MANCHESTER
1824
The University of Manchester

The University of Manchester Research

Reputation-based Interaction Promotes Cooperation with Reinforcement Learning

DOI:
10.1109/TEVC.2023.3304911

Document Version
Accepted author manuscript

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Ren, T., & Zeng, X. J. (2023). Reputation-based Interaction Promotes Cooperation with Reinforcement Learning. *IEEE Transactions on Evolutionary Computation*. https://doi.org/10.1109/TEVC.2023.3304911

Published in:
IEEE Transactions on Evolutionary Computation

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

OPEN ACCESS

Download date:12. Sep. 2023

Reputation-based Interaction Promotes Cooperation with Reinforcement Learning

Tianyu Ren, Xiao-Jun Zeng

Abstract—**Dynamical interaction represents a fundamental coevolutionary rule that addresses the intricacies of cooperation in social dilemmas. It provides a normative account for the changes in ties within interaction networks in response to the behaviour of social partners. While considerable efforts have explored the role of partner selection in fostering cooperation, there remains a limited understanding of how agents learn to establish effective interaction patterns and adapt their connections accordingly. To bridge this knowledge gap, we leverage recent advancements in reinforcement learning and propose an adaptive interaction mechanism to investigate self-organization behaviour in the iterated prisoner's dilemma game. Within this framework, artificial agents are trained using a self-regarding Roth-Erev algorithm, utilizing reputation as a dynamic signal to update their willingness to engage with neighbours. Additionally, these agents are endowed with the capability to sever inactive connections. Simulation results demonstrate the effectiveness of utilizing reinforcement learning and local information from reputation to capture the dynamics of interactions. Notably, we discover that the entangled coevolution of strategy and interaction network can facilitate the emergence and maintenance of cooperation, despite the optimal tolerance threshold for ineffective neighbours varying depending on the strength of the social dilemma. Furthermore, the emerging network topology presented in this work accurately captures the assortative mixing pattern observed in previous experiments and realistic evidence. Finally, we validate the simulation results through theoretical analysis and confirm the robustness of the proposed mechanism across populations of varying sizes and initial structures.**

I. Introduction

FOSTERING cooperation among self-interested agents is a challenging task, as natural selection favours free-riding on others' efforts. These myriad scenarios are characterized by so-called social dilemmas, where short-term individual incentives can conflict with long-term group interests, leading to collective irrationality [1]–[3]. Despite this, cooperation is ubiquitous in both natural and artificial systems [4], and it has played a vital role in the evolution of social species, chief among all in human social progress and civilization. Therefore, understanding the necessary conditions for cooperation has been an active topic, given the contradiction involved in such contexts [5]. The evolutionary game theory (EGT) provides a comprehensive theoretical framework to address social dilemmas. Among these, the prisoner's dilemma (PD) is widely recognized as one of the most challenging scenarios for pairwise cooperation and has received significant attention. In PD scenarios, the selection often leads to a reduction in the abundance of cooperators until the population is predominantly composed of defectors. Therefore, a mechanism to facilitate the evolution of cooperation is required, and many scholars have dedicated significant effort to this pursuit. The pioneering research of Nowak [6] identified five mechanisms that can cause cooperation to be favoured over defection.

One essential consideration that cannot be overlooked is that individuals typically engage in interactions within their immediate vicinity [7]. Network reciprocity, as an expanded framework that integrates EGT with complex network theory, enables studying cooperative behaviour in real-world systems. Research has demonstrated that the introduction of nontrivial populations can help individuals achieve higher payoff and resist exploitation from defectors by forming stable clusters [8]. However, the assumption that the underlying interaction network remains constant over time is often violated. In reality, interactions are typically formed from an ever-evolving amalgamation [9]. It is reasonable to assume that individuals often have control over their interactions and that the patterns of interaction change over time in response to the behaviour of their social partners. This nature should be considered to better understand the complexities of cooperation. Recently, coevolution dynamics have received increasing attention, involving not only the evolution of strategies but also the evolution of game environments [10]. Among that, one of the most effective coevolutionary strategies is the dynamical interaction rule, where unsatisfied players sever their links and seek more advantageous interactions with other partners. Breaking links can be viewed as expulsion, which has been demonstrated to be an effective means of promoting cooperation [11]. The positive effect of link reciprocity in facilitating the evolution of cooperation has been demonstrated through theoretical analyses and behavioural experiments [12]–[14]. Furthermore, the adaptability of networks provides an explanation for the coexistence of heterogeneous characteristics at various scales within realistic networks [15].

However, the aforementioned research findings, along with other profound insights, are just a few of many factors highlighting the significance of studying coevolutionary dynamics. One may question whether more intelligent agents rely solely on such a simplistic mechanism. In reality, individuals have the ability to gradually adjust their relationships with neighbours, referred to as interaction intensity, rather than abruptly changing them all at once. By combining adjustable interaction intensity with link rewiring, players can continuously adapt and optimize their potential interaction intensity, ultimately

The authors would like to acknowledge the use of the Computational Shared Facility at The University of Manchester. (Corresponding author: Xiao-Jun Zeng.)

T. Ren, X. Zeng are with the Department of Computer Science, The University of Manchester, Manchester M13 9PL, U.K. (email: tianyu.ren@manchester.ac.uk and x.zeng@manchester.ac.uk).

forming partnerships that maximize their fitness. Within such models, agents modify their partnerships by utilizing local information rather than merely copying or mimicking their neighbour's actions [16]. By observing past events, an intelligent agent can derive valuable insights to make informed judgments about the actions of its neighbours [17]. This information empowers the intelligent agent to distinguish between individuals, enabling it to respond in a more sophisticated and intricate manner, thus improving the overall adaptability of agents. In this context, indirect reciprocity, which relies on reputation and norms, is regarded as a powerful mechanism that aids individuals in distinguishing between good and bad actions. There have been several precursors who explored the fusion of network reciprocity combing with indirect reciprocity mechanisms. A comprehensive discussion regarding the reputation and reciprocity dynamic is expounded in a very recent review paper [18]. Notably, Hu et al. [19] exemplified this integration by incorporating adaptive reputation into a trust game based on network interactions, thus formalising fundamental trust mechanisms in social networks. Conversely, Tanimoto [20] reported that the effectiveness of combining network and indirect reciprocity mechanisms in fostering cooperation varies depending on the specific assessment system employed for evaluating actions. Furthermore, scholarly investigations have affirmed that reputation offers individuals sufficient information to adapt their interaction intensity [21].

One major weakness of EGT, however, is its proclivity to oversimplify real-life social phenomena by reducing them to abstract choices, neglecting the underlying structures and self-learning adaptive behaviour involved [22]. This limitation highlights the necessity of a more comprehensive and nuanced description of interaction behaviour, which can reveal new microfoundational mechanisms that shed light not only on the interaction relationship agents choose but also on how they adjust them. Modern methods based on Reinforcement Learning (RL) make this possible [23]. The learning dynamic of RL agents is aligned with real-life situations or empirical experiments, as they must learn to take future action through observations from the environment and the rewards received for those actions. The reimagining of artificial intelligence as deeply social can be considered a valuable tool for resolving fundamental challenges of cooperation [24]–[26]. Wang et al. [27] incorporated payoff noise and RL into a structured population and revealed a positive influence of Lévy noise promoting cooperation in PD games. Jia et al. [28] classified RL agents into two categories, namely global players and local players, based on the origin of the stimulus they encounter. Their findings highlighted the pivotal role of global players in driving cooperation forward. Furthermore, RL has been recently applied to gain a better understanding of the evolution of adaptive interaction in social dilemmas [29]–[31]. For instance, Anastassacos et al. [32] trained RL agents in decentralized multi-agent scenarios and found that encouraging norm-inducing behaviours and adopting a bottom-up approach to partner selection can effectively promote agent cooperation. Additionally, recent literature introduced the Bush-Mostelle algorithm to explore the evolutionary process of adaptive interaction intensity, demonstrating how RL can contribute to the self-organization of social fabric [33].

We would like to point out that, based on our understanding, the existing studies on the evolution of cooperation through RL primarily focus on static networks. Although prior studies on coevolutionary games have explored the dynamics of interaction networks, there has been limited attention given to understanding the rewiring characteristics that arise in such scenarios. Building upon these insights, we propose a novel coevolutionary rule that incorporates reputation as observable information, affecting the interaction behaviour of RL agents. The purpose of this paper is to investigate the emergence of cooperation in a decentralized society and to deepen our understanding of coevolutionary dynamics through the introduction of RL with individual learning properties. Remarkably, our observations reveal that RL agents exhibit two distinct interaction patterns, which subsequently shape the characteristics of the emergent network as either assortative or disassortative. Moreover, we find the optimal tolerance threshold and social norm, which govern the evolution of cooperation, is contingent upon the dilemma strength.

The contribution of this work makes two-fold. Firstly, it proposed an insightful approach that utilizes RL to capture the dynamic nature of interaction intensity, enabling a comprehensive examination of the interplay between individual learning and social learning in a decentralized multi-agent system. The integration of RL and EGT demonstrates superior performance in promoting cooperation, thereby highlighting the effectiveness of combining these two approaches. Secondly, we explain why an assortative mixing pattern emerges in self-organizing populations with a reputation-based interaction. It offers a fresh understanding of the underlying adaptive interaction and partner selection driving cooperative behaviour in multi-agent systems. We believe these findings hold implications not only for comprehending the emergence of cooperation in human societies but also for enhancing cooperation in artificial intelligence systems.

The rest of this paper is organized as follows. We begin in Section II by introducing some formal definitions, including PD game, adjustable interaction intensity and Roth-Erev algorithm. Following that, in Section III, we describe our model in detail. Section IV presents our simulation results followed by a discussion. Finally, conclusions are drawn in Section V.

II. Background Knowledge

A. *Prisoner's Dilemma Game*

The story of the PD was first introduced by mathematician Merrill Flood and Melvin Dresher in 1950 and later formalized by Albert Tucker. In its classical form, the prisoner dilemma describes a situation that two burglars are arrested and held separately by the police. The prosecutor offers each prisoner the same deal: if one confesses and the other remains silent, the silent accomplice receives a three-year sentence while the confessor goes free. However, both would receive a less severe sentence if they both confess. This paradox has been increasing attention since it arose and is considered one of the most challenging cooperative dilemmas [34].

Game theory provided a fundamental framework for investigating the social conflict between cooperative and selfish behaviour. The PD game can be described as a scenario where two players, namely the Donor and the Recipient, face a decision-making situation. In this paradigmatic model, the players are confronted with the choice of either cooperating (remaining silent, denoted by C) and incurring a cost c to assist their counterpart, who would then receive a temptation benefit b $(b > c > 0)$; or defecting (confess, denote by D), which entails receiving the benefit without providing any benefit in return. Consider a game between two strategies, C and D, the standard payoff matrix between cooperators and defectors is given by

$$M = \begin{array}{cc} & \begin{matrix} C & D \end{matrix} \\ \begin{matrix} C \\ D \end{matrix} & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{array} \tag{1}$$

where mutual cooperation and mutual defection lead to the reward $R = b - c$ and punishment $P = 0$, respectively. While unilateral cooperation yields a cost $S = c$ when confronting a defector, who gets the highest payoff $T = b$. Within the PD setting, these payoffs satisfy the conditions: $T > R > P > S$ and $2R > T+S$. Since D strictly dominates C for both players, defection is the only Nash Equilibrium, despite all individuals would be better if they all choose cooperation. It is also worth noting that the general version of the PD game is a one-off, many of the situations are alleged to have an iterated structure [35]. Thus, the one-off game can be extended into a repeated PD form, where each player i can follow one of two strategies, denoted by the two-dimensional unit vector, $s_i(t) = [1, 0]^T$ corresponds to choose cooperation strategy, and $s_i(t) = [0, 1]^T$ for defection, respectively, at each time step t.

B. Adjustable Interaction Intensity

Traditional PD games are based on a prior assumption that players interact with their paired partners in a deterministic relationship. However, the interactions between participants are not always at the same level. Consider two paired individuals i and j, playing an iterated prisoner's game (IPD) in several time series. At each time step t, they are eligible to adjust their willingness to interact with each other dynamically. Referring to Ref. [21], we denote the willingness that individual i $[j]$ to interact with j $[i]$ at time t by $w_{i\to j}(t)$ $[w_{j\to i}(t)]$, where $w_{i\to j}(t) \in [0, 1]$. Therefore, according to probability theory, the interaction intensity between agents i and j, i.e., the probability that these two agents successfully interact with each other, can be denoted as follows:

$$W_{i,j}(t) = w_{i\to j}(t) \times w_{j\to i}(t). \tag{2}$$

The interaction intensity $W_{i,j}$ depends on the interaction willingness from both sides. If $W_{i,j} > 0$, player j becomes the effective neighbor of player i (vice versa). Initially, all individuals have a strong interaction willing, that is $w(0) = 1$. After each time step, players independently revise and adapt their willingness to interact with paired counterparts. The adjustment for individual i and j is measured by $\triangle w_{i\to j}$ and $\triangle w_{j\to i}$, respectively, where $\triangle w \in [-1, 1]$. Consequently, the interaction intensity is updated in the subsequent time step based on the revised interaction willingness between individuals i and j, which is defined as follows:

$$\begin{aligned} W_{i,j}(t+1) &= w_{i\to j}(t+1) \times w_{j\to i}(t+1) \\ &= [w_{i\to j}(t) + \triangle w_{i\to j}] \times [w_{j\to i} + \triangle w_{j\to i}]. \end{aligned} \tag{3}$$

C. Roth-Erev Algorithm

Compared with the Q-learning environment, the adjustable interaction intensity task requires a dynamically varying environment with a more complex set of actions. For this purpose, the Roth-Erev (RE) algorithm, a classical reinforcement learning algorithm proposed by Roth and Erev [36], proves to be more suitable [37]. The RE algorithm is specifically designed based on observations of human behaviour in iterated game play involving multiple strategically interacting players across various game contexts. The key idea of the RE algorithm is: The probability of choosing an action is proportional to the total accumulated rewards from choosing it in the past. The schematic model of the RE algorithm includes three basic elements for agent i: action choice (a_i), choice propensity vector (q_i) and normalized choice probability (p_i). These game-theoretic models consist of a probabilistic decision rule and a learning algorithm in which game payoffs are evaluated relative to an aspiration level, and the corresponding choice propensities are updated accordingly. Assuming participant i select an action $(A \in (C, D))$ can receive a payoff x from the game setting, the reinforcement function of agent i is described as follows:

$$R(x) = x - x_{min} \tag{4}$$

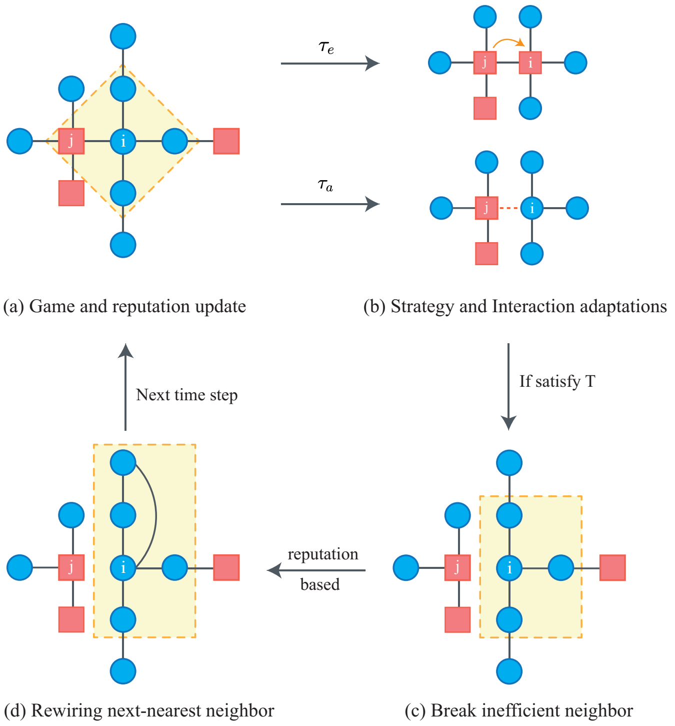
where x_{min} is the smallest possible payoff for the agent during one round game. Afterward, the agent i updates his propensity according to the equations:

$$\begin{cases} q_{i,A}(t+1) = (1-\xi)[q_{i,A}(t) + E_t[A, R(x)]] \\ q_{i,\neg A}(t+1) = (1-\xi)q_{i,\neg A}(t) \end{cases} \tag{5}$$

where $q_{i,A}$ is the propensity for agent i's action A, and $q_{i,A}(0)$ denotes the initial propensity. The initial propensity level act as an 'aspiration level' and can be classified into two distinct levels based on their characteristics. For the high initial propensities, the effect of the payoff is diminished, thereby allowing for slower learning and more experimentation. Conversely, low initial propensities encourage premature fixation on one strategy. The forgetting (or recency) parameter ξ reduces the influence of past experiences. When $\xi \to 0$, equal weight is given to all rewards received to date by the agent. And $E(A, R)$ is a function which determines how the experience of adopting action A and receiving reward $R(x)$ is generalized to update each strategy.

III. MODEL

In the former section, we detailed the IPD among the well-mixed population. However, in realistic multi-agent systems, players do not interact with all other participants but only with their immediate neighbours. We now consider a spatially structured population in which individuals are confined on L^2



(a) Game and reputation update

(b) Strategy and Interaction adaptations

(d) Rewiring next-nearest neighbor

(c) Break inefficient neighbor

Fig. 1. **The coevolutionary dynamic of strategy and topology through reputation-based interaction.** The yellow area represents the effective interaction range of player i. (a) Game interaction and reputation updating stages. Individuals engage in several PD games with their effective neighbours and update their future reputations according to social norms. (b) Strategy and interaction adaptations stage. Entangled coevolution of individual strategy and interaction willingness depends on the ratio $\Phi = \tau_e/\tau_a$. (c) Link dismisses stage. Individuals break the connection with one ineffective neighbour who has not been interacted with for T rounds. (d) Neighbour rewiring stage. Individuals link with the next-nearest neighbour based on reputation preference.

square lattices with periodic boundary conditions and interact only with their neighbours in the set Ω. Initially, we employ square lattices where each player has four neighbours. The nodes represent players, and the edges refer to the connection between the corresponding players. Yet, we assume that the network continues to evolve after the initial setting [15]. Each agent not only engages in the strategy-switching process but also the topological evolution process. Consequently, the coevolutionary dynamic of network heterogeneity and the strategy complexity become the crucial factors affecting the evolution of cooperation.

A. Framework of model

For each time step, the entangled coevolutionary dynamics combined with strategy and topological evolution are illustrated and described in Fig. 1. Accordingly, the evolutionary process can be divided into the following stages.

Stage 1 (Game Interaction): During the game interaction process, each individual plays several PD games. The fitness of a certain player corresponds to the total accumulated payoff resulting from all effective pairwise interactions (Fig. 1a). For instance, individual i obtain an overall income at time step t as [38],

$$\Pi_i(t) = \sum_{j\in N_i(t)} s_i^T M s_j \tag{6}$$

where $N_i(t)$ denotes the number of effective neighbor of agent i at time step t. M is the payoff matrix for the PD game. For the sake of simplicity, we accept the idea proposed by Nowak and May [39] and adopt the weak PD by setting $R = 1, P = S = 0, T = b$ $(b \geq 1)$.

Stage 2 (Reputation Updating): In accordance with the convention [40], we employ a simple model based on binary scores, taking only the value ”good” and ”bad” to assess the reputation of individuals. For the sake of clarity, we refer to a score of 1 as good and 0 as bad. At the outset, the good and bad labels are devoid of any prior meaning. The assessment of binary reputation depends on the actions and reputation between the player and the co-player. For instance, the action of player i to j is captured and evaluated by the bystander, who attributes a new reputation to the player according to her/his action, as well as the reputation of co-player. Players can communicate what has transpired or their assessment to others, meaning reputation becomes common knowledge throughout the system. This can be viewed as a rudimentary system of moral assessment. The significance of the reputation label emerges in conjunction with individual behaviour. To perform this task, the bystander uses a second-order social norm that encodes and translates the information about the player's action and the co-players reputation involved in an interaction into the future reputation for the player. It is worthwhile to notice that the social norm can be extended to the third order by including the current reputation of the central player or considering past reputation [41]. During the main experiments, we employ an efficient second-order social norm, called stern-judging (SJ), to generate the reputation dynamic, which assigns a good reputation to agents with prosocial behaviour and a bad reputation to agents performing antisocial behaviour [42]. Accordingly, the reputation of individual i at time step t can be determined using the following utility function:

$$R_i(t) = (1-\theta)R_i(t-1) + \theta \sum_{j\in N_i(t)} \frac{d[s_i(t), R_j(t-1)]}{k_i} \tag{7}$$

where θ is the evaluation factor regarding to individual behaviour, and k_i is the actual degree of i. Here the $d(s, R)$ function determined the updated reputation according to the SJ norm, which can be described as:

$$d(s_i, R_j) = \begin{cases} 1, & s_i = C\ , R_j > 0 \\ 1, & s_i = D\ , R_j \leq 0 \\ -1, & s_i = D\ , R_j > 0 \\ -1, & s_i = C\ , R_j \leq 0 \end{cases}. \tag{8}$$

Stage 3 (Coevolution of Strategy and Interaction): The coevolutionary game model proposed in this study integrates individual strategy and interaction intensity during pairwise interactions. By combining link reciprocity and indirect reciprocity, the model entangles strategy and interaction, allowing for updates in both variables. The update dynamics employed in this study were derived from the method suggested by Perc and Szolnoki[10], which does not rely on predetermined discriminative rules for cooperation and defection strategies.

We first introduce two different time scales: strategy update process (τ_s) and interaction intensity revise process (τ_i). Depending on the ratio between them, $\Phi = \tau_s/\tau_i$ determines the coevolutionary dynamics of strategy and interaction intensity, which can be updated asynchronously (Fig. 1b). The probability of selecting a strategy update process is $(1+\Phi)^{-1}$, and an interaction intensity revision process is chosen otherwise. (The specific approaches to implement the above coevolution process, see Refs.[7], [43], [44]). The parameter Φ reflects the inertia of individuals to react to rational choices, and its value can influence the fate of the evolution. Specifically, as $\Phi \to 0$, the evolution process approaches the static graph cases [45], whereas with increasing Φ, individuals become more efficient in adapting their interaction intensity toward neighbours.

Stage 4 (Linking Updating): To achieve the evolution of network structure, we introduce an additional mechanism for updating links. At the end of each time step, a rewiring round takes place, during which players decide whether to alter their network connections with existing neighbours. In this study, we aim to implement a reputation-based rewiring model based on an extension of a previous empirical experiment [12]. For each edge with paired individuals, i break the link with j if this neighbour has been inefficient for the last T rounds (Fig. 1c), maintaining the edge otherwise. If i dismisses the link, i switches from this inefficient neighbour either to another player chosen among i's next-nearest neighbour preferentially according to their reputation or to a random member of the entire population (Fig. 1d. The specific approaches for the rewiring process can be found at Ref. [46]). Without loss of generality, a randomly rewiring action happens with probability $1-\epsilon$; otherwise, with probability ϵ, the player forms a new connection with the highest reputation among all neighbours' neighbours. The intuition behind this reasoning is as follows: rational individuals tend to maximize their game payoff with limited information and are more likely to choose a good-reputation partner who can potentially enhance their future payoff. In this model, individuals cannot reject the formation of new connections, and there are no costs for disconnection and rewiring links. However, we do enforce a crucial condition: nodes in the graph must remain connected at all times. To achieve this condition, we stipulate that individuals with only one edge cannot disconnect or be disconnected.

B. Method of Strategy Updating

In the evolutionary game, all individuals experience the strategy updating phase synchronously. Specifically, an individual i revises its strategy s_i by pairwise comparison with randomly chosen neighbour j, who can choose to pass their strategy with a probability of [47]

$$f(s_i \leftarrow s_j) = (\Pi_{s_j} - \Pi_{s_i})/(\triangle \cdot k_q) \tag{9}$$

where k_q represents the larger of the two degrees k_i and k_j, and $\triangle$ is defined as the difference between temptation to defection T and the sucker's payoff S. Since we adopt a weak PD game in this model, we have $\triangle = b$.

C. Method of Adjusting Interaction Intensity

As discussed in Section II-B, the interaction intensity between pairwise players is determined by their mutual interaction willingness. To obtain adaptive dynamics of interaction intensity along with the proposed reputation mechanism, we employ the RE algorithm to train a population of independent RL agents. These agents have different propensities regarding two optional actions toward their neighbours: interact (I, denoted by 0) or separate (S, denoted by 1). For instance, the propensity vector of player i toward its neighbor j at a given time step t is denoted as: $q_{i,j}(t) = [q_{I_{i,j}}(t), q_{S_{i,j}}(t)]^T$. Initially, at $t=0$, every player has an equal propensity to actions I and S, denoted by $q_{A_{i,j}}(0) = [0.5, 0.5]^T$. An individual i engages in several games with its effective neighbours and collects their current reputation. Assuming that individual i chooses action A towards j during the learning process, then i updates its propensity vector for j using this information according to

$$\begin{cases} q_{A_{i,j}}(t+1) = (1-\xi)q_{A_{i,j}}(t) + \mathrm{sgn}(A)\tanh(\beta[R_j(t) - R_{i,\Omega}(t)]) \\ q_{\neg A_{i,j}}(t+1) = (1-\xi)q_{\neg A_{i,j}}(t) \end{cases} \tag{10}$$

where $R_{i,\Omega}(t) = \sum_{j\in\Omega_i(t)} R_j/k_i$ is the aspiration level of reputation average over all its neighbours, β represents the degree of reputation sensitivity, which also reflects the learning intensity of the RL agent. $\mathrm{sgn}(x)$ is a sign function: $\mathrm{sgn}(x) = 1$ if $x = 0$, otherwise $\mathrm{sgn}(x) = -1$. Subsequently, the interaction willingness of individual i concerning j at time t can be derived from the corresponding propensity vector $q_{i,j}(t)$. Since the reputation could be negative or zero, we use the *Softmax* function to normalize the above propensity vector. Therefore, the willingness of individual i to interact with j in the next time step is

$$w_{i\to j}(t+1) = w_{i\to j}(t) + \triangle w_{i\to j} \tag{11}$$

where

$$\triangle w_{i\to j} = \mathrm{sgn}(A)\left[\frac{\exp(q_{A_{i,j}}(t+1))}{\sum_{A\in\{I,S\}}\exp(q_{A_{i,j}}(t+1))} - \frac{1}{2}\right].$$

See Section I of the supplementary material for detailed information regarding the RL step diagram and training algorithm, involving the process of updating the interaction intensity.

IV. Results

In this section, we will show the detailed results of the evolution of cooperation induced by the proposed reputation-based interaction mechanism. We started the simulation from a homogeneous partner network [48], where the linear size of the entire lattice L was chosen between 100 and 400, and the population size was denoted as $N = L \times L$. Initially, cooperators and defectors are randomly distributed in the population with the same probability. To facilitate the following discussion, we used the fraction of cooperators f_c in the population to characterize the general cooperative level.

We maintain the parameters as $\xi = 0.01, \epsilon = 0.9, \theta = 0.5$ and $\Phi = 0.5$ throughout this work unless explicitly stated otherwise. To ensure validity and minimize variability, the results are computed by averaging over the last 10^3 generations across the entire simulation, which is encompassed a range of time spanning from 10^4 to 10^6. Moreover, the final stable states are obtained by conducting up to 10 independent runs for each set of parameters to eliminate deviations.

To capture the evolving nature of the networks as individuals alter their connections, we employ two metrics to assess network dynamics. First, the degree of heterogeneity in the networks is as follows

$$h = \frac{\sum_k k^2 N(k) - \langle k \rangle}{N} \tag{12}$$

where $N(k)$ determine the number of vertices with k degree, and $\langle k \rangle = \sum_{i=1}^{N} k_i / N$ represents the average degree. Furthermore, the degree-degree pattern of the emerging network is investigated using the Pearson associativity coefficient r [49], which quantifies the correlation between the degrees of adjacent nodes. Here, we use p_k to denote the degree distribution of the graph as a whole, which is the probability that a randomly chosen vertex will have degree k. Accordingly, the excess degree of the vertex at the end of an edge is distributed according to

$$\mu_k = \frac{(k+1)p_{k+1}}{\sum_j j p_j}. \tag{13}$$

Then the assortativity coefficient for mixing by vertex degree is

$$r = \frac{\sum_{jk} jk(e_{jk} - \mu_j \mu_k)}{\sigma_\mu^2} \tag{14}$$

where e_{jk} refers to the joint probability distribution of the remaining degrees of the two vertices, and σ_μ is the standard deviation of the distribution μ_k. This equation allows us to evaluate the mixing pattern for a given network, where $r > 0$ indicates that nodes with similar degrees tend to be connected to each other. In contrast, $r < 0$ indicates that nodes tend to connect to other nodes with dissimilar degrees. $r = 0$ means the network is non-assortative.

A. Effectiveness of RL in promoting cooperation

We begin by examining the overall performance of RL agents, who can dynamically adjust their interaction willingness and neighbours based on reputation information. Consider that temptation to defection b and reputation sensitivity β are two key parameters for the weak PD and RE algorithm, respectively, the contour plots in Fig. 2 demonstrate how cooperator survive as a function of b and β for different tolerance threshold values. Specifically, when $\beta = 0$, indicating that the interaction willingness of players is not influenced by the reputation of their neighbours. In this case, our results follow the conventional expectations of the standard iterated PD game, as reported in [39].

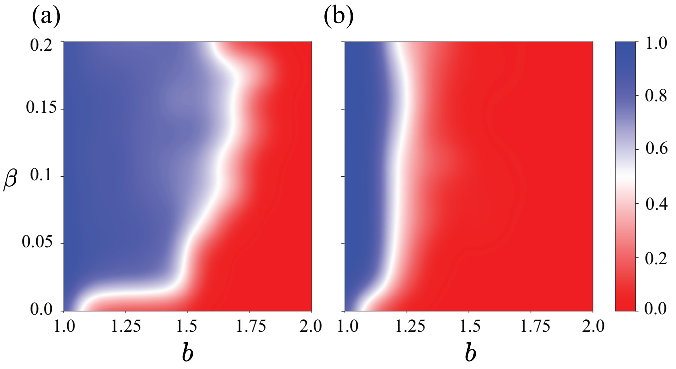


Fig. 2. **Equilibrium fraction of cooperators in dependence on the temptation and reputation sensitivity for different tolerance threshold values.** Incorporating adaptive interaction intensity and partner selection has effectively enhanced cooperation. There are three distinct phases: full cooperation, mixed strategy, and full defection phase. (a) Reputation-based coevolutionary dynamics with ineffective threshold $T = 10$. (b) Adjustable interaction intensity without evolving population ($T = \infty$). From red to blue, the colour bar indicates that the cooperation level changes from 0 to 1 accordingly.

As β becomes greater, the "wave of cooperation" gradually spreads to the east in each contour plot, demonstrating that RL cooperators can outcompete defectors by adjusting the interaction willingness without considering structural evolution. This result is consistent with previous research in EGT [21]. However, the fraction of cooperators is greatly enhanced if agents have the ability to rewire links. As shown in Fig. 2a, the positive effect of increased reputation sensitivity on the evolution of cooperation is more evident under coevolutionary dynamics. Additional insights provided in Fig. 5 also demonstrate the critical role of structural evolution, where the entangled coevolutionary dynamics promote the evolution of cooperation. Similar conclusions have been reported in previous research [10]. An intriguing aspect that deserves attention within coevolutionary dynamics is that even if cooperators make up the majority, they are usually unable to completely eliminate defectors (represented by the light and white area). However, we can observe in Fig. 2b a sharper transition from the phase of full cooperation, represented by the blue area, to the phase of full defection, represented by the red area. Altogether, individuals utilizing reinforcement learning can achieve greater fitness by adjusting interaction intensity with neighbours, and the coevolution of population structure leads to the strengthening of the cooperation level.

B. Coevolution of strategy and structure

1) Dynamic change of individual behaviour: We first selected two typical temptation values to study the evolution of the strategy resulting in the coexistence (see Fig. 3a) and full defection phases (see Fig. 3d). It is worth noticing that under such a reputation-based adjustable interaction model, the learning process of RL agents can be characterized by two evident EGT processes: enduring (END) period and expanding (EXP) period [50]. As depicted in the graphs, cooperation experienced a sharp decline across all cases during the former period. In the bottom graph, the subsequent period illustrates instances where the defectors prevailed in the preceding period and absorbed, while the upper graph showed an increase in the proportion of cooperation. Considering we select a critical

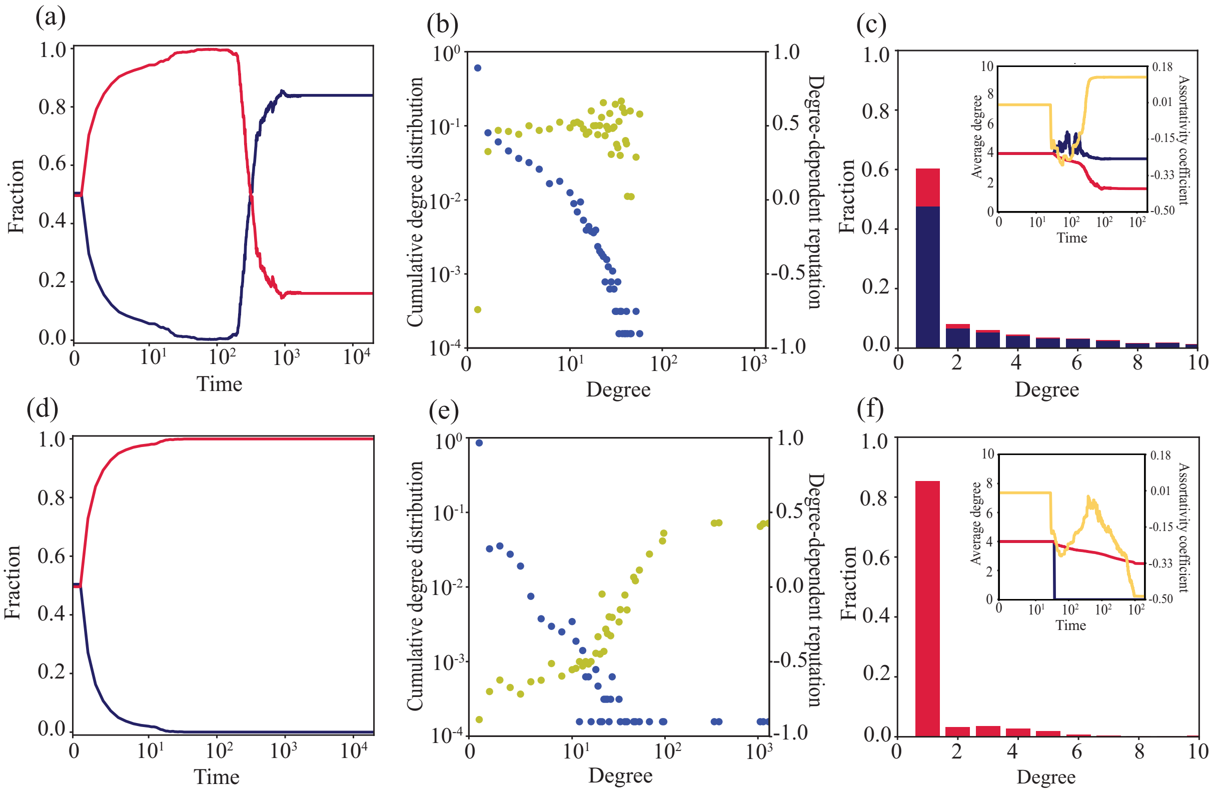


Fig. 3. **Coevolutionary dynamics of strategy and network structure.** The utilization of reputation-based interaction fosters heterogeneity. Evolving networks exhibit assortativity in weak dilemma-strength scenarios but disassortativity in strong dilemma-strength scenarios. The temptation values are set to $b = 1.45$ in the upper row and $b = 1.80$ in the bottom row, respectively, with $\beta = 0.1$ and $T = 10$. Each column presents the fraction of cooperators and defectors, indicated by the blue and red lines; the cumulative degree distribution of the evolving network and the degree-dependent reputation score at the final state, depicted by blue and yellow dots; the distribution of cooperators and defectors with degree $k \leq 10$, represented by blue and red bars. The subgraph in (c) and (f) displays the average degree of each strategy and the evolution of the associativity coefficient (yellow line) over time.

value as $b = 1.45$, cooperators take a considerable amount of time to endure the invasion of defection effectively. As a consequence, during the END period, defectors almost dominate the population, with cooperators struggling to gain a foothold. However, in the EXP period, cooperators become more evolutionary competitive and expand by dynamically adjusting their interaction willingness and neighbours. Notably, the reputation mechanism provides positive feedback to promote cooperation. As shown in Figs. 3b and 3c, the degree of individuals is positively correlated with their reputation level. This finding is consistent with previous reputation research on reputation [46], which suggests that individuals with higher degrees tend to have correspondingly higher reputation levels. Moreover, the highest degree nodes in the network are predominantly occupied by cooperators, in accord with empirical evidence [12]. Indeed, the survival and expansion of cooperation are largely attributed to the heterogeneity of the network (see Refs [14], [51], [52]), whereby the remaining defectors among the population can only survive on low-degree nodes. Therefore, we confirm that the well-established result that a heterogeneity degree distribution enhances cooperation in static networks also holds true for evolving networks [47].

2) Evolution of network structure: To provide a more precise quantification of the entangled dynamics between strategies and networks in promoting cooperation, we examined several network characteristics. The degree distribution had a sharp cutoff around $k = 10$, with two distinct types of heterogeneity: single-scale heterogeneity (Fig. 3b) and broad-scale heterogeneity (Fig. 3e). Furthermore, the associativity of the evolving network is tuned by the temptation parameter b, as illustrated in Figs. 3c and 3f. During the END period, the resulting networks generally displayed assortative mixing, although disassortativity was occasionally observed. In the following EXP period, the network continues to assortativity mix for $b = 1.45$, while becoming extremely heterogeneous for $b = 1.8$, with network heterogeneity degrees of $h = 34.3$ and $h = 972.53$, respectively. Similar phenomena have been reported in previous coevolutionary models [53]. The intuition behind this phenomenon is as follows: in order to avoid exploitation by defectors, the surviving cooperators tend to form clusters, which typically maintain good reputations according to social norms. As a result, individuals from outside the cluster who have severed ties with their defective neighbours prefer to join these alliances. However, the initial clusters can be too fragile to withstand invasion due to the high temptation value, and defectors occupy the hub node. It is important to note that the hub defectors may also have a good reputation (Fig. 3e) as long as they adopt defective behaviour toward other bad-reputation neighbours. In such cases, the reputation cannot maintain a consistent level among the entire population.

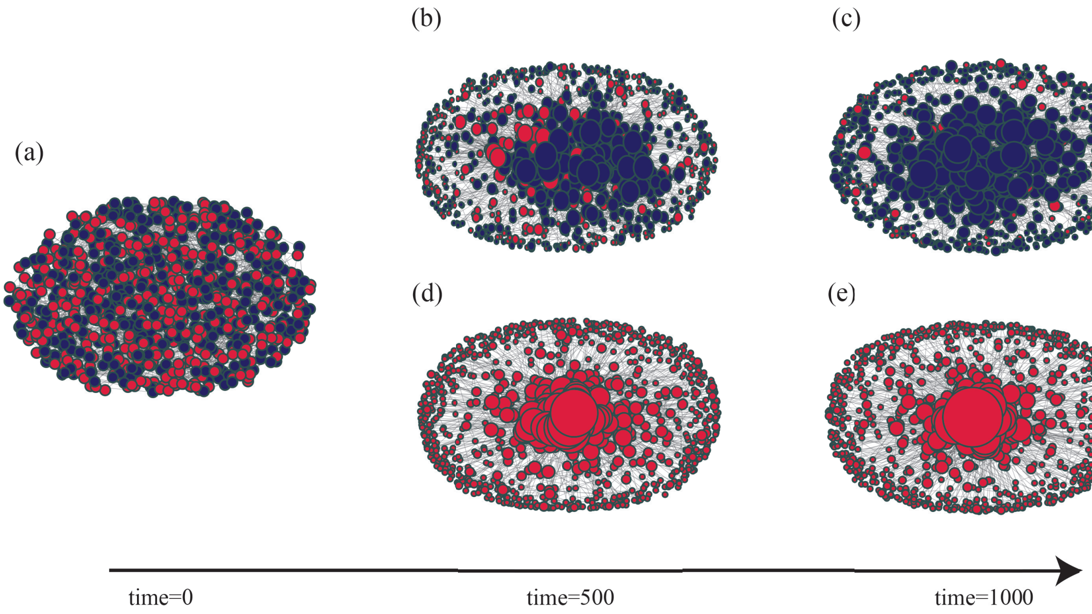


Fig. 4. **Snapshots of structure and strategy for a small subset $N = 900$ individuals from the entire population.** In the presence of reputation-based interaction, network heterogeneity is further exacerbated when the temptation value is high. Snapshots are taken at time steps 0, 500, and 1000, and are displayed from left to right. Panel (a) illustrates the initial state for all cases. Panels (b) and (c) depict the evolution of the network with a temptation value of $b = 1.45$, while panels (d) and (e) show the network evolution with a temptation value of $b = 1.80$. Blue and red nodes represent cooperators and defectors, respectively, with node size indicating the number of connections. The parameters are fixed at $T = 10$ and $\beta = 0.1$.

Consequently, frequent changes in interaction willingness and link occur, further exacerbating network heterogeneity.

To provide an intuitive understanding of the interplay between strategy and structure, a series of evolving network snapshots are presented in Figure 4. The network layout is achieved using a Force-directed algorithm, where edges are modelled as springs between nodes, and nodes with high degrees are positioned closer together. The snapshots support the earlier conclusions and illustrate that network updates promote assortative interactions between good-reputation individuals and disassortative interactions between good and bad-reputation individuals. As the evolutionary trail depicted in Figs. 4b and 4c, the evolving cooperation causes assortative topology [53]. However, disassortative interactions are favoured in the full-defection network, leading to a gradual increase in network heterogeneity over time. Altogether, these results indicate that adaptive interactions between RL agents increase network heterogeneity, which is instrumental for the evolution of cooperation if cooperative clusters are capable of forming stable alliances and reaching a consistent reputation level.

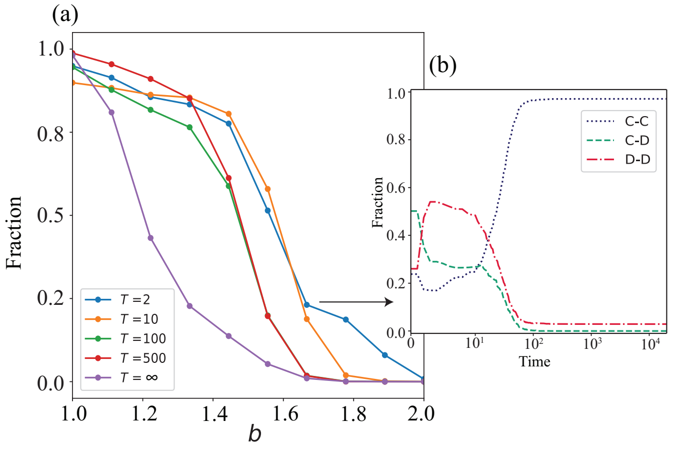


Fig. 5. **The influence of tolerance thresholds in relation to different temptation values.** Cooperation is favoured when individuals are able to adapt their interactions based on the reputation of their partner. As the temptation value increases, individuals are compelled to switch partners more frequently. The curve for $T = \infty$ represents dynamical interactions without link adaptations. The right-hand subgraph illustrates the fraction of CC/CD/DD links over time for $b = 1$ and $T = 2$. The reputation-sensitive parameter is fixed at $\beta = 0.1$.

C. Relation of tolerance threshold

As illustrated in Fig. 2, when the dilemma strength is strong, the coevolutionary model that relies solely on adjustable interaction intensity depicts inferior cooperation compared to the model that incorporates both strategy and topological structure. This highlights the crucial role of severing adverse connections in maintaining cooperation. In order to assess the significance of rewiring behaviour, we evaluate the effect of tolerance threshold in the partner-switching process, as shown in Fig. 5. A comparison between numerical simulation and theoretical analysis is given in Section II of the supplementary material. Clearly, if no link adaptations are involved ($T = \infty$), the cooperation level is much lower. The introduction of an evolving network mechanism greatly enhances cooperation; however, the optimal threshold condition may vary depending on the temptation to defect. As demonstrated, if $b > 1.6$, increased sensitivity to ineffective neighbours (at $T = 2$) can enhance the survival of cooperation. On the flip side,

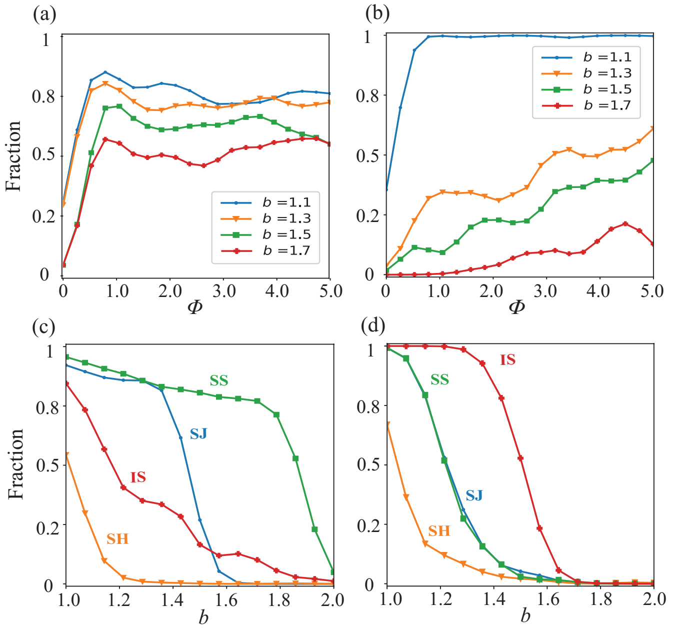


Fig. 6. **The effect of time scale separation and social norm on cooperation levels under varying dilemma strength.** Cooperation is facilitated when individuals become more efficient in adapting their interaction intensity. The less benevolent social norm demonstrates better performance under strong dilemma conditions. The upper panels display the evolution of cooperation with respect to Φ in two scenarios: $T = 10$ [panel (a)] and $T = \infty$ [panel (b)]. The lower panels (c) and (d) illustrate the performance of four social norms when Φ is fixed at 0.5 and T is set to 10 and ∞, respectively. The reputation-sensitive parameter remains fixed at $\beta = 0.1$

when faced with a low temptation to defect, a population is more likely to achieve the highest level of cooperation if they rarely switch neighbours, as in the case where $T = 500$. Noteworthy, in the no-dilemma case depicted in the leftmost area, frequent rewiring behaviour inhibits the emergence of a full-cooperation phase. The interplay between the tolerance threshold and the temptation to defect can be understood as follows: for the fixed parameter setting of $T = 2$ and $b = 1$, network adaptability enables cooperators to interact with their cooperative neighbours in an assortative mixing pattern, leading to a rapid decrease in the fraction of defectors to a significantly low level. However, complete elimination of defectors is hard to achieve, and the link-strategy configuration plotted in Fig. 5b confirms that the number of C-D links is limited once the system reaches a stable state. As a result, the failure of high-connected cooperative clusters to convert defectors to cooperators leads to the coexistence equilibrium of cooperation and defection. However, the lower tolerance threshold is required in addition to strategy updating to promote cooperation as b increases. Ultimately, the present results convey a simple yet powerful message for the evolving network: as the temptation to defect increases, individuals need to adjust the connections of ineffective neighbours more frequently.

D. The role of time scale and social norm

During the aforementioned experiments, we maintained a fixed $\Phi = 0.5$ ratio for the strategy update process and the interaction intensity revision process. Previous studies have demonstrated that coevolutionary dynamics can be significantly affected by the time scale separation [15], [46]. Therefore, we investigate the effect of the update time scale between strategy and interaction intensity on the evolution of cooperation, as shown in the upper panels of Fig. 6. As the time scale Φ increase, a corresponding increase in the cooperation level within the population. However, the impact of time scale separation appears to be highly dependent on the temptation to defect b. Specifically, as b increases, the optimal cooperation level decreases, and the interaction intensity updating needs to occur more frequently to ensure optimal conditions are met. Furthermore, the temptation of defection plays a decisive role in the evolution of cooperation, as achieving full cooperation becomes increasingly challenging even for sufficiently large values of Φ. It is noteworthy that, although introducing link rewiring ($T = 10$) results in a higher fraction of cooperation with only a small Φ, the conclusion drawn from Fig. 5 is still applicable. Frequently rewiring ineffective neighbours may not necessarily favour cooperation in weak dilemma conditions. As shown in Fig. 6b, for $b = 1.1$, frequently adjusting interaction intensity without modifying the network structure can encourage higher cooperation.

Thus far, we have provided aggregate information on the promotion effect of adjustable interaction intensity and partner selection, yet we did not pay sufficient attention to the generation of reputation. Specifically, the reputation of each participant is attributed solely according to the SJ norm, while other social norms could be considered. In the lower panels of Fig. 6, we present a comparison of the SJ norm with three other effective norms, as described in Ref. [54]. The first norm, Simple-Standing (SS), is similar to SJ but assigns a good reputation to any players who cooperate, making it more benevolent. The second norm, Shunning (SH), is similar to SJ but assigns a bad reputation to any player who exhibits defective behaviour, making it less benevolent. The third norm, Image Score (IS), is based solely on a player's actual behaviour and is independent of their opponent's behaviour. Based on the information presented in Fig. 6c, it is clear that despite SJ leading to a cooperation level systematically lower than SS in strong dilemma situations, these two norms are still the most effective in promoting cooperation within the coevolutionary interplay of strategy and network. The advantage of SS lies in its greater benevolence towards unconditional cooperation compared to SJ, which enhances the reputation of cooperators. As a result, cooperators can form stable clusters to resist exploitation by defectors and prevent the emergence of highly heterogeneous interaction networks. Conversely, the IS norm, in which only the action of the focal agent matter, has been proven to be the most effective in promoting cooperation under time-invariant interaction network conditions, as shown in Fig. 6d. Furthermore, SH is detrimental to cooperation in all cases, as it often results in the indiscriminate assignment of negative reputations. This negative effect has been observed not only in small-scale societies but also in more complex environments [41], [54].

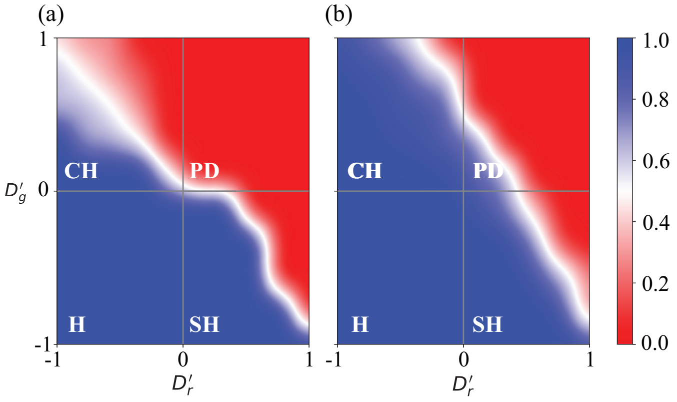


Fig. 7. **Equilibrium fraction of cooperators in $D'_r - D'_g$ diagrams.** A similar promotion effect of the proposed mechanism in a weak PD game is observed in three social dilemma models. (a) Conventional interaction model with network reciprocity ($T = \infty$ and $\beta = 0$). (b) Reputation-based interaction model with RL for the fixed parameters at $T = 10$ and $\beta = 0.1$. From red to blue, the colour bar indicates that the cooperation level changes from 0 to 1 accordingly.

E. Universal scaling for dilemma strength

It should be noted that the discussions regarding reputation-based interaction have primarily focused on a simplified payoff matrix M, where the value for R and S are set to 0. Although this weak version of the PD game has been widely used due to its consistent qualitative outcome, it would be beneficial to compare the findings presented in this study with Donor and Recipient (D&R) game, as well as other 2×2 games. Nowak [55] utilized c/b as a scaling parameter to quantify the dilemma strength as indicated in the payoff matrix mentioned in Eq. 1. However, this particular representation is applicable only within the context of the D&R game. Inspired by this concept, Tanimoto and Sagara [56] categorized pairwise games into two types: gamble-intending dilemmas (GID) and risk-averting dilemmas (RAD), defined by $D_g = T - R$ and $D_r = P - S$, respectively. However, D_r and D_g are insufficient for accurately indicating the dilemma strength when a specific reciprocity mechanism is introduced into the game. To address this limitation, Wang et al. [57] introduced a new set of scaling parameters to measure the dilemma strength, $D'_g = (T - R)/(R - P)$ and $D'_r = (P - S)/(R - P)$ for any matrix game. Consequently, the payoff matrix is rescaled as

$$M = \begin{pmatrix} R & P - (R - P)D'_r \\ R + (R - P)D'_g & P \end{pmatrix}. \tag{15}$$

Depending on these two dilemma strengths, the game can be classified into four classes using a RAD-GID diagram [58]. The quadrants in the diagram represent different game structures: the PD, chicken (CH), harmony (H) and stag hunt (SH) games, respectively. A detailed description of each game class and its corresponding region can be found in Ref. [59]. To maintain mathematical generality, we assume $P = 0$ and $R = 1$, consistent with the setting defined in Eq. 6.

In Fig. 7, we verify the promoting effect of the proposed model on above 2×2 games by adjusting the parameters of dilemma strength (D'_g and D'_r). Aligning with previous conclusions drawn from the weak PD, the cooperation level is still enhanced as reputation-based interaction is considered, effectively mitigating both GID and RAD. As shown in Fig. 7b, the boundary points between the four game classes shift upward and to the right. Additionally, this mechanism exhibits superior performance in the CH game compared to the PD and SH games. Last but not least, although our simulation results are obtained on a large interaction network where the initial structure is a square lattice with von Neumann neighbourhood, we conducted validation experiments to assess the robustness and generalizability of our findings. The results demonstrate the consistent performance and effectiveness of our proposed model across various population sizes and structures. Further details and numerical analyses can be found in Section III of the supplementary material.

V. Conclusion

To investigate the interplay between evolving networks and the evolution of cooperation, we study the reputation-based interaction dynamics of RL agents in PD games. Our proposed model aligns conceptually with the interaction-updating coevolutionary rules proposed in Ref. [10]. Building upon prior work regarding strategy-dependent propensity to change links, we developed a novel framework that allows for modifying social connections based on observed reputation information. Specifically, RL agents can adjust their interaction intensity with neighbours and rewire connections accordingly. The fundamental concept for severing a link remains the same, where an agent disconnects with the most significantly malfunctioning relationship to gain a better payoff. The combination of RL and traditional EGT in this study could contribute to the advancement of cooperative behaviour modelling and sheds light on the potential applications in artificial intelligence systems.

It is worth mentioning that disconnecting from dissatisfied neighbours in a dynamic network can be viewed as a type of expulsion [11], [60]. These studies assume that players have the ability to relocate uncooperative individuals to other inactive nodes within the network. However, network heterogeneity in our experiments is exacerbated because individuals are able to form new links based on reputation. Indeed, the success of our proposed coevolutionary rule in prompting cooperation can be attributed to the emergence of heterogeneous interaction networks and the assortative mixing among cooperative agents (more detail can be found in Ref. [61]). Moreover, the resulting interaction topology of a heterogeneous network varies under different dilemma strengths, which is consistent with previous findings showing that cooperators are more likely to establish relationships and exhibit loyalty due to their consistently positive reputation levels [62].

Our proposed model incorporates social norms that govern reputation dynamics through the mechanism of indirect reciprocity [40]. The emergent dynamic population structure is the outcome of reputation-based partner switching and interaction processes, which depends on a simple rule: helping others increases the likelihood of receiving help from someone else later on. Our works highlight the significant role of indirect reciprocity [63] in enabling RL agents to adjust their interaction intensity and social ties with others. While our experiments have primarily focused on a reputation dynamic in which individuals help their right neighbours to receive a good reputation, we find a more benevolent norm in the adaptive interaction pattern can create more effective social selection forces, thereby promoting cooperation. Noteworthy, the framework regarding reputation assignment developed here can be extended to higher-order social norms, such as past reputation [41]. Meanwhile, the use of local reputation and selection may lead to different outcomes in establishing and maintaining large-scale cooperation [64].

Furthermore, our investigation into the interplay of adaptive interaction intensity and link adaption has confirmed that dynamic networks produce a substantial amount of cooperation [12]. The optimal threshold for ineffective neighbours is contingent upon changes in the dilemma strength. The emergence of the interaction network in our model also gives an explanation regarding the formation of high heterogeneity networks in realistic societies. It is noteworthy to acknowledge that our approach does not impose any constraints on the growth of node degrees within the network, thereby generating a highly heterogeneous network structure. This, in turn, yields a negative assortative coefficient when the dilemma conditions are strong. Consequently, introducing a maximal degree may yield significantly different outcomes in the formation of the interaction system [65]. Furthermore, in our model, reputation assessment is grounded in the social preference hypothesis. A potential avenue for future research could involve investigating whether the domain of morality enhances the understanding of human decision-making in artificial intelligence systems [66]. Ongoing studies have commenced analyzing the emergent behaviour of intrinsically-motivated RL agents whose rewards are derived from moral theories [67]. As intelligent artificial agents are anticipated to engage in various coevolutionary dynamics, how cooperation can evolve has become increasingly complex.

References

[1] G. Hardin, "The tragedy of the commons," *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968.

[2] E. Pennisi, "On the origin of cooperation," *Science*, vol. 325, no. 5945, pp. 1196–1199, 2009.

[3] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, "Statistical physics of human cooperation," *Physics Reports*, vol. 687, pp. 1–51, 2017.

[4] S. A. Rhoads, K. M. Vekaria, K. O'Connell, H. S. Elizabeth, D. G. Rand, M. N. Kozak Williams, and A. A. Marsh, "Unselfish traits and social decision-making patterns characterize six populations of real-world extraordinary altruists," *Nature Communications*, vol. 14, no. 1807, pp. 1–15, 2023.

[5] D. G. Rand and M. A. Nowak, "Human cooperation," *Trends in Cognitive Sciences*, vol. 17, no. 8, pp. 413–425, 2013.

[6] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, 2006.

[7] N. Hanaki, A. Peterhansl, P. S. Dodds, and D. J. Watts, "Cooperation in evolving social networks," *Management Science*, vol. 53, no. 7, pp. 1036–1050, 2007.

[8] A. Szolnoki and X. Chen, "Alliance formation with exclusion in the spatial public goods game," *Physical Review E*, vol. 95, no. 5, p. 052316, 2017.

[9] A. Li, L. Zhou, Q. Su, S. P. Cornelius, Y.-Y. Liu, L. Wang, and S. A. Levin, "Evolution of cooperation on temporal networks," *Nature Communications*, vol. 11, no. 2259, pp. 1–9, 2020.

[10] M. Perc and A. Szolnoki, "Coevolutionary games—a mini review," *Biosystems*, vol. 99, no. 2, pp. 109–125, 2010.

[11] T. Ren and J. Zheng, "Evolutionary dynamics in the spatial public goods game with tolerance-based expulsion and cooperation," *Chaos, Solitons & Fractals*, vol. 151, p. 111241, 2021.

[12] D. G. Rand, S. Arbesman, and N. A. Christakis, "Dynamic social networks promote cooperation in experiments with humans," *Proceedings of the National Academy of Sciences*, vol. 108, no. 48, pp. 19 193–19 198, 2011.

[13] Y. Murase, C. Hilbe, and S. K. Baek, "Evolution of direct reciprocity in group-structured populations," *Scientific Reports*, vol. 12, no. 18645, pp. 1–16, 2022.

[14] J. Zheng, Y. He, T. Ren, and Y. Huang, "Evolution of cooperation in public goods games with segregated networks and periodic invasion," *Physica A: Statistical Mechanics and its Applications*, vol. 596, p. 127101, 2022.

[15] F. C. Santos, J. M. Pacheco, and T. Lenaerts, "Cooperation prevails when individuals adjust their social ties," *PLOS Computational Biology*, vol. 2, no. 10, pp. 1284–1291, 2006.

[16] I. S. Lim and N. Masuda, "To trust or not to trust: evolutionary dynamics of an asymmetric n-player trust game," *IEEE Transactions on Evolutionary Computation*, 2023 (Early Access).

[17] J. Wang and C. Xia, "Reputation evaluation and its impact on the human cooperation—a recent survey," *Europhysics Letters*, vol. 141, no. 2, p. 21001, 2023.

[18] C. Xia, J. Wang, M. Perc, and Z. Wang, "Reputation and reciprocity," *Physics of Life Reviews*, vol. 46, pp. 8–45, 2023.

[19] Z. Hu, X. Li, J. Wang, C. Xia, Z. Wang, and M. Perc, "Adaptive reputation promotes trust in social networks," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 4, pp. 3087–3098, 2021.

[20] J. Tanimoto, "Does information of how good or bad your neighbors are enhance cooperation in spatial prisoner's games?" *Chaos, Solitons & Fractals*, vol. 103, pp. 184–193, 2017.

[21] J. Li, C. Zhang, Q. Sun, Z. Chen, and J. Zhang, "Changing the intensity of interaction based on individual behavior in the iterated prisoner's dilemma game," *IEEE Transactions on Evolutionary Computation*, vol. 21, no. 4, pp. 506–517, 2016.

[22] R. Köster, D. Hadfield-Menell, R. Everett, L. Weidinger, G. K. Hadfield, and J. Z. Leibo, "Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents," *Proceedings of the National Academy of Sciences*, vol. 119, no. 3, p. e2106028118, 2022.

[23] K. R. McKee, E. Hughes, T. O. Zhu, M. J. Chadwick, R. Koster, A. G. Castaneda, C. Beattie, T. Graepel, M. Botvinick, and J. Z. Leibo, "Deep reinforcement learning models the emergent dynamics of human cooperation," *arXiv preprint arXiv:2103.04982*, 2021.

[24] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, and T. Graepel, "Cooperative ai: machines must learn to find common ground," *Nature*, vol. 593, no. 7857, pp. 33–36, 2021.

[25] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, J. Z. Leibo, and N. De Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 3040–3049.

[26] Y. Hou, M. Sun, Y. Zeng, Y.-S. Ong, Y. Jin, H. Ge, and Q. Zhang, "A multi-agent cooperative learning system with evolution of social roles," *IEEE Transactions on Evolutionary Computation*, 2023 (Early Access).

[27] L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, and Z. Wang, "Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning," *Nonlinear Dynamics*, vol. 108, no. 2, pp. 1837–1845, 2022.

[28] D. Jia, H. Guo, Z. Song, L. Shi, X. Deng, M. Perc, and Z. Wang, "Local and global stimuli in reinforcement learning," *New Journal of Physics*, vol. 23, no. 8, p. 083020, 2021.

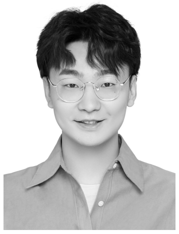
[29] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems*, 2017, pp. 464–473.

[30] R. Merhej, F. P. Santos, F. S. Melo, M. Chetouani, and F. C. Santos, "Cooperation and learning dynamics under risk diversity and financial incentives," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022, pp. 908–916.

[31] A. D. Martinez, J. Del Ser, E. Osaba, and F. Herrera, "Adaptive multifactorial evolutionary optimization for multitask reinforcement learning," *IEEE Transactions on Evolutionary Computation*, vol. 26, no. 2, pp. 233–247, 2021.

[32] N. Anastassacos, S. Hailes, and M. Musolesi, "Partner selection for the emergence of cooperation in multi-agent systems using reinforcement

learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 7047–7054.
[33] Z. Song, H. Guo, D. Jia, M. Perc, X. Li, and Z. Wang, "Reinforcement learning facilitates an optimal interaction intensity for cooperation," *Neurocomputing*, vol. 513, pp. 104–113, 2022.
[34] A. Rapoport, A. M. Chammah, and C. J. Orwant, *Prisoner's dilemma: A study in conflict and cooperation*. University of Michigan Press, 1965.
[35] D. W. Stephens, C. M. McLinn, and J. R. Stevens, "Discounting and reciprocity in an iterated prisoner's dilemma," *Science*, vol. 298, no. 5601, pp. 2216–2218, 2002.
[36] A. E. Roth and I. Erev, "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, vol. 8, pp. 164–212, 1995.
[37] P. Kiran and K. V. Chandrakala, "New interactive agent based reinforcement learning approach towards smart generator bidding in electricity market with micro grid integration," *Applied Soft Computing*, vol. 97, p. 106762, 2020.
[38] G. Szabó and C. Tőke, "Evolutionary prisoner's dilemma game on a square lattice," *Physical Review E*, vol. 58, pp. 69–73, 1998.
[39] M. A. Nowak and R. M. May, "The spatial dilemmas of evolution," *International Journal of Bifurcation and Chaos*, vol. 3, no. 01, pp. 35–78, 1993.
[40] M. A. Nowak and K. Sigmund, "Evolution of indirect reciprocity," *Nature*, vol. 437, no. 7063, pp. 1291–1298, 2005.
[41] F. P. Santos, F. C. Santos, and J. M. Pacheco, "Social norm complexity and past reputations in the evolution of cooperation," *Nature*, vol. 555, no. 7695, pp. 242–245, 2018.
[42] J. M. Pacheco, F. C. Santos, and F. A. C. Chalub, "Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity," *PLOS Computational Biology*, vol. 2, no. 12, pp. 1634–1638, 2006.
[43] B. Skyrms and R. Pemantle, "A dynamic model of social network formation," *Proceedings of the National Academy of Sciences*, vol. 97, no. 16, pp. 9340–9346, 2000.
[44] H. Ebel and S. Bornholdt, "Coevolutionary games on networks," *Physical Review E*, vol. 66, no. 5, p. 056118, 2002.
[45] G. Szabó and G. Fath, "Evolutionary games on graphs," *Physics Reports*, vol. 446, no. 4-6, pp. 97–216, 2007.
[46] F. Fu, C. Hauert, M. A. Nowak, and L. Wang, "Reputation-based partner choice promotes cooperation in social networks," *Physical Review E*, vol. 78, no. 2, p. 026117, 2008.
[47] F. C. Santos and J. M. Pacheco, "Scale-free networks provide a unifying framework for the emergence of cooperation," *Physical Review Letters*, vol. 95, no. 9, p. 098104, 2005.
[48] F. C. Santos, J. F. Rodrigues, and J. M. Pacheco, "Epidemic spreading and cooperation dynamics on homogeneous small-world networks," *Physical Review E*, vol. 72, no. 5, p. 056128, 2005.
[49] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, 2002.
[50] Z. Wang, S. Kokubo, J. Tanimoto, E. Fukuda, and K. Shigaki, "Insight into the so-called spatial reciprocity," *Physical Review E*, vol. 88, no. 4, p. 042145, 2013.
[51] Q. Su, B. Allen, and J. B. Plotkin, "Evolution of cooperation with asymmetric social interactions," *Proceedings of the National Academy of Sciences*, vol. 119, no. 1, p. e2113468118, 2022.
[52] K. Zhou and T. Ren, "Low-carbon technology collaborative innovation in industrial cluster with social exclusion: An evolutionary game theory perspective," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 3, p. 033124, 2021.
[53] J. Tanimoto, "Difference of reciprocity effect in two coevolutionary models of presumed two-player and multiplayer games," *Physical Review E*, vol. 87, no. 6, p. 062136, 2013.
[54] F. P. Santos, F. C. Santos, and J. M. Pacheco, "Social norms of cooperation in small-scale societies," *PLOS Computational Biology*, vol. 12, no. 1, pp. 1–13, 2016.
[55] M. A. Nowak, *Evolutionary dynamics: Exploring the equations of life*. Harvard University Press, 2006.
[56] J. Tanimoto and H. Sagara, "Relationship between dilemma occurrence and the existence of a weakly dominant strategy in a two-player symmetric game," *BioSystems*, vol. 90, no. 1, pp. 105–114, 2007.
[57] Z. Wang, S. Kokubo, M. Jusup, and J. Tanimoto, "Universal scaling for the dilemma strength in evolutionary games," *Physics of Life Reviews*, vol. 14, pp. 1–30, 2015.
[58] J. Tanimoto, *Sociophysics approach to epidemics*. Springer, 2021.
[59] H. Ito and J. Tanimoto, "Scaling the phase-planes of social dilemma strengths shows game-class changes in the five rules governing the evolution of cooperation," *Royal Society Open Science*, vol. 5, no. 10, p. 181085, 2018.
[60] S. Zhuo, J. Liu, T. Ren, and J. Sun, "Evolution dynamics with the switching strategy of punishment and expulsion in the spatial public goods game," *New Journal of Physics*, vol. 24, no. 12, p. 123020, 2022.
[61] J. Tanimoto, "The effect of assortative mixing on emerging cooperation in an evolutionary network game," in *Proceedings of the IEEE Congress on Evolutionary Computation*, 2009, pp. 487–493.
[62] S. Van Segbroeck, F. C. Santos, A. Nowé, J. M. Pacheco, and T. Lenaerts, "The evolution of prompt reaction to adverse ties," *BMC Evolutionary Biology*, vol. 8, no. 287, pp. 1–8, 2008.
[63] M. A. Nowak and K. Sigmund, "The dynamics of indirect reciprocity," *Journal of Theoretical Biology*, vol. 194, no. 4, pp. 561–574, 1998.
[64] S. Podder, S. Righi, and K. Takács, "Local reputation, local selection, and the leading eight norms," *Scientific Reports*, vol. 11, no. 16560, pp. 1–10, 2021.
[65] A. Szolnoki, M. Perc, and Z. Danku, "Making new connections towards cooperation in the prisoner's dilemma game," *Europhysics Letters*, vol. 84, no. 5, p. 50007, 2008.
[66] V. Capraro and M. Perc, "Mathematical foundations of moral preferences," *Journal of the Royal Society interface*, vol. 18, no. 175, p. 20200880, 2021.
[67] E. Tennant, S. Hailes, and M. Musolesi, "Modeling moral choices in social dilemmas with multi-agent reinforcement learning," *arXiv preprint arXiv:2301.08491*, 2023.



Tianyu Ren received the B.Sc degree in Management Information Systems from Wuhan University of Technology, and the M.Sc degree in Management Science and Engineering from Wuhan University, Wuhan, China, in 2019 and 2022, respectively. He is currently pursuing a Ph.D. degree in the Department of Computer Science, University of Manchester, Manchester, U.K.

His research interests include reinforcement learning, evolutionary game dynamics and cooperative intelligence.

Xiao-Jun Zeng received the B.Sc. degree in Mathematics and the M.Sc. degree in Control Theory and Operation Research from Xiamen University, Xiamen, China, and the Ph.D. degree in Computation from the University of Manchester, Manchester, U.K.

He has been with the University of Manchester since 2002, where he is currently a professor of Machine Learning in the Department of Computer Science. His current research interests include computational intelligence, machine learning, data mining, decision support systems, game theory, and their applications.

Prof. Zeng is an Associate Editor of the IEEE Transactions on Fuzzy Systems and several other journals. He received the European Information Society Technologies Award in 1999 and the Microsoft European Retail Application Developer Awards in 2001 and 2003 with KSS Ltd. His research in intelligent pricing decision support systems was selected by UK Computing Research Committee, Council of Professors and Heads of Computing, and British Computer Society Academy as one of 20 impact cases to highlight the impact made by UK academic Computer Science Research within the UK and worldwide over the period 2008 – 2013.