Enhancing Cooperation through Selective Interaction and Long-term Experiences in Multi-Agent Reinforcement Learning

Tianyu Ren, **Xiao-Jun Zeng**
University of Manchester
{tianyu.ren, x.zeng}@manchester.ac.uk

Abstract

The significance of network structures in promoting group cooperation within social dilemmas has been widely recognized. Prior studies attribute this facilitation to the assortment of strategies driven by spatial interactions. Although reinforcement learning has been employed to investigate the impact of dynamic interaction on the evolution of cooperation, there remains a lack of understanding about how agents develop neighbour selection behaviours and the formation of strategic assortment within an explicit interaction structure. To address this, our study introduces a computational framework based on multi-agent reinforcement learning in the spatial Prisoner's Dilemma game. This framework allows agents to select dilemma strategies and interacting neighbours based on their long-term experiences, differing from existing research that relies on preset social norms or external incentives. By modelling each agent using two distinct Q-networks, we disentangle the coevolutionary dynamics between cooperation and interaction. The results indicate that long-term experience enables agents to develop the ability to identify non-cooperative neighbours and exhibit a preference for interaction with cooperative ones. This emergent self-organizing behaviour leads to the clustering of agents with similar strategies, thereby increasing network reciprocity and enhancing group cooperation.

1 Introduction

The emergence of cooperation is fundamental to human civilization and evident in various biological and multi-agent systems [Rand and Nowak, 2013; Kramár *et al.*, 2022]. However, maintaining cooperation in social dilemmas is challenging due to the conflict between individual interests and collective welfare [Sigmund, 2010]. A key focus in cooperation research involves understanding the conditions under which individuals favour altruism over self-interest, with reciprocity identified as a crucial element. In this context, reciprocity is examined through repeated interactions limited to direct neighbours due to physical or social constraints, where the behaviours of individuals are shaped by the actions of their counterparts [Perc *et al.*, 2013]. Studies using evolutionary game theory (EGT) demonstrate that network structure can promote altruistic behaviour through spatial reciprocity, clustering similar strategies and reducing exploitation by free-riders [Santos *et al.*, 2006; Wang *et al.*, 2013]. Furthermore, considering individuals' capability to interact with neighbours selectively, recent research emphasizes the significance of dynamic interaction mechanisms in the evolution of cooperation [Su *et al.*, 2022; Rand *et al.*, 2011; Sylwester and Roberts, 2013]. Such neighbour selection behaviours notably transform learning dynamics, thereby allowing agents to modify their interaction structures in response to evolving cooperative scenarios.

Despite advancements, explanations for the spontaneous emergence of cooperative behaviours and selective interactions remain limited. Traditional game theory and EGT emphasize social learning, where agents imitate nearby successful strategies [Sigmund *et al.*, 2010; Li *et al.*, 2016], overlooking learning through trial-and-error [Metcalfe, 2017]. Additionally, empirical studies employing experience-based learning face challenges in developing enduring strategies, particularly due to the complexities encountered in large population iterations [McKee *et al.*, 2023]. To unravel the intricate coevolutionary dynamics of agent behaviours, there is an increasing interest in leveraging advanced deep reinforcement learning (RL) algorithms [Perolat *et al.*, 2017; Du *et al.*, 2023; Willis *et al.*, 2023]. These methods are not only used to examine agents' decision-making processes but also to investigate the emergence of their behaviours [Köster *et al.*, 2022]. However, RL agents typically optimize personal policies, which may lead to a reduction in global optimization [Lowe *et al.*, 2017]. Although studies introducing mechanisms like reputation [Anastassacos *et al.*, 2020] and moral rewards [Tennant *et al.*, 2023] to address these issues, these specialized approaches exhibit restricted applicability. Therefore, it is crucial to have a deeper understanding of the learning dynamics of agent behaviours and encourage cooperation.

In this study, we construct a computational model using deep Q-learning (DQN) [Mnih *et al.*, 2015] to explore how agents can simultaneously learn both interaction and dilemma strategies from their long-term experiences. These agents, modelled as artificial neural networks, learn behavioural poli-

Version with Appendix: https://arxiv.org/abs/2405.02654 cies and obtain rewards in a spatial Prisoner's Dilemma Game (PDG) setting within a multi-agent reinforcement learning (MARL) environment. Unlike previous studies that depended on predefined social norms or explicit external incentives, our approach highlights the significance of temporal factors and historical information in influencing agent decision-making. Initially, agents have no prior knowledge regarding the actions or game states, hindering their ability to assess the consequences of their actions and respond effectively. Throughout the training process, they must learn the causality between their actions, observations, and rewards from local environment observation. To aid this learning process, we introduce a utility function that integrates self-learning and social learning, reflecting the interplay of personal preferences and the influence of others in shaping human behaviour.

The experimental results demonstrate that RL agents trained in our framework effectively differentiate between cooperative neighbours and those who are free-riders. Their preference for building connections with cooperators bolsters network reciprocity, contributing to the formation of strategy clusters in network-structured populations. This finding aligns with existing EGT research, emphasizing the importance of strategy assortment in promoting cooperation. Moreover, the trained agents achieve superior cooperation levels and greater average payoff compared to the EGT baseline. Further, we observe that increased efficiency in learning is correlated with the length of memory experiences. For a detailed comparison between our RL model and EGT approaches, refer to the Supplementary Information (SI) A.2.

Our work offers three key contributions. Firstly, it reveals the coevolutionary dynamics of cooperation and interaction strategies within a spatial PDG framework, demonstrating that RL agents can learn effective interaction mechanisms to enhance network reciprocity and cooperation. Secondly, it sheds light on how extensive long-term experiences positively influence group cooperation. Finally, the MARL training environment we developed sets the stage for future explorations into diverse aspects of pro-social cooperative behaviour.

2 Related Works

2.1 Evolutionary Game Theory

EGT is crucial for exploring the evolution of cooperation among self-interested individuals. It expands on traditional game theory by considering extended interactions and strategy dynamics, exploring the emergence and stability of cooperative behaviours. Notably, Nowak [2006] identifies five mechanisms central to understanding cooperation evolution. One vital aspect noted is the emergence of cooperation on network structures, wherein individuals predominantly interact with their immediate neighbours [Perc *et al.*, 2017].

Inspired by the dynamic nature of social interactions, numerous studies have explored the coevolutionary dynamics of cooperation by integrating strategy evolution with network changes. In this domain, the concept of network assortativity has been highlighted, revealing that agents with similar strategies often connect, thereby boosting group cooperation [Tanimoto, 2013; Ren and Zheng, 2021]. Research like Su et al. [2022] explores how cooperative strategies evolve and spread, particularly in unidirectional interactions, shedding light on shaping social interactions to promote cooperation. However, these often neglect the self-learning aspect of agents, which is crucial in real-life where directly copying strategies is impractical. To address this gap and accurately depict cooperation emergence, we applied the MARL framework, enabling agents to learn both dilemma and interaction strategies through environmental observation independently.

2.2 Human Experiments

While EGT is essential for examining evolutionary trajectories and conditions favouring cooperation, incorporating empirical data brings psychological nuances to these models [Köbis *et al.*, 2019], which focus on imposed interaction structures and psychological mechanisms beyond laboratory settings [Rand and Nowak, 2013]. Surprisingly, behavioural studies indicate that participants in structured settings tend to randomly change strategies instead of copying higher-payoff neighbours [Traulsen *et al.*, 2010], disrupting the clustering process and rendering cooperation less advantageous. Addressing this, Rand et al. [2011] found that dynamic interactions enhance multilateral cooperation by encouraging links with cooperators over defectors, promoting strategy clustering. However, such laboratory experiments often encounter challenges in terms of scalability and struggle with complex, larger-scale, or long-term scenarios [Moffatt *et al.*, 2009].

2.3 Multi-agent Reinforcement Learning

Recent RL applications also focus on understanding the emergence of cooperation by integrating spatial and temporal dynamics relevant to realistic scenarios [Ren and Zeng, 2023; Vinitsky *et al.*, 2023; Tennant *et al.*, 2023], moving beyond traditional matrix games through the fusion of complex incentive structures [Leibo *et al.*, 2017; Jaques *et al.*, 2019; Abeywickrama *et al.*, 2023]. Studies have applied the intrinsic trial-and-error learning characteristic of RL to reformulate agent interaction strategies. For instance, Anastassacos et al. [2020] demonstrate that RL agents can learn an interaction strategy akin to Tit-for-Tat when partner selection is present, aiding in the maintenance of cooperation. Meanwhile, McKee et al. [2023] used a graph neural network-based agent as a social planner, demonstrating the ability of deep RL to foster coordination and cooperation in a group.

Concurrently with our work, Ueshima [2023] employed two distinct Q-networks for each agent, differentiating interaction and dilemma strategies. However, our approach varies as follows: (1) we extend the model to allow each agent to interact with four potential neighbours, unlike their paired interaction focus; (2) we incorporate environments with an explicit interaction structure, considering network reciprocity; (3) they consider single-round observation input, while we take into account the agents' long-term experiences.

3 Background

3.1 Prisoner's Dilemma Game

The PDG is a fundamental paradigm in EGT [Rapoport *et al.*, 1965], representing a typical decision-making scenario where agents balance individual benefits against collective welfare.

Fundamentally, the PDG is characterized as a symmetric matrix game, representing interactions between pairs of individuals within a population. Each participant faces a choice: to cooperate (C), incurring a cost c while providing a benefit b to others, or to defect (D), avoiding the cost while exploiting those who cooperate (with $b > c > 0$). The corresponding payoff matrix can be summarized as

$$\mathcal{M}_p = \begin{bmatrix} R & S \\ T & P \end{bmatrix} \tag{1}$$

where mutual cooperation yields a reward $R = b - c$, while mutual defection result in $P = 0$. Unilateral cooperation against a defector incurs a cost $S = c$, wherease the defector gains $T = b$. This payoff matrix forms four classical game structures [Wang *et al.*, 2015]: PD, chicken, harmony and stag hunt, each defined by specific payoff inequalities. In the PD, the conditions $T > R > P > S$ and $2R > T + S$ hold. Our model employs a weak PDG with $T = b$ $(1 \leq b \leq 2)$, $R = 1$, and $P = S = 0$ [Nowak and May, 1993]. Here, the parameter b directly assesses the strength of the dilemma, given $c = 0$. In a single-shot PDG, defection is the dominant strategy, leading to a Nash Equilibrium of defection, despite cooperation could yield a Pareto improvement. Considering that players often engage in repeated iterations with the same counterparts, the conventional PD can be extended to the iterated prisoner's dilemma (IPD) format. In this work, the dilemma strategy of agent i at timestep t is represented by a two-dimensional unit vector $a_{d_i}(t)$, with $a_{d_i} = [1, 0]^T$ indicating cooperation and $a_{d_i} = [0, 1]^T$ signifying defection.

3.2 MARL Markov Game

Within MARL, the IPD is conceptualized as a multi-agent extension of Markov decision processes (MDPs), termed partially observable general-sum Markov games [Littman, 1994]. Here, agents have observations limited to their local environment. Formally, a N-player MDP is defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \mathcal{T}, \gamma, \mathcal{R} \rangle$, where $\mathcal{S}$ is a set of joint states for all agents, $\mathcal{A}_1, \ldots, \mathcal{A}_N$ represent joint actions, and $\mathcal{R}$ is the reward function. The function $\mathcal{O} : \mathcal{S} \times \{1 \ldots, N\} \to \mathbb{R}^d$ maps each player's d-dimensional view of the state space. In a given state, each agent i selects an action from $\mathcal{A}_i$, and the dynamics of MDP are determined by the stochastic transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_N \to \triangle(\mathcal{S})$, where $\triangle(\mathcal{S})$ represents the set of discrete probability distributions over $\mathcal{S}$. The objective of each agent i is to learn a policy $\pi_i : \mathcal{O}_i \to \triangle(A_i)$ to maximize its extrinsic reward $r_i(s, a^1, \ldots, a^N)$ based on the agent's own observation, simplified as $\pi(a^i | o^i)$. This optimization is conducted while adhering to the joint policy of all agents. The long-term γ-discounted payoff for agent i under the joint policy $\overrightarrow{\pi} = (\pi_1, \ldots, \pi_N)$ from an initial state s_0 can be defined as:

$$V_i^{\overrightarrow{\pi}}(s_0) = \mathbb{E}_{\overrightarrow{a_t} \sim \overrightarrow{\pi}(\mathcal{O}(s_t)), s_{t+1} \sim \mathcal{T}(s_a, \overrightarrow{a}_t)} [\sum_{t=0}^{T} \gamma^t r_i(s_t, \overrightarrow{a}_t)] \tag{2}$$

where $\gamma \in [0, 1]$ represents the temporal discount factor, and T denotes the time horizon. Policies are optimized using trial-and-error interactions within the MARL environment to maximise cumulative long-term rewards.

3.3 Deep Q-Network

As an extension of Q-learning, DQN stands out as one of the most popular off-policy Deep RL algorithms, which utilizes an independent deep neural network to estimate Q-values [Mnih *et al.*, 2015]. Departing from the traditional tabular representation for Q-values of state-action pairs, DQN utilizes a parametrized Q-function $Q_\theta(s, a)$ to approximate Q-values. Each Q-network is parameterized by θ, representing the weights of the neural network. This approach incorporates the utilization of a replay memory buffer $\mathcal{D}$ to store past experiences and a target Q function $\overline{Q}$ to mitigate the risk of overestimating Q-values. The learning process for the optimal action-value function Q^* involves minimizing the loss.

$$\mathcal{L}_\theta = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} [(r + \gamma \max_{a'} \overline{Q}(s', a') - Q(s, a))^2]. \tag{3}$$

DQN can be directly extended to a multi-agent setting by having each agent i learn an independent Q function denoted as $Q_i : \mathcal{O}_i \times \mathcal{A}_i \to \mathbb{R}$. In line with the traditional Q-learning approach, DQN also adopts an ϵ-greedy policy to promote exploration. The policy for the i-th agent is parameterized as:

$$\pi_i(s) = \begin{cases} \arg\max_{a_i \in \mathcal{A}_i} Q_i(s, a) & \text{with probability } 1 - \epsilon \\ \mathcal{U}(\mathcal{A}_i) & \text{with probability } \epsilon \end{cases} \tag{4}$$

where $\mathcal{U}(\mathcal{A}_i)$ signifies a sample drawn from the uniform distribution over the action space $\mathcal{A}_i$.

4 Methodology

This model employs value-based optimization utilizing the DQN approach within a MARL environment. Throughout the training, agents interact with neighbours, exhibiting cooperative or defective behaviours across various episodes. As illustrated in Fig. 1, each agent i aims to learn a joint policy π_i concerning dilemmas and selection actions, informed by their local observations and long-term experiences. A comprehensive explanation of our methodology follows. [1]

4.1 Game Environment

Agents in our experiment are situated within an $L \times L$ square lattice with periodic boundary conditions. They are placed at specific spatial coordinates and can engage in interactions limited to their von Neumann neighbourhood. The graphical representation employs vertices to denote agents and edges to indicate the relationships between an agent and its four neighbours. The action space for each agent encompasses two strategies: the overall dilemma strategy and the specific interaction selection strategy. This dual contributes to the formulation of the agent's policy π_i expressed as:

$$\pi_i = (\pi_{s_i}, \pi_{d_i}) \tag{5}$$

where π_{s_i} refers to the interaction selection policy dictating whether to interact with its neighbours, while π_{d_i} is the dilemma policy that guides the agent in choosing between a cooperative or defective strategy.

[1]Code: https://github.com/itstyren/InteractionMARL-Coop

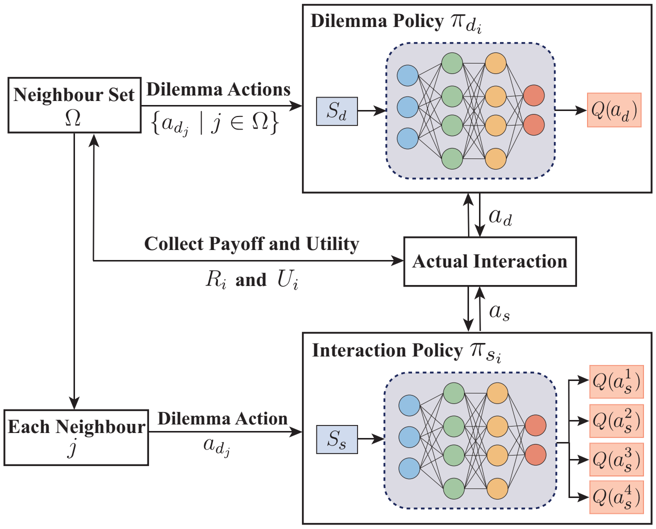

Figure 1: **Training framework for developing dilemma and interaction strategies.** Each iteration involves agent i choosing a dilemma strategy and selecting neighbouring agents for PDG engagement. Each agent uses two Q-networks: the dilemma policy network, which processes long-term actions in dilemmas by the agent and its neighbours, and the interaction selection network, which assesses neighbours' dilemma actions alongside the agent's previous interactions. The agent calculates the utility of its actions based on the rewards accumulated from past encounters.

Interaction and Dilemma Actions

During each timestep of an episode, agents participate in multiple pairwise interactions within the IPD game framework, consisting of two phases: neighbour selection and dilemma strategy determination. In the first phase, agent i chooses game partners through an interaction selection action denoted as $a_{s_i} \in \{0,1\}^4$, informed by previously observed game experiences with its neighbours. This selection action a_s^i is represented by the following bitstring of size 4:

$$a_{s_i} = (a_{s_i}^1, a_{s_i}^2, a_{s_i}^3, a_{s_i}^4). \tag{6}$$

Specifically, a bit value of 1 for $a_{s_i}^j$ signifies agent i choosing to interact with neighbour j, whereas a 0 implies no interaction. For example, a selection action of $a_s = (1,0,0,1)$ implies interaction only with the first and last neighbours. Importantly, actual interaction occurs only when both players decide to interact with each other, i.e., when $a_{s_i}^j = a_{s_j}^i = 1$.

In the second phase, each player selects either cooperation or defection as their dilemma strategy. Following the interaction phase, paired co-players engage in one round of PDG pairwise, utilizing the action pair $[a_{d_i}, a_{d_j}]$ where $a_d \in [C, D]$. Critically, the chosen binary dilemma strategy remains consistent across all interacted neighbours. In other words, a player cannot cooperate with one neighbour while defecting against another neighbour simultaneously.

Game Formulation and Reward

According to their selected dilemma and interaction actions, agent i receives an accumulative payoff at each timestep t by participating in multiple rounds of the PDGs with its currently interacted neighbours, as shown by:

$$r_i(t) = \sum_{j=0}^{n_i^t} a_{d_i}^T \mathcal{M}_p a_{d_j} \tag{7}$$

where $n_i^t = \sum_{j \in \Omega_i} a_{s_i}^j \times a_{s_j}^i$ denotes the number of interacted neighbours for agent i at timestep t, and Ω_i is the set of neighbors. In this model, we incorporate learning from previous interactions by employing a weighted moving average of past payoffs [Danku *et al.*, 2019]. Thus, the final payoff of agents i at each timestep not just based on the current round's payoff but also includes payoffs from the past m rounds:

$$R_i(t) = \frac{r_i^t + \sum_{m=1}^{M} \alpha^m R_{i,m}}{1 + \sum_{m=1}^{M} \alpha^m} \tag{8}$$

where α is a parameter that controls the rate of weight decay with increasing m, indirectly determining the memory length M. To effectively assess the emergent behaviours, M is restricted by the condition $M = \min\{n \mid \alpha^n < 0.01\}$. With $\alpha = 0$, agents focus only on the current round, indicating short memory. Conversely, as $\alpha \to 1$, agent memory extends to include all previous timesteps, capturing a comprehensive history of experiences.

In EGT research, the Fermi rule is commonly employed to model the dynamics of strategy evolution [Szabó and Tőke, 1998], reflecting social learning wherein neighbours tend to imitate the most successful observed policy. A detailed description is provided in SI B. To adapt these imitation dynamics to the RL context, we have modified the utility function of agent i to align with game payoffs, as formulated below:

$$U_i(t) = \frac{[\omega_i(a_d^t) + 1] R_i(a_{d_i}^t, a_{s_i}^t, s_i^t) - \omega_i(\tilde{a}_d^t) \overline{R}(\tilde{a}_d^t, a_s^t)}{\sum_{a_d \in A_d} \omega_i(a_d) + 1} \tag{9}$$

where $\tilde{a}_{d_i}$ represents a counterfactual dilemma action, condition on the actual action a_d taken by agent i. The function $\omega_i(\cdot)$ returns the number of neighbours performing a specific action. The term $\overline{R}(\tilde{a}_d^t, a_s^t)$ denotes the average payoff associated with the counterfactual action $\tilde{a}_d^t$ across the population at timestep t. Essentially, the agent raises a retrospective question: "Would a different past action have led to a more advantageous outcome?" This setting integrates aspects of social learning and RL, allowing agents to compare global information from group about the performance of different actions with their own localized experiences.

4.2 Training Approach

In our multi-agent PDG framework, the training methodology aligns with the well-established DQN approach, typically used in single-agent tasks. Our focus is on formulating a joint policy (π_s, π_d), which utilizes the combined action utilities to calculate the Q-loss for each policy network, guiding both dilemma strategy and interaction selection processes. See SI A.1 for a detailed elucidation of the training procedure.

Network Architecture

Each agent in our independent MARL setup is equipped with a memory buffer, storing experiences from the last M rounds, encompassing a record of both the agent's own actions and those of adjacent agents. The agent updates its memory at each timestep with recent feedback from the local environment, ensuring an accurate representation of its state. The selection and dilemma networks process inputs from long-term neighbour interactions and prior dilemma strategies, respectively. Additionally, both networks consider the dilemma strategies of four neighbouring agents, with actions represented through one-hot encoding and sequenced together.

In the selection phase, agents evaluate the state $s_s \in \mathbb{R}^{2\times16\times M}$, and in the dilemma phase $s_d \in \mathbb{R}^{2\times5\times M}$. They operate with two Q-networks configurations, formulating policies independently and without sharing parameters across agents. The architecture of each Q-network includes a dual-layer perception with 32 hidden units and employs the *tanh* activation function for nonlinear transformations.

Experiment Setup

During the training stage of our RL experiments, we employ a centralized training with decentralized execution approach [Lowe *et al.*, 2017]. This method allows agents to access global information regarding the average payoff for potential action in the PDG among the population, thereby facilitating an effective evaluation of their action utilities.

The training involved 900 agents, each equipped with two neural networks, ensuring a broad representation of cooperation dynamics at the population level. To generate experiences for agents, 10 parallel arenas were established. In each arena during every experimental trial, agents engage in interactions with their neighbours over $6,000$ episodes, each comprising 10 timesteps, resulting in a total of $60,000$ steps. At the end of each episode, sampled trajectories for agents were aggregated and subsequently forwarded to the respective learner. We compute the gradient by using the Adam optimizer [Kingma and Ba, 2014] with a linear annealing schedule of learning rate. To enhance the efficiency of the Q-learner in learning from experience replay, we also implement a common practice known as prioritized experience replay [Schaul *et al.*, 2015] within the DQN framework. Additional details on hyperparameters, please refer to the SI A.2.

5 Results and Discussion

In this section, we present the outcomes of our experimental investigations, which provide evidence supporting the hypothesis that incorporating the interaction selection action with the dilemma action through RL promotes network reciprocity and cooperation evolution in a spatial PDG setting.

5.1 Experimental Setup

Our experiment employs a square lattice setup, randomly assigning agents as cooperators or defectors with equal likelihood. The primary evaluation metric is the fraction of cooperative agents in the population, representing the achieved level of overall cooperation. For robustness, we average the outcomes of the final 10 episodes over the entire training duration. All experiments were replicated five times to ensure replicability. Unless otherwise stated, a memory weight of 0.6 is assigned, incorporating experiences from the previous four rounds as network input.

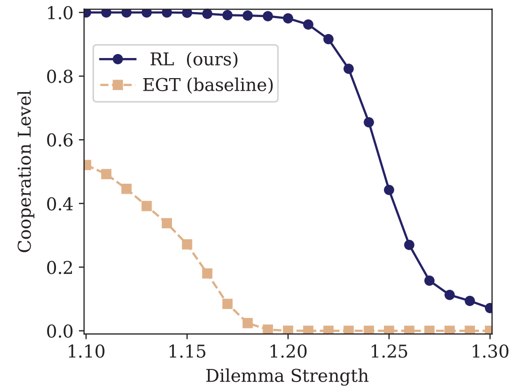

Figure 2: **RL-based training approach (ours) promotes cooperation more effectively than the EGT (baseline) method.** The EGT (orange) represents agents solely calculating cumulative payoffs and adjusting dilemma actions through social learning. In contrast, the implementation of effective dilemma and selection policies, guided by RL (blue), has significantly enhanced the level of cooperation within the population. Our RL-based method maintains full cooperation in the population until dilemma strength exceeds 1.2.

5.2 Promotional Effect of RL on Cooperation

To evaluate the effectiveness of our proposed method, we first train a population to learn a combined policy, including both dilemma and interaction strategies under various dilemma strength conditions. Figure 2 demonstrates that the application of RL in coordinating interaction and dilemma strategies enables the population to sustain a high cooperation level successfully. Notably, this approach proves robust, maintaining its efficacy even in scenarios characterized by increased dilemma intensity. As shown, the MARL system transitions from a complete cooperation phase to a mixed strategy phase exclusively, when the dilemma strength $b > 1.2$.

For comparison, the evolutionary outcomes of conceptually similar models from existing literature are used as a benchmark [Danku *et al.*, 2019]. Within EGT framework, agents evaluate equivalent lengths of past payoffs and emulate the most successful dilemma strategy observed in their neighbourhood (a detailed description of EGT methodology, refer to the SI B). It is noteworthy that even when the dilemma intensity is reduced to $b = 1.1$, only 54% EGT agents opt for cooperative strategies. Moreover, ablation experiments detailed in SI D.3 demonstrate that agents utilizing RL to learn dilemma strategies exclusively underperform in comparison to the model proposed in this study. These results suggest that lacking selective interaction in the PDG and the mere imitation of neighbouring successful strategies are insufficient to achieve optimal performance within the population. Additional MARL-based benchmarks are reported in SI D.5.

5.3 Evolutionary Dynamics of Cooperation

We next investigate the evolutionary dynamics and outcomes of overall dilemma strategies within the population, focusing on the evolution trajectories and average payoffs in four representative dilemma conditions. Figure 3(a) reveals a rapid initial decline in the population's cooperation level as the dilemma intensity increases. However, RL agents employing

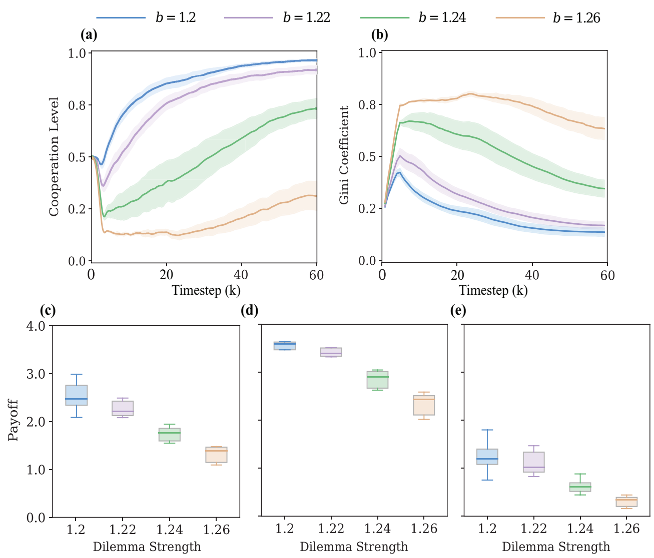

Figure 3: **The evolution of cooperation and associated payoffs across varying dilemma strengths.** In all scenarios, the fraction of cooperators first decreases and then increases over time, coinciding with reduced average individual payoffs and increased inequality as dilemma strength intensifies. The evaluation encompasses evolutionary trajectories of (a) cooperation level and (b) the Gini Coefficient, alongside metrics including (c) average group payoffs and (d)-(e) payoffs for trained cooperators and defectors, with dilemma strength varying from $b = 1.20$ to 1.26.

additional interaction strategies exhibit two evident phases in their learning process, aligning with observations from previous EGT experiments [Wang *et al.*, 2013]: the END period and the EXP period. During the END period, cooperative agents resist the invasion of defectors, and successful cooperators convert those defectors into cooperators in the EXP period. In our experiments, the former period is characterized by a rapid decrease in cooperation levels in the first $3,000$ training step, followed by a phase where these levels rise unless defectors completely dominate in the early stages. Consequently, at the end of the EXP period, the cooperation level in the population significantly decreases, falling from 0.987 to 0.294 as b rises from 1.20 to 1.26.

The dilemma strength also significantly affects the distribution of individual payoffs, leading to a bifurcated process. As shown in Figure 3(b), the group Gini Coefficient exhibits a temporal evolution, initially increasing and subsequently decreasing, hinting at a correlation between payoff equality and the evolution of cooperation. Notably, a large fraction of cooperators contributes to high levels of group equality. In Figures 3(c)-(e), the focus is on evaluating the average payoff for the population, as well as the separate payoffs for cooperative and defective individuals. There is a general decrease in the average payoff as the dilemma increases, which aligns with expectations, given that cooperators are primarily the contributors to the group payoff. Cooperative individuals, however, show greater resilience to tougher dilemma conditions. Specifically, with an increase in dilemma strength from $b = 1.20$ to 1.22, the average payoff per episode decreases from 2.67 to 2.25, but this can be lessened by adopting an effective interaction mechanism, resulting in a smaller decrease in cooperative payoff from 3.52 to 3.41, compared to a substantial drop for defectors from 1.44 to 1.10. This trend indicates the potential of RL agents to develop interaction policies for dilemma scenarios, mitigating adverse effects on their payoff. For detailed information on the average payoff from the trained population, refer to Table S1 in the SI.

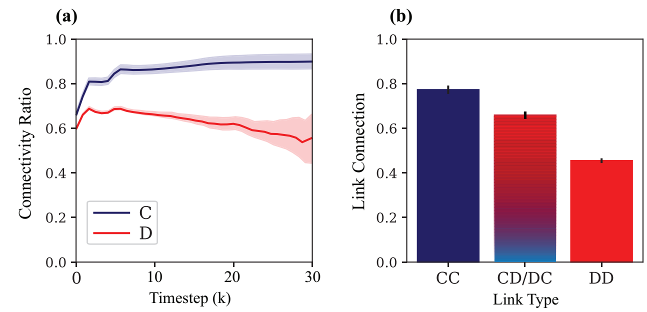

Figure 4: **Temporal evolution of strategy connectivity and actual link interactions.** RL agents demonstrate enhanced interaction capabilities, increasing connections with cooperative neighbours. (a) The average connectivity ratio for cooperators and defectors in the first half of the total timestep. (b) The frequency of actual link connections between dilemma strategies during the first ten episodes. The dilemma strength is set to $b = 1.20$.

5.4 Efficiency of Learned Interaction Patterns

To elucidate the role of interaction selection in network reciprocity and cooperation, we analyze strategy connection and distribution patterns in Figures 4 and 5. Through participation in PDG with selective neighbours, RL agents develop policies for distinguishing cooperators from defectors, thereby boosting spatial reciprocity and cooperation. Figure 4(a) depicts how the disparity in average connections between cooperators and defectors increases during the initial half of the total timestep. This leads to a preference for forming connections with cooperators, irrespective of the dilemma strategy chosen by agents. In Figure 4(b), we evaluate the actual link connections across different link types during the first 10 training episodes, signifying frequency where the chosen co-player reciprocally opts for interaction within the same round. As illustrated, the occurrence of interaction between two neighbouring individuals who both employ the defective strategy (DD link) is merely 45.48%. In contrast, interactions between two cooperators (CC link) can increase to as high as 77.29%. For measurement metrics of agent interactions, see SI A.3.

We next provide intuitive evidence regarding the previously described learned interaction policy and its role in enhancing spatial reciprocity by illustrating the spatial coevolution of the dilemma and interaction strategies within the population over time. Initially, in the END phase depicted in Figures 5(a) and (e), cooperators resist the invasion of defectors by forming small clusters, yet lack an effective neighbour selection strategy. As training progresses, RL agents learn to adapt their interaction strategies in response to neighbouring dilemma strategies, enhancing the influence of spatial reciprocity in promoting the evolution of cooperation. Figures 4 and 5 (f)-(h) show that, in the EXP phase, individuals within

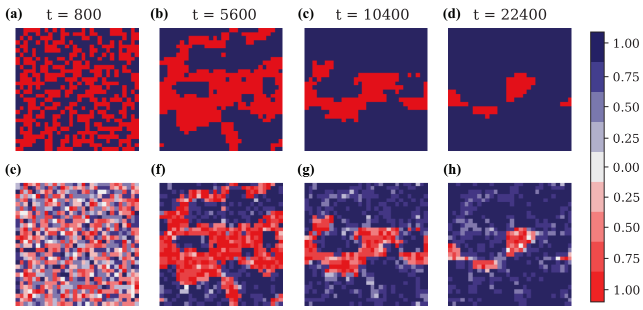

Figure 5: **Snapshots of the spatial evolution of strategies and their connections.** Cooperative individuals resist defector incursions by forming and expanding clusters. Panels (a)-(d) depict strategy distributions; (e)-(h) illustrate corresponding strategy connections at identical timesteps. Pixels represent agents as cooperators (blue) and defectors (red), with strategy connectivity ratio varying from 0 (shallow) to 1 (deep). The results are obtained for $b = 1.20$.

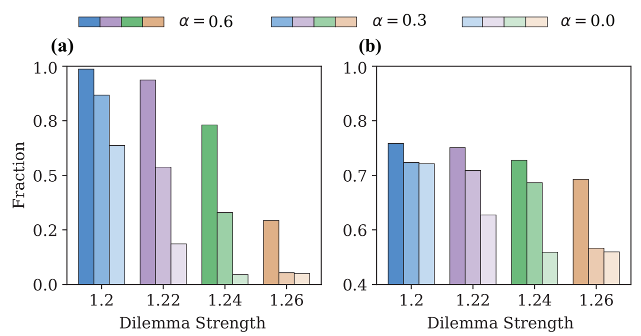

Figure 6: **Average cooperation level and effective connection for cooperators with varying experience lengths.** Incorporating longer experiences as input enhances group cooperation and cooperative interactions. (a) Post-training cooperation levels in the population. (b) Frequency of cooperators participating in PDG across the initial 20 episodes. These results are obtained for four dilemma strength scenarios, with memory weight α ranging from 0 to 0.6.

the cooperative cluster show a higher tendency to engage in the PDG with nearby cooperators, while those at the periphery are more likely to avoid interactions with defectors. This selective interaction mechanism favours cooperators and limits the payoff obtained from free-riding behaviours. Consequently, defectors gradually switch their dilemma strategies, leading to the expansion of cooperative clusters.

5.5 Role of Long-term Experiences

In prior experiments, RL agents use experiences from the last four rounds as network input. Finally, we investigated the influence of varying memory lengths on their learning dynamics, with findings presented in Figure 6. Three memory weights were assessed, with α varying from 0.6 to 0, representing input scenarios based on observations from the last, second, and fourth rounds, respectively. Notably, a positive correlation is observed between the group cooperation level and the memory length of trained agents. Figure 6(a) illustrates that, for instance, when dilemma strength increased from 1.2 to 1.22, agents recalling four-time steps maintained group cooperation effectively. Conversely, those relying solely on current information saw a substantial decrease in cooperation levels, dropping from 0.64 to 0.19.

Furthermore, Figure 6(b) echoes the findings of Figure 4, demonstrating that successful cooperation benefits from a preference for efficient communication with neighbouring cooperators. Moreover, the efficiency of interaction selection is also influenced by the length of agent memory. Longer observation inputs are shown to improve the average interaction ratio among cooperators in the END and initial EXP stages, which supports network reciprocity and contributes to the formation and development of cooperative clusters. Interestingly, our findings also reveal that RL individuals consistently and effectively avoid interactions with defectors, irrespective of memory length and dilemma intensity (detailed in SI D.2). Finally, experiments conducted in SI D.4 demonstrate that the dual Q-network configuration outperforms its single Q-network counterpart, highlighting the advantage of our proposed dual network approach.

6 Conclusion

Our computational model enables agents to interact selectively with their neighbours and protect cooperative behaviours from antisocial influences by MARL and iterative trial-and-error in simulations. Unlike existing RL cooperation studies, our approach does not rely on predefined social norms or external incentives [Ueshima *et al.*, 2023; Tennant *et al.*, 2023]. By integrating a spatial PDG into the training environment, we extend focus from paired interactions to those within a spatial structure, enabling a deeper analysis of how long-term learning impacts the coevolution dynamics of cooperation and selective interaction.

Our findings reveal that trained agents are capable of differentiating between neighbouring cooperators and defectors by observing information from their immediate surroundings, which enhances the network reciprocity and the associated group cooperation. Consistent with theoretical models [Szolnoki and Chen, 2020], our findings suggest that the performance of the learned interaction mechanism in promoting cooperation is attributed to its ability to help populations form homogeneous strategic clusters. These clusters are crucial for resisting invasions by defectors, especially in the early stages of development. Additionally, we confirm a positive correlation between agent memory length and the effectiveness of interaction selection, which in turn, aids the evolution of cooperative behaviors [Park *et al.*, 2022].

In conclusion, we emphasize the significance of understanding the learning dynamics in interaction selection and their contribution to fostering cooperation. These insights offer new insights into the emergence of cooperation within social dilemmas. Moreover, the computational framework developed here has broader implications, providing a versatile tool for investigating and examining mechanisms behind the evolution of cooperation in spatially structured environments. Integrating psychological complexity in our training framework emerges as a promising direction for future research. Understanding these mechanisms may provide solutions to social dilemmas and strengthen cooperation within both human societies and artificial intelligence systems.

Acknowledgments

The authors would like to thank anonymous reviews for their constructive comments. We also appreciate the insightful discussions with Xiaomei Mi and Yang Li related to this paper, and are grateful Haotong Zhang and Runqi Zhao for their support with computational resources. Finally, we acknowledge the assistance given by Research IT and the use of the Computational Shared Facility at The University of Manchester.

References

[Abeywickrama *et al.*, 2023] Dhaminda B Abeywickrama, Nathan Griffiths, Zhou Xu, and Alex Mouzakitis. Emergence of norms in interactions with complex rewards. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2280–2282, 2023.

[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 *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7047–7054, 2020.

[Danku *et al.*, 2019] Zsuzsa Danku, Matjaž Perc, and Attila Szolnoki. Knowing the past improves cooperation in the future. *Scientific reports*, 9(1):262, 2019.

[Du *et al.*, 2023] Yali Du, Joel Z Leibo, Usman Islam, Richard Willis, and Peter Sunehag. A review of cooperation in multi-agent learning. *arXiv preprint arXiv:2312.05162*, 2023.

[Jaques *et al.*, 2019] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Köbis *et al.*, 2019] Nils C Köbis, Bruno Verschuere, Yoella Bereby-Meyer, David Rand, and Shaul Shalvi. Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science*, 14(5):778–796, 2019.

[Köster *et al.*, 2022] Raphael Köster, Dylan Hadfield-Menell, Richard Everett, Laura Weidinger, Gillian K Hadfield, and Joel Z Leibo. Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents. *Proceedings of the National Academy of Sciences*, 119(3):e2106028118, 2022.

[Kramár *et al.*, 2022] János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications*, 13(1):7214, 2022.

[Leibo *et al.*, 2017] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[Li *et al.*, 2016] Jiaqi Li, Chunyan Zhang, Qinglin Sun, Zengqiang Chen, and Jianlei Zhang. Changing the intensity of interaction based on individual behavior in the iterated prisoner's dilemma game. *IEEE Transactions on Evolutionary Computation*, 21(4):506–517, 2016.

[Littman, 1994] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[Lowe *et al.*, 2017] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6382–6393, 2017.

[McKee *et al.*, 2023] Kevin R McKee, Andrea Tacchetti, Michiel A Bakker, Jan Balaguer, Lucy Campbell-Gillingham, Richard Everett, and Matthew Botvinick. Scaffolding cooperation in human groups with deep reinforcement learning. *Nature Human Behaviour*, 7(10):1787–1796, 2023.

[Metcalfe, 2017] Janet Metcalfe. Learning from errors. *Annual review of psychology*, 68:465–489, 2017.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Moffatt *et al.*, 2009] Peter Moffatt, Chris Starmer, Robert Sugden, Nicholas Bardsley, Robin Cubitt, and Graham Loomes. *Experimental economics: Rethinking the rules*. Princeton University Press, 2009.

[Nowak and May, 1993] Martin A Nowak and Robert M May. The spatial dilemmas of evolution. *International Journal of Bifurcation and Chaos*, 3(01):35–78, 1993.

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

[Park *et al.*, 2022] Peter S Park, Martin A Nowak, and Christian Hilbe. Cooperation in alternating interactions with memory constraints. *Nature Communications*, 13(1):737, 2022.

[Perc *et al.*, 2013] Matjaž Perc, Jesús Gómez-Gardenes, Attila Szolnoki, Luis M Floría, and Yamir Moreno. Evolutionary dynamics of group interactions on structured populations: a review. *Journal of the royal society interface*, 10(80):20120997, 2013.

[Perc *et al.*, 2017] Matjaž Perc, Jillian J Jordan, David G Rand, Zhen Wang, Stefano Boccaletti, and Attila Szolnoki.

Statistical physics of human cooperation. *Physics Reports*, 687:1–51, 2017.

[Perolat *et al.*, 2017] Julien Perolat, Joel Z Leibo, Vinicius Zambaldi, Charles Beattie, Karl Tuyls, and Thore Graepel. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3646–3655, 2017.

[Rand and Nowak, 2013] David G Rand and Martin A Nowak. Human cooperation. *Trends in cognitive sciences*, 17(8):413–425, 2013.

[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.

[Rapoport *et al.*, 1965] Anatol Rapoport, Albert M Chammah, and Carol J Orwant. *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan Press, 1965.

[Ren and Zeng, 2023] Tianyu Ren and Xiao-Jun Zeng. Reputation-based interaction promotes cooperation with reinforcement learning. *IEEE Transactions on Evolutionary Computation*, 2023.

[Ren and Zheng, 2021] Tianyu Ren and Junjun Zheng. Evolutionary dynamics in the spatial public goods game with tolerance-based expulsion and cooperation. *Chaos, Solitons & Fractals*, 151:111241, 2021.

[Santos *et al.*, 2006] Francisco C Santos, Jorge M Pacheco, and Tom Lenaerts. Cooperation prevails when individuals adjust their social ties. *PLoS computational biology*, 2(10):e140, 2006.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Sigmund *et al.*, 2010] Karl Sigmund, Hannelore De Silva, Arne Traulsen, and Christoph Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466(7308):861–863, 2010.

[Sigmund, 2010] Karl Sigmund. *The calculus of selfishness*. Princeton University Press, 2010.

[Su *et al.*, 2022] Qi Su, Benjamin Allen, and Joshua B Plotkin. Evolution of cooperation with asymmetric social interactions. *Proceedings of the National Academy of Sciences*, 119(1):e2113468118, 2022.

[Sylwester and Roberts, 2013] Karolina Sylwester and Gilbert Roberts. Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, 34(3):201–206, 2013.

[Szabó and Tőke, 1998] György Szabó and Csaba Tőke. Evolutionary prisoner's dilemma game on a square lattice. *Physical Review E*, 58:69–73, 1998.

[Szolnoki and Chen, 2020] Attila Szolnoki and Xiaojie Chen. Blocking defector invasion by focusing on the most successful partner. *Applied Mathematics and Computation*, 385:125430, 2020.

[Tanimoto, 2013] Jun Tanimoto. Difference of reciprocity effect in two coevolutionary models of presumed two-player and multiplayer games. *Physical Review E*, 87(6):062136, 2013.

[Tennant *et al.*, 2023] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Modeling moral choices in social dilemmas with multi-agent reinforcement learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 317–325, 2023.

[Traulsen *et al.*, 2010] Arne Traulsen, Dirk Semmann, Ralf D Sommerfeld, Hans-Jürgen Krambeck, and Manfred Milinski. Human strategy updating in evolutionary games. *Proceedings of the National Academy of Sciences*, 107(7):2962–2966, 2010.

[Ueshima *et al.*, 2023] Atsushi Ueshima, Shayegan Omidshafiei, and Hirokazu Shirado. Deconstructing cooperation and ostracism via multi-agent reinforcement learning. *arXiv preprint arXiv:2310.04623*, 2023.

[Vinitsky *et al.*, 2023] Eugene Vinitsky, Raphael Köster, John P Agapiou, Edgar A Duéñez-Guzmán, Alexander S Vezhnevets, and Joel Z Leibo. A learning agent that acquires social norms from public sanctions in decentralized multi-agent settings. *Collective Intelligence*, 2(2):26339137231162025, 2023.

[Wang *et al.*, 2013] Zhen Wang, Satoshi Kokubo, Jun Tanimoto, Eriko Fukuda, and Keizo Shigaki. Insight into the so-called spatial reciprocity. *Physical Review E*, 88(4):042145, 2013.

[Wang *et al.*, 2015] Zhen Wang, Satoshi Kokubo, Marko Jusup, and Jun Tanimoto. Universal scaling for the dilemma strength in evolutionary games. *Physics of Life Reviews*, 14:1–30, 2015.

[Willis *et al.*, 2023] Richard Willis, Yali Du, Joel Z Leibo, and Michael Luck. Resolving social dilemmas with minimal reward transfer. *arXiv preprint arXiv:2310.12928*, 2023.