzan, Peleteiro, and Burguillo 2011; Fan et al. 2022).

Reinforcement Learning (RL) agents augmented with partner selection mechanisms have been shown to successfully learn cooperative behaviour in social dilemmas, operating in a fully decentralised fashion (Anastassacos, Hailes, and Musolesi 2020). This holds true even when partner selection capabilities are minimal (Leung and Turrini 2024) and under challenging, network-constrained interaction settings (Leung, Lenaerts, and Turrini 2024). These findings rest on the assumption that agents can observe the past behaviour of others. The ability to identify and preferentially interact with past cooperators serves as a powerful heuristic for forming stable, cooperative groups (Santos, Pacheco, and Lenaerts 2006). Moreover, after partners are selected, tracking others' actions enables direct reciprocity: strategies such as Tit-for-Tat (Axelrod 1984) can further discourage selfish behaviour and promote cooperation.

Although information on past interactions can be used to select partners, accessing prior behaviours of agents, without having interacted with them directly, is challenging in real-world settings: interactions can be private (Hilbe et al. 2018; Perret, Krellner, and Han 2021), individuals might refuse to share their experiences (Santos, Pacheco, and Santos 2018), and direct observability can be limited to narrow settings. This raises a critical question: can cooperation still emerge when agents lack such observational capabilities? Exploring what happens when agents are limited to their own direct experiences, or otherwise constrained in their ability to observe others, is a crucial step toward developing a robust and generalizable theory of cooperation in learning agents.

Contribution. In this paper, we study independent RL agents playing a social dilemma and capable of selecting other agents to interact with. In contrast with the literature, these agents, which we model as epsilon-greedy Q-learners, can only rely on limited observations and experiences about the others' behaviour. We show that direct experience alone is insufficient for the emergence of cooperation, particularly in large societies. We propose a novel *Observer Model*, allowing as few as one observer per gameplay, and show that this enables the population to achieve and sustain cooperation across varying population sizes. Through the analysis of policies' evolution under different levels of observability, we trace the co-evolution of strategy and interaction structure and disentangle how learning happens under various settings. By inspecting the interaction graph, which agents form by repeatedly selecting other agents to interact with, we identify key reasons for cooperation to emerge, and we explore how different learning and exploration rates affect the outcome of social dilemmas played among RL agents.

Related Work. Understanding why self-interested individuals cooperate is a key question for many fields of science, such as economics, evolutionary and social psychol-

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

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

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

ogy or biology (Nowak 2006; Kollock 1998; Sigmund 2010; Van Lange et al. 2013; Bowles and Gintis 2013). Various studies, empirical (Barclay and Willer 2007; Rand, Arbesman, and Christakis 2011; Wang, Suri, and Watts 2012; Zhang et al. 2016) and theoretical (Segbroeck et al. 2009; Sylwester and Roberts 2010; Zheng et al. 2017; Santos, Pacheco, and Lenaerts 2006; Bara, Turrini, and Andrighetto 2022; Graser et al. 2025), have shown how the capacity of individuals to choose reliable partners is key for this to happen. In computational social science, agent-based simulation models are widely used to study the emergence of cooperation in social dilemmas (e.g., (Gilbert 1995; Salazar et al. 2011; Santos and Pacheco 2005)), typically featuring agents governed by heuristic decision-making rules.

RL has become a key paradigm for studying the emergence of cooperation among autonomous learning agents (Sandholm and Crites 1996; Dafoe et al. 2020; Dasgupta and Musolesi 2025; Leibo et al. 2017; Smit and Santos 2024; Barfuss et al. 2025). Breakthroughs have been achieved in scenarios involving the exploitation of common resources, particularly through the development of trust mechanisms among agents (Pérolat et al. 2017; Leibo et al. 2017; Tennant, Hailes, and Musolesi 2024). Partner selection has emerged as a key mechanism driving cooperation among independent Q-learners in SDs (Anastassacos, Hailes, and Musolesi 2020; Leung and Turrini 2024; Leung, Lenaerts, and Turrini 2024). However, these studies typically rely on global information, assuming that agents can access the most recent behaviour of all potential partners prior to making their selection, an assumption that we remove in this paper.

The importance of studying mechanisms that achieve cooperation in the presence of imperfect and noisy information is widely acknowledged. Reputation and indirect reciprocity have long been recognised as key mechanisms (Sabater and Sierra 2002; Pujol, Sangüesa, and Delgado 2002; Perreau de Pinninck, Sierra, and Schorlemmer 2010; Perret, Krellner, and Han 2021; Santos, Pacheco, and Santos 2021; Ren et al. 2025). (Anastassacos et al. 2021) show that when RL agents adopt a shared social norm to assign reputations, cooperative behaviour emerges within the population. While reputation typically relies on a shared understanding of what constitutes good/bad actions (Santos, Pacheco, and Santos 2018), in our study, agents base their decisions solely on the most recent observed actions of others.

Preliminaries

Social Dilemmas. Social dilemmas (SD) are a benchmark model for the emergence of cooperation between autonomous agents, with the Prisoner's Dilemma (PD) serving as the main example and the hardest to solve. The PD is a 2-player symmetric game where both players can choose to Cooperate (C) or Defect (D). Players receive a payoff based on the game outcome: R (mutual cooperation (C, C)), P (mutual defection punishment (D, D)), S (being cheated on (C, D)), or T (cheating (D, C)), from Player 1's perspective. The payoff matrix is presented in Figure 1. While mutual cooperation provides the best collective outcome, defection is the dominant strategy, leading to a Nash equilibrium of mutual defection, encoding the tension between societal and individual interest that is typical of social dilemmas.

Partner Selection. Following the partner selection mechanism introduced in (Anastassacos, Hailes, and Musolesi 2020), each agent has the opportunity to select another agent to interact with at each round. Once selections are made, every resulting pair plays a Prisoner's Dilemma, after which agents update both their game and partner selection strategies based on the obtained payoffs. Note that while every agent participates in at least one game per round, some may play multiple times if repeatedly chosen as partners.

Q-learning

We train our agents using Q-learning (Watkins and Dayan 1992), a well-established reinforcement learning algorithm that operates within the framework of a Markov Decision Process (MDP). Each agent i learns its policy independently. At each time step, the agent observes the current state of the environment, denoted as $s_t \in S$, where S is the set of all possible states. The agent then chooses an available action $a_t \in A(s_t)$, where $A(s_t)$ returns the set of available actions in s_t. We denote the corresponding Q-value as $Q^i(s_t, a_t)$, estimating the expected accumulated discounted reward of choosing action a_t at state s_t. After performing the action, the agent receives an immediate reward $r_t \in \Re$ and enters the next state $s_{t+1} \in S$. The Q-value of the chosen action of the agent is updated as follows:

$$Q^i(s_t, a_t) \leftarrow Q^i(s_t, a_t) + \alpha[G_t - Q^i(s_t, a_t)], \quad (1)$$

where $G_t = r_t + \gamma max_{a'} Q^i(s_{t+1}, a')$ the estimated accumulated discounted reward, $\alpha \in [0, 1]$ is the learning rate, and $\gamma \leq 1$ is the discount factor.

An exploration mechanism aims to strike a good balance between exploitation and exploration such that the performance of the agent is maximised during learning while ensuring that desirable convergence guarantees are met. Epsilon-greedy exploration is a commonly used mechanism, where the stochastic policy $\boldsymbol{\pi}^i(s) = (\pi(s, a_1), ..., \pi(s, a_{|A|})) \in \Delta$ is evaluated as:

$$\pi(s, a_k) = \frac{\epsilon}{|A|} + \frac{\mathbb{1}\{a_k \in arg\,max_a Q^i(s, a)\}}{|\{arg\,max_a Q^i(s, a)\}|}(1 - \epsilon) \quad (2)$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, $\epsilon \in [0, 1]$ is the exploration rate.

Deep Q-learning. Function approximation is adopted when the state space is large and standard Q-learning becomes infeasible. The deep Q-network (DQN) proposed by (Mnih et al. 2015) utilises a neural network to approximate

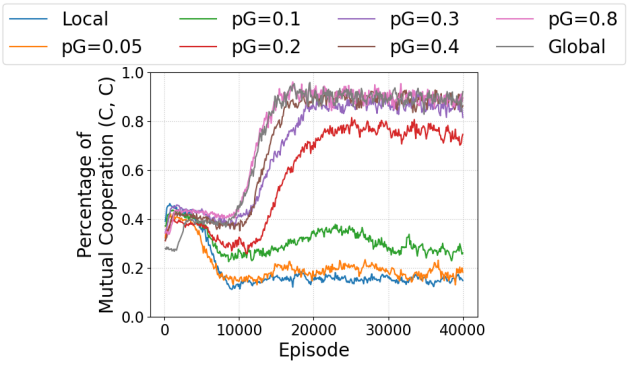


Figure 2: The percentage of mutual cooperation (C, C) across episodes with agents acting based on different probabilities p_G of utilising global information. The cooperation rate rises with p_G. Local means $p_G = 0$ and Global $p_G = 1$.

the Q-value $Q(s_t, a_t; \theta^i)$, together with the idea of experience replay and target network, has achieved remarkable results. The parameters update under DQN is as follows:

$$\theta^i \leftarrow \theta^i + \alpha[Y_t - Q(s_t, a_t; \theta^i)]\nabla_\theta Q(s_t, a_t; \theta^i), \qquad (3)$$

where $Y_t = r_t + \gamma max_{a'} Q(s_{t+1}, a'; \theta^{i-})$, and θ^{i-} are the parameters from the target network updated periodically.

Observability and Cooperation

In this section, we study how the capacity of agents to observe what others have done affects their capacity to learn cooperative strategies. We do so through our **Observer Model**, which encodes the agents' observability mechanism. Before doing that we show that observation is essential for cooperation and local models, even when accelerated by mutual learning, fail to achieve it.

Restricting Agents' Observability

Partner selection was shown to enable cooperation in populations of decentralised self-interested RL agents (Anastassacos, Hailes, and Musolesi 2020). This and follow-up works (Leung and Turrini 2024; Leung, Lenaerts, and Turrini 2024) implicitly assume that interactions happen publicly, and everyone has observed what their current partner last did. In this section, we remove this assumption completely, assuming that agents can only count on their own direct experience when evaluating others.

In this 'local' information setting, each agent is able to recall the last action an agent played against them. As soon as the agent interacts with the same agent twice, only the last action is recorded. Such records are the only information used to decide what to do when facing them or whether to select them as partners. We refer to the model in (Anastassacos, Hailes, and Musolesi 2020) as the **Global Information Model**, and this local restriction as the **Local Information Model**. To better illustrate how the availability of global information can alter the convergence towards cooperative policies, we conducted experiments in settings combining the Local and Global Information models. To this end, Figure 2 plots the percentage of mutual cooperation (C, C) across episodes with different probabilities p_G. We can see that cooperation is hindered under the Local Information Model ($p_G = 0$) and that its rate rises with p_G. When up to $p_G = 0.3$, cooperation pairs emerge and stabilise around 85%. Further increasing p_G will speed up the convergence rate. When $p_G = 0.8$, we retrieve the cooperation rates from the literature, under the Global Information Model (Anastassacos, Hailes, and Musolesi 2020). The availability of global information plays a key role in promoting cooperation, yet the full availability seems unnecessary, a key finding to motivates the Observer Model introduced.

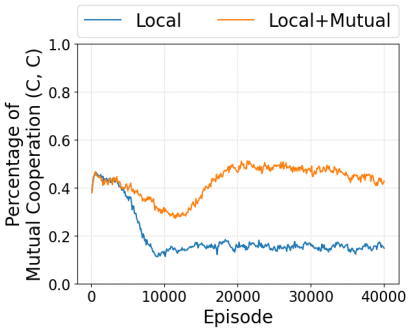


Figure 3: Percentage of mutual cooperation (C, C) across episodes with mutual partner selection policy update. Mutual policy update corresponds to a higher cooperation rate.

Mutual Update of Partner Selection Policy

When agents choose each other to play the game, they hold Q-values that reflect their opinion about the current partner. In our model we assume that once a pair is formed and interacts, *both* agents learn from that experience. This is an addition to the partner selection model studied in the literature (Anastassacos, Hailes, and Musolesi 2020), where only the *selecting* agent updates Q-value, while the *selected* agent does not. Besides being a reasonable addition, which is in line with what happens for the gameplay strategies, where both agents learn from direct experience, this also ends up accelerating learning and improving overall performance. To see why this is the case we conducted the experiments on the Local Information Model with mutual update on partner selection. Figure 3 presents the percentage of mutual cooperation (C, C) across episodes with mutual partner selection update. When mutual update is implemented, the overall learning effectiveness for partner selection improves, and the percentage of mutual cooperation has risen from below 20% to above 40%. This finding will play a crucial role in our results, but it also demonstrates that mutual update alone is insufficient to foster cooperation; it must be complemented by a degree of limited observability, as we explore next.

The Observer Model

Building on the previous findings, we propose the Observer Model, which constrains the observability of agents' interactions while retaining the ability to achieve cooperation.

Algorithm 1: The Observer Model

Input: learning rate α, exploration rate ϵ, number of rounds per episode n_R, number of agents n_A

1: **for each** episode **do**
2: **for each** round **do**
3: **for each** $agent_i$ in $population$ **do**
4: $s_{PS_i} \leftarrow agent_i.memory.o_{LastActions}$
5: $a_{PS_i} \leftarrow agent_i.partnerSelect(s_{PS_i})$
6: $agent_j \leftarrow mapping(a_{PS_i})$
7:
8: $agent_k \leftarrow selectObsever(population)$
9: $s_{PD_i} \leftarrow agent_i.memory.o_{LastActions_j}$
10: $s_{PD_j} \leftarrow agent_j.memory.o_{LastActions_i}$
11: $a_{PD_i} \leftarrow agent_i.gameSelect(s_{PD_i})$
12: $a_{PD_j} \leftarrow agent_j.gameSelect(s_{PD_j})$
13: $r_i, r_j \leftarrow playGame(a_{PD_i}, a_{PD_j})$ (Fig.1)
14:
15: $agent_i.StoreAndTrain(s_{PS_i}, a_{PS_i}, 0)$ (Eq.3)
16: $agent_i.StoreAndTrain(s_{PD_i}, a_{PD_i}, r_i)$ (Eq.1)
17: $s_{PS_j} \leftarrow agent_j.memory.o_{LastActions}$
18: $a_{PS_j} \leftarrow mapping(agent_i)$
19: $agent_j.StoreAndTrain(s_{PS_j}, a_{PS_j}, 0)$ (Eq.3)
20: $agent_j.StoreAndTrain(s_{PD_j}, a_{PD_j}, r_j)$ (Eq.1)
21:
22: $agent_i.memoryUpdate(j, a_{PD_j})$
23: $agent_j.memoryUpdate(i, a_{PD_i})$
24: $agent_k.memoryUpdate(i, a_{PD_i})$
25: $agent_k.memoryUpdate(j, a_{PD_j})$
26: **end for**
27: **end for**
28: **end for**

Consider a population of n_A agents learning to play several rounds of the Prisoner's Dilemma (PD), where each agent can select a partner to play with in every round. The agents' goal is to earn the highest payoff across n_R rounds of the game. The best outcome for the population (in terms of total rewards of all agents) is achieved when all agents cooperate every time. However, such an outcome is hard to achieve since the immediate reward for defection is higher, especially in a highly cooperative society. With the partner selection mechanism, an agent has to choose between the immediate reward of defecting against its cooperative partner and the future rewards generated by stable cooperation. The Observer Model is described in Algorithm 1.

Each round of the game comprises two stages. In the first stage (partner selection, Algorithm 1 lines $4-6$), each agent i can choose a partner whom they would like to play with. The decision is made based on the last actions in the PD of other agents stored in the memory $s_{PSi} \in \{o_C, o_D\}^{n_A-1}$. The agent cannot select itself, and the selected agent j cannot refuse to play. No reward is received at this stage. During the second stage (gameplay: line $8-13$), agent i and j will play a PD (a_{PD_i}, a_{PD_j}), where $a_{PD_i}, a_{PD_j} \in \{C, D\}$ and receive their rewards (r_i, r_j). The decisions are made based on the last actions in the PD of their partners, stored in their own memory $s_{PDi}, s_{PDj} \in \{o_C, o_D\}$. During the gameplay, a third agent is randomly selected as the 'observer' to witness the actions played by the pair and record such observation in their memory.

Each agent maintains two policies for action selections, π_{PS} for partner selection, π_{PD} for the action selection in the PD game. The policies are updated based on the Q-learning algorithms discussed in Section 2 (Algorithm 1, lines $15-20$). In line with (Anastassacos, Hailes, and Musolesi 2020), we adopt DQN for partner selection and standard Q-learning for the PD. For DQN, the neural network is parameterised with one hidden layer of size 256 with the ReLU activation function[1]. The policies are generated by epsilon-greedy exploration with $\epsilon = 0.01$, the learning rate $\alpha = 0.005$, and the discount rate $\gamma = 1$[2]. Finally, the players i, j and the observer k will update their memory of the last actions performed (Algorithm 1, lines $22-25$).

Experimental Results

In this section, we present our experimental results, showing how cooperation can emerge under the Observer Model. We then identify four phases of the population strategy's development, characterised by different outcome distributions in the PD. We analyse the types of strategies that are learnt, both during the game stage and the partner selection stage. We then look at the robustness of the model under varying population size, as well as modifying classical Q-learning hyperparameters such as learning and exploration rates.

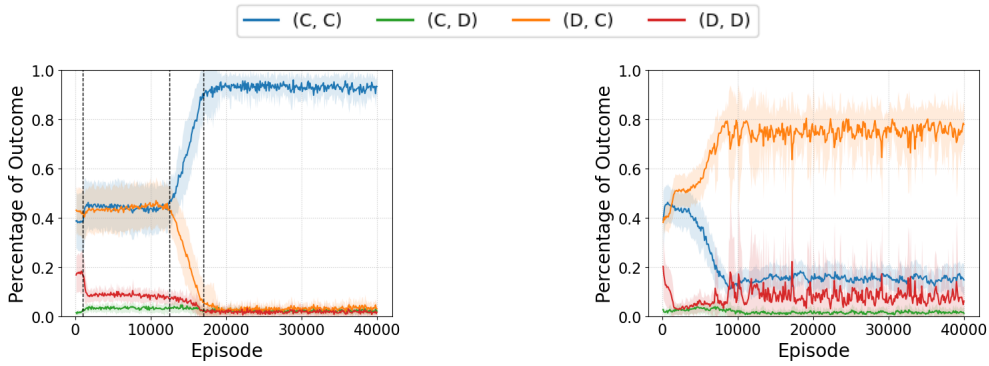
Emergence of Cooperation

Under the Observer Model, full cooperation emerges and stabilises. Figure 4 compares the percentage of outcomes in PD across episodes between the Observer Model and the Local Information Model. The results are averaged over 20 simulations for a population of 30 agents. Under the Observer Model, the outcome of mutual cooperation (C, C) is slightly less popular than unilateral exploitation (D, C) at first. After episode $12,500$, cooperative behaviour begins to dominate. On the other hand, under the Local Information Model, the outcome of exploitation (D, C) becomes dominant quickly after the $5,000$ episode, whereas there is still some level of mutual cooperation (C, C) and mutual defection (D, D) throughout the learning. Note that this is not equivalent to the result we will get under random matching, where mutual defection (D, D) prevails (Anastassacos, Hailes, and Musolesi 2020). This reflects the fact that some agents have learnt to be a cooperator when partner selection is allowed, yet the unavailability of the global information has prevented this group from further growing.

Zooming in on the result in sub-Figure 4a, we notice four different phases of learning, characterised by different outcome distributions in the PD game, under the Observer Model. Each phase corresponds to the development of a population strategy as follows:

[1]The replay memory size is 160, batch size 32, target network update interval 100, optimised over multiple configurations.

[2]We optimise the hyperparameters through a grid search, with α, ϵ ranging from 0.001 to 0.05.



(a) Observer Model

(b) Local Information Model

Figure 4: The mean and standard deviation of the percentage of outcomes in PD across episodes under the Observer Model and the Local Information Model. When agents can only make decisions based on local information, the outcome of exploitation (D, C) quickly dominates. With the Observer Model, the outcome of mutual cooperation (C, C) is slightly less popular at first, and starts to take over after $12,500$ episodes. The population size is 30, and the results are averaged over 20 simulations. Vertical dashed lines represent distinct phases of learning discussed in our Results section.

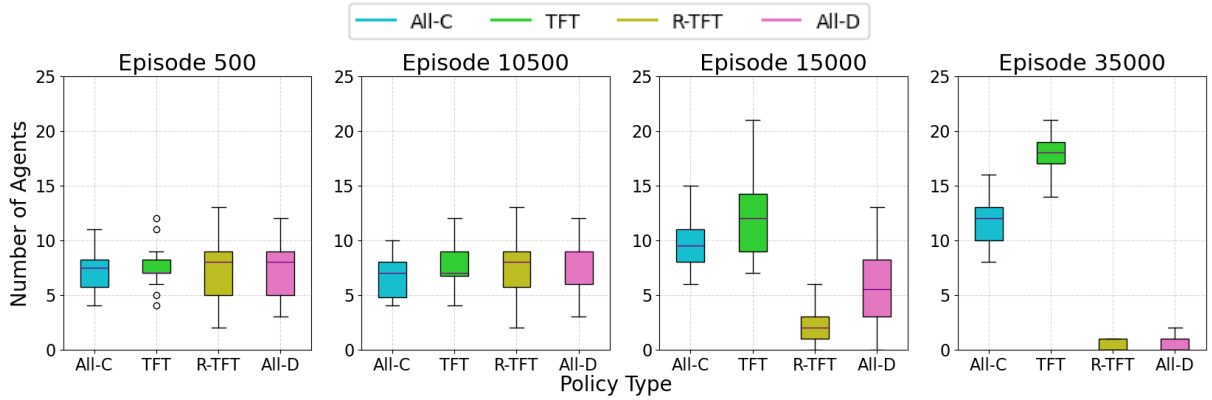


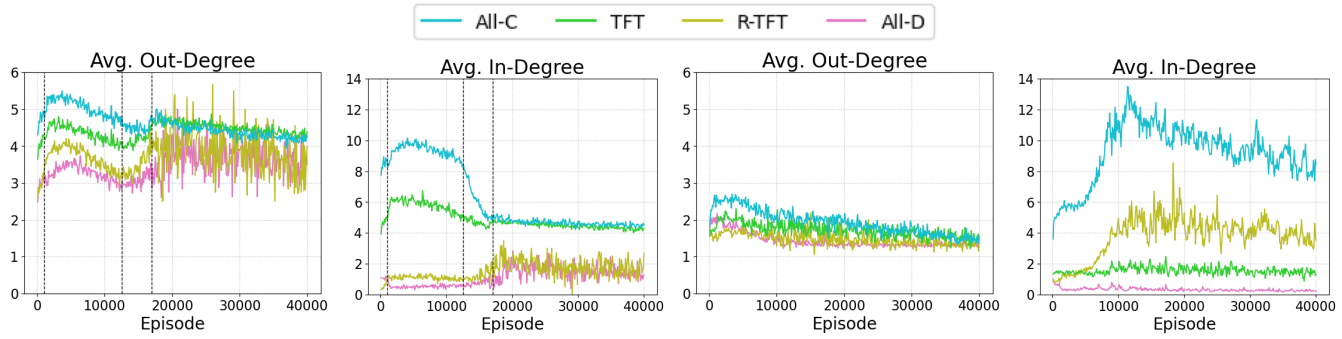
Figure 5: Box plots representing the number of agents adopting different policy types across learning phases. The policy types are quite even in the earlier phase, whereas later on *TFT* agents start to take over, consequently pushing up the *ALL-C* agents.

- Phase 1 (episodes 0 to $1,000$): Exploitation (D, C) is popular, followed by mutual cooperation (C, C) and then mutual defection (D, D).
- Phase 2 (episodes $1,000$ to $12,500$): Mutual defection (D, D) has significantly dropped, mutual cooperation (C, C) and exploitation (D, C) are fairly balanced.
- Phase 3 (episodes $12,500$ to $17,000$): Mutual cooperation (C, C) grows sharply while exploitation (D, C) and mutual defection (D, D) decrease.
- Phase 4 (episodes $17,000$ onwards): Cooperation emerges and is sustained, as over 90% of interactions result in mutual cooperation (C, C).

Emergent Policy Types

In our setting, the agents' decisions are only based on the last recorded PD action performed by their partner. Therefore, there are only two states and four Q-values in the gameplay stage. By analysing the Q-values for different states, we can categorise the agent's policy into distinct policy types. For example, if the Q-value of the cooperation action C is larger than that of defection D regardless of the partner's last action ($Q(C|o_C) > Q(D|o_C)$, $Q(C|o_D) > Q(D|o_D)$), we classify the agent as adopting the Always-Cooperate strategy; if the Q-value of the cooperation C action is larger than that of the defection action D when the partner cooperated last, but reverse it otherwise ($Q(C|o_C) > Q(D|o_C)$, $Q(D|o_D) > Q(C|o_D)$), we classify the agent as adopting the Tit-for-Tat strategy; and so forth. Thus, the agents' policies can be classified into four policy types: (1) Always-Cooperate (*ALL-C*), (2) Tit-for-Tat (*TFT*), where the agent copies the partner's last recorded action, (3) reverse Tit-for-Tat (*R-TFT*), where the agent does exactly the opposite, and (4) Always-Defect (*ALL-D*).

To analyse how the gameplay strategies develop and stabilise within the population, we created a box plot to show the number of agents that adopted each policy type during the training across various learning phases (Figure 5). In the



(a) Observer Model

(b) Local Information Model

Figure 6: Average Out-Degree and In-Degree of the unweighted partner selection graph across episodes by policy types under the Observer Model and the Local Information Model. The figures reflect the number of distinct agents an agent selects actively or is selected passively in one episode. For the out-degree, agents' partner selection under the Observer Model is constantly more dispersed than that of the Local Information Model in all policy types. For the in-degree, the *ALL-C* agents receive more selection under both models. The selection received by *TFT* agents is significantly higher under the Observer Model.

first phase, the learnt policy types are essentially uniform in the population. Agents have to spend some time identifying the *ALL-C* agents at the beginning, thus we can see a higher number of mutual defections (D, D) in the interaction. In the second phase, the distribution of learnt policy types has changed slightly. We can observe that the outcomes of mutual cooperation (C, C) and exploitation (D, C) are equally popular, showing that partner selection is learnt effectively. During the period, *ALL-C* and *TFT* agents are able to identify and cooperate with each other, *ALL-D* and *R-TFT* agents are able to identify the *ALL-C* agents and attempt to exploit them. Thus, the number of *ALL-C* agents has slightly decreased. In the third phase, the number of *TFT* agents starts to rise. This is in line with the observation that mutual cooperation (C, C) grows sharply. When agents are switching their policy type during the learning, it appears that *TFT* agents are getting an advantage with respect to the other policy types, which contributes to the rise in the number of *TFT* agents. In the fourth phase, the number of *TFT* agents becomes predominant, followed by the *ALL-C* agents. As agents cannot infer others' policy types through observation, the increase of *TFT* agents has largely affected the payoff of *ALL-D* agents, and this drives the defectors to cooperate over time; thus, cooperation emerges and is sustained.

Tracing the Interaction Graph

In this subsection, we compare how the partner selection dynamics evolve between agents interacting under the Observer Model and the Local Information Model. By examining the partner selection graph in a representative simulation (please refer to the Technical Appendix for a graphical representation), we observe that partner selection is more diverse under the Observer Model, whereas in the Local Information Model, agents frequently choose the same partners. In the Local Information Model, limited information about others leads agents to adopt a more conservative strategy, often preferring to select the partner they have already interacted with. This becomes a disadvantage when their partner

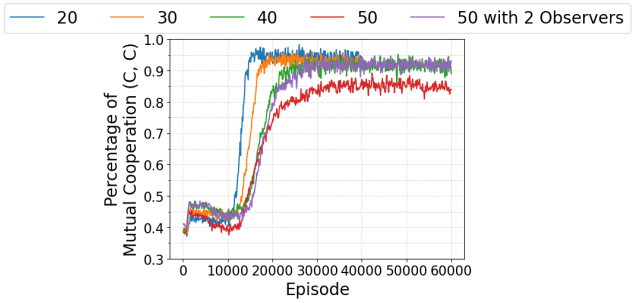
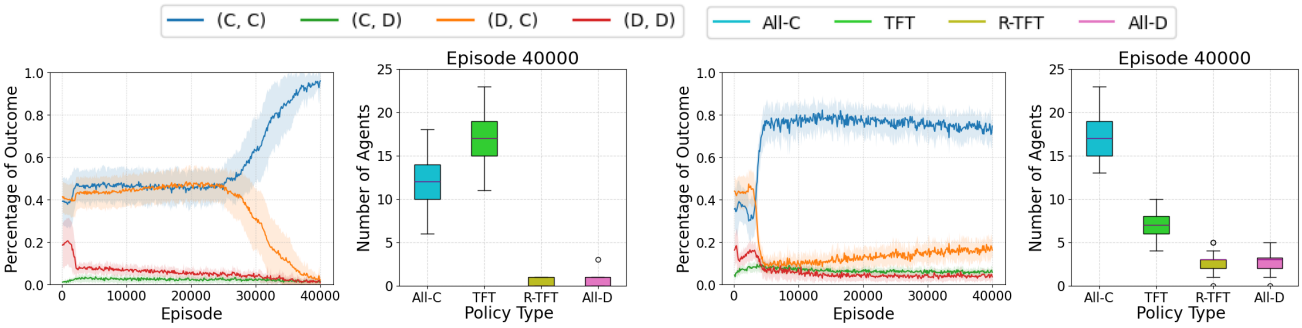


Figure 7: Percentage of mutual cooperation (C, C) across episodes with different population sizes. The percentage slightly deteriorates as population size increases. The percentage drops to about 85% for 50 agents. It rises back to 90% when we increase the number of observers to 2.

switches policy type. When the information becomes less reliable, the majority of agents will choose to defect. Under the Observer Model, each agent has a chance, on average, to become an observer at every round and thus receives two additional observations of others' actions. This additional information encourages agents to select different agents during the partner selection stage; therefore, we can see that the partner selection graph assumes a more regular structure.

The same conclusion is supported by looking at the out-degree of the unweighted partner selection graph. Figure 6 plots the average Out-Degree and In-Degree of the unweighted partner selection graph across episodes by policy types under the Observer Model and the Local Information Model. The figures reflect the number of distinct agents an agent selects actively or is selected by in an episode. As for the out-degree, agents' partner selection under the Observer Model is constantly more dispersed than in the Local Information Model, and this is true for all policy types. Agents are selecting on average 4.26 different partners in the Observer Model, while only 1.5 in the Local Information Model. As



(a) $\epsilon = 0.005$ (b) $\epsilon = 0.05$

Figure 8: The percentage of outcomes in PD across episodes and the box plots of policy type at the end of simulations under different exploration rates. A small exploration rate slows the rate of convergence. A large exploration rate promotes the adoption of the *ALL-C* policy type, limiting the cooperation rate from growing in the long run.

for in-degree, *ALL-C* agents are selected more often under both models. The average in-degree for *ALL-C* agents starts around 10 and gradually drops to 4 under the Observer Model, where the population is mainly composed of *ALL-C* and *TFT* agents at the end of the simulations. For the Local Information Model, the average in-degree for *ALL-C* agents rises to 18 and then stabilises at around 9, where the population is mainly composed of *ALL-D* agents with a few other types at the end of the simulations. Selection of *TFT* agents is significantly higher under the Observer Model across the learning period. The successful partnership of *ALL-C* and *TFT* agents is essential to the survival of *TFT* agents, which in turn is crucial for punishing the *ALL-D* policy.

The Effect of Population Size

In this and the next subsection, we study the robustness of our results, focusing first on the effect of population size in the Observer Model. We repeat the same experiment with population sizes of $20-50$. Figure 7 presents the percentage of mutual cooperation (C, C) across episodes with different population sizes. We see that cooperation still emerges even when the population size rises to 50. As the population size increases, the rate of convergence slightly deteriorates. The cooperation rate drops to about 85% in the case of 50 agents. We conduct further simulations increasing the number of observers. When this is increased to 2 (purple line), the cooperation rate rises back to above 90%.

The Effects of Learning and Exploration Rates

We look at two key hyperparameters: learning and exploration rates. We then test a lower and a higher learning rate ($\alpha = 0.001, 0.01$) and exploration rate ($\epsilon = 0.005, 0.05$).

Figure 8 shows the percentage of outcomes in PD across episodes and the box plots of policy type at the end of simulations under different exploration rates. The effects of changing the magnitude of both parameters are quite similar; for space reasons, we relegate the representation of learning rates to the Appendix. When agents adopt a lower exploration rate, the rate of convergence slows down. Mutual cooperation starts to grow at $25,000$ instead of at $12,500$ in the original setting. On the other hand, when agents adopt higher learning and exploration rates, we can see that the convergence rate speeds up. However, the percentage of mutual cooperation is kept below 80%. A high learning or exploration rate causes more agents to adopt the *ALL-C* policy, and limits the capacity of the population to resist defectors.

Conclusions and Future Work

Observability was identified as an important feature to sustain cooperation using Reinforcement Learning. But to what extent can we reduce it and still achieve socially desirable behaviour? In our novel Observer Model agents gather information through direct gameplay with others and by (limitedly) observing others' gameplays. We show that direct experience alone is not sufficient for the emergence of cooperation, particularly in large societies. However, allowing as few as one observer per gameplay leads to significant improvements, enabling the population to achieve and sustain robust cooperation across varying population sizes. By inspecting different phases of learning, we showed the nuanced dynamics of partner selection under limited observability and how it nurtures cooperation-promoting strategies.

This work demonstrates the key role of information availability for strategic partner selection geared to the emergence of cooperation. However, many questions still need to be answered. First and foremost, further robustness analysis is required, where the game itself is modified, e.g., through rewards or allowed selection frequency. Future work shall theoretically study the exact critical level of observability required to trigger cooperation. Potential directions also include investigating alternative observation mechanisms, such as recommendations from third-party, or via constrained observation networks where agents can only observe their neighbours. An extension to multi-player dilemmas, such as the common-pool resource problem (Pérolat et al. 2017), is also an important avenue to test the robustness of the Observer Model. Finally, exploring the impact of long-term observations and their impact on the emergent strategy types is a natural direction.

Acknowledgments

CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG-2023-050. FPS acknowledges the support of the Dutch Research Council (NWO) through project OCENW.M.22.322. MM acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) EP/X028569/1.

References

Anastassacos, N.; García, J.; Hailes, S.; and Musolesi, M. 2021. Cooperation and Reputation Dynamics with Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'21)*.

Anastassacos, N.; Hailes, S.; and Musolesi, M. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*.

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic.

Bara, J.; Turrini, P.; and Andrighetto, G. 2022. Enabling imitation-based cooperation in dynamic social networks. *Autonomous Agents and Multi-Agent Systems*, 36(2): 34.

Barclay, P.; and Willer, R. 2007. Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, 274(1610): 749–753.

Barfuss, W.; Flack, J.; Gokhale, C. S.; Hammond, L.; Hilbe, C.; Hughes, E.; Leibo, J. Z.; Lenaerts, T.; Leonard, N.; Levin, S.; et al. 2025. Collective cooperative intelligence. *Proceedings of the National Academy of Sciences*, 122(25): e2319948121.

Bazzan, A. L.; Peleteiro, A.; and Burguillo, J. C. 2011. Learning to cooperate in the Iterated Prisoner's Dilemma by means of social attachments. *Journal of the Brazilian Computer Society*, 17(3): 163–174.

Bowles, S.; and Gintis, H. 2013. *A Cooperative Species*. Princeton University Press.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.

Dasgupta, N.; and Musolesi, M. 2025. Investigating the impact of direct punishment on the emergence of cooperation in multi-agent reinforcement learning systems. *Autonomous Agents and Multi-Agent Systems*, 39(1): 1–37.

Fan, L.; Song, Z.; Wang, L.; Liu, Y.; and Wang, Z. 2022. Incorporating social payoff into reinforcement learning promotes cooperation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(12).

Fatima, S.; Jennings, N. R.; and Wooldridge, M. J. 2024. Learning to Resolve Social Dilemmas: A Survey. *Journal of Artificial Intelligence Research*, 79: 895–969.

Gilbert, N. 1995. Emergence in social simulation. In Gilbert, N.; and Conte, R., eds., *Artificial Societies: The Computer Simulation Of Social Life*. Routledge.

Graser, C.; Fujiwara-Greve, T.; García, J.; and Van Veelen, M. 2025. Repeated games with partner choice. *PLOS Computational Biology*, 21(2): e1012810.

Hilbe, C.; Schmid, L.; Tkadlec, J.; Chatterjee, K.; and Nowak, M. A. 2018. Indirect reciprocity with private, noisy, and incomplete information. *Proceedings of the national academy of sciences*, 115(48): 12241–12246.

Kollock, P. 1998. Social dilemmas: The Anatomy of Cooperation. *Annual Review of Sociology*, 24(1): 183–214.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS'17)*.

Leung, C.; Lenaerts, T.; and Turrini, P. 2024. To Promote Full Cooperation in Social Dilemmas, Agents Need to Unlearn Loyalty. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI'24)*, 111–119.

Leung, C.; and Turrini, P. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS'24)*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Nowak, M. A. 2006. Five Rules for the Evolution of Cooperation. *Science*, 314(5805): 1560–1563.

Pérolat, J.; Leibo, J. Z.; Zambaldi, V. F.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, 3643–3652.

Perreau de Pinninck, A.; Sierra, C.; and Schorlemmer, M. 2010. A multiagent network for peer norm enforcement. *Autonomous Agents and Multi-Agent Systems*, 21(3): 397–424.

Perret, C.; Krellner, M.; and Han, T. A. 2021. The evolution of moral rules in a model of indirect reciprocity with private assessment. *Scientific Reports*, 11(1): 23581.

Peysakhovich, A.; and Lerer, A. 2018. Towards AI that Can Solve Social Dilemmas. In *AAAI Spring Symposia*.

Pujol, J. M.; Sangüesa, R.; and Delgado, J. 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)*.

Rand, D. G.; Arbesman, S.; and Christakis, N. A. 2011. Dynamic social networks promote cooperation in experiments

with humans. *Proceedings of the National Academy of Sciences*, 108(48): 19193–19198.

Ren, T.; Yao, X.; Li, Y.; and Zeng, X.-J. 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)*.

Sabater, J.; and Sierra, C. 2002. Reputation and Social Network Analysis in Multi-Agent Systems. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'02)*.

Salazar, N.; Rodriguez-Aguilar, J. A.; Arcos, J. L.; Peleteiro, A.; and Burguillo-Rial, J. C. 2011. Emerging Cooperation on Complex Networks. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'11)*.

Sandholm, T. W.; and Crites, R. H. 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems*, 37(1-2): 147–166.

Santos, F. C.; and Pacheco, J. M. 2005. Scale-Free Networks Provide a Unifying Framework for the Emergence of Cooperation. *Physical Review Letters*, 95(9): 098104.

Santos, F. C.; Pacheco, J. M.; and Lenaerts, T. 2006. Cooperation Prevails When Individuals Adjust Their Social Ties. *PLOS Computational Biology*, 2(10).

Santos, F. P.; Pacheco, J. M.; and Santos, F. C. 2018. Social Norms of Cooperation With Costly Reputation Building. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18)*.

Santos, F. P.; Pacheco, J. M.; and Santos, F. C. 2021. The complexity of human cooperation under indirect reciprocity. *Philosophical Transactions of the Royal Society B*, 376(1838): 20200291.

Segbroeck, S. V.; Santos, F. C.; Nowé, A.; Pacheco, J. M.; and Lenaerts, T. 2009. The coevolution of loyalty and cooperation. In *Proceedings of the 11th IEEE Congress on Evolutionary Computation (CEC'09)*.

Sigmund, K. 2010. *The Calculus of Selfishness*. Princeton University Press.

Smit, M.; and Santos, F. P. 2024. Learning fair cooperation in mixed-motive games with indirect reciprocity. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI'24)*.

Sylwester, K.; and Roberts, G. 2010. Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, 6(5): 659–662.

Tennant, E.; Hailes, S.; and Musolesi, M. 2024. Dynamics of Moral Behavior in Heterogeneous Populations of Learning Agents. In *Proceedings of the 7th AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024)*.

Van Lange, P. A.; Joireman, J.; Parks, C. D.; and Van Dijk, E. 2013. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, 120(2): 125–141.

Wang, J.; Suri, S.; and Watts, D. J. 2012. Cooperation and assortativity with dynamic partner updating. *Proceedings of the National Academy of Sciences*, 109(36): 14363–14368.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine Learning*, 8: 279–292.

Zhang, B.-Y.; Fan, S.-J.; Li, C.; Zheng, X.-D.; Bao, J.-Z.; Cressman, R.; and Tao, Y. 2016. Opting out against defection leads to stable coexistence with cooperation. *Scientific Reports*, 6: 35902.

Zheng, X.-D.; Li, C.; Yu, J.-R.; Wang, S.-C.; Fan, S.-J.; Zhang, B.-Y.; and Tao, Y. 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: 12–17.