Applied Mathematics and Computation 463 (2024) 128364

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Interaction state Q-learning promotes cooperation in the spatial prisoner's dilemma game

Zhengzhi Yang [a], Lei Zheng [a], Matjaž Perc [b,c,d,e,f], Yumeng Li [a,*]

[a] Beihang University, Beijing 100191, PR China
[b] Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia
[c] Department of Medical Research, China Medical University Hospital, China Medical University, Taichung 404332, Taiwan
[d] Alma Mater Europaea, Slovenska ulica 17, 2000 Maribor, Slovenia
[e] Complexity Science Hub Vienna, Josefstädterstraße 39, 1080 Vienna, Austria
[f] Department of Physics, Kyung Hee University, 26 Kyungheedae-ro, Dongdaemun-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Evolutionary games
Cooperation
Prisoner's dilemma game
Reinforcement learning

ABSTRACT

Many recent studies have used reinforcement learning methods to investigate the behavior of agents in evolutionary games. Q-learning, in particular, has become a mainstream method during this development. Here we introduce Q-learning agents into the evolutionary prisoner's dilemma game on a square lattice. Specifically, we associate the state space of Q-learning agents with the strategies of their neighbors, and we introduce a neighboring reward information sharing mechanism. We thus provide Q-learning agents with the payoff information of their neighbors, in addition to their strategies, which has not been done in previous studies. Through simulations, we show that considering neighborhood payoff information can significantly promote cooperation in the population. Moreover, we show that for an appropriate strength of neighborhood payoff information sharing, a chessboard pattern emerges on the lattice. We analyze in detail the reasons for the emergence of the chessboard pattern and the increase in cooperation frequency, and we also provide a theoretical analysis based on the pair approximation method. We hope that our research will inspire effective approaches for resolving social dilemmas by means of sharing more information among reinforcement learning agents during evolutionary games.

1. Introduction

Cooperative behavior is widespread in both the natural world and human society, seemingly challenging Darwin's theory of evolution and the process of natural selection. Evolutionary game theory provides a practical framework to investigate the stable presence of cooperative behavior among rational agents [1–5]. The prisoner's dilemma game (PDG) is a typical game model that often used to illustrate the conflicts that arise between individuals and populations. [6–8]. In the classic model, two agents decide to cooperate (C) or defect (D) simultaneously. The reward (punishment) for both agents is R (P) if they cooperate (defect) mutually. If one agent cooperates and the other defects, the former obtains the sucker's payoff S, and the latter receives the temptation to defect T. The size relationship of the four parameters is $T > R > P > S$. Obviously, each agent tends to defect in order to maximize its own payoff, resulting in a social dilemma of mutual defection.

* Corresponding author.
E-mail address: liyumeng@buaa.edu.cn (Y. Li).

https://doi.org/10.1016/j.amc.2023.128364
Received 7 July 2023; Received in revised form 19 September 2023; Accepted 24 September 2023

Nowak and May explored PDG on square lattice network, which has shown that the spatial structure can promote the evolution of cooperation [9]. Following their pioneering work, many researches investigated role of different topology structures in evolutionary games, such as regular lattice [10–15], scale-free networks [16–21], *etc.* On the basis of network structures, numerous different mechanisms have been suggested to encourage cooperative behavior [22–41]. For example, Qin *et al.* empowered players to make decisions based on cumulative payoffs stored in memory found that moderate memory promotes cooperative behaviors [24]. Li *et al.* showed that continuously adjusting interaction intensity based on neighbor's reputation promotes formation and maintenance of cooperator clusters [28]. The central idea runs through these works is that players learn to make better decisions by interacting with their neighbors during evolution.

As a machine learning method, reinforcement learning focuses on how intelligent agents can make better decisions through iterated interactions with environment to maximize cumulative rewards. It has experienced rapid advancements in recent years and been successfully applied in many fields, including robot control [42–44], recommender systems [45–47], and so on [48–53]. Generally, a reinforcement learning agent can observe the surrounding environment, try to take action, and obtain the corresponding rewards. Through this iterative trial-and-error procedure, it can learn strategies leading to the maximum cumulative rewards. The evolution of game strategies of reinforcement learning agents has attracted great interest, and many studies have explored along this direction [54–63].

Among all the attempts mentioned above, Q-learning has gained significant attention due to its characteristics of simplicity, efficiency, and powerful ability to handle state-action spaces. For instance, Ding *et al.* utilized Q-learning method to play PDG with extortion action, and found that Q-learning can significantly promote cooperation compared to other traditional strategy updating rules [59]. Geng *et al.* investigated mixed agents taking different strategy update mechanisms (i.e., the Fermi rule and Q-learning) to promote cooperation. [63].

In order to enhance the decision-making capabilities of reinforcement learning agents, such as those based on Q-learning, it is customary to integrate supplementary information to assist them in better comprehending the details patterns in the environment [64,65]. This practice can accelerate the convergence of agents' strategies and improve their decision-making performance. In the evolutionary prisoner's dilemma game on the regular lattice, one crucial aspect of observational information that can be provided to the agents is knowledge of their own and the neighbor's strategies in the previous actual round, or the strategy predicted for this current round, which has been taken into account in previous research [62,63,66,67]. Furthermore, the payoff information of neighboring agents, which has been neglected in previous studies, also plays a significant role in shaping the agents' decision-making strategies. Therefore, in addition to providing the strategy information of the neighbors to the intelligent agent, we also provide the information of the neighbors' payoffs, and investigate how this additional information affects the agents' behavior.

In the remaining part of this article, we will begin by introducing the game model, Q-learning method, and neighboring reward information sharing mechanism used to incorporate neighbor's payoff information. Following that, we present our simulation results and analyses in detail. In the last section, our conclusions are summarized.

2. The model

In this paper, we consider the prisoner's dilemma game (PDG) on the $L \times L$ regular square lattice with periodic boundary conditions and von Neumann neighborhood, where each vertex represents an agent that can interact with the four nearest agents simultaneously. Maintaining generality, we use the simplified PDG framework, wherein the reward for mutual cooperation (R) is set to 1, the sucker's payoff (S) is 0, the temptation to defect is denoted as $T = b$ (where $1 \leq b \leq 2$), and the punishment for mutual defection (P) is fixed at 0. Then the payoff matrix can be expressed as:

$$A = \begin{pmatrix} 1 & 0 \\ b & 0 \end{pmatrix} \tag{1}$$

Each agent is designated to cooperate or defect with a coin toss initially, then can make decisions using a reinforcement learning method called ϵ-greedy Q-learning. Q-learning finds the optimal policy maximizing the expectation of the total reward for a given Markov Decision Process (MDP) denoted by a tuple $(\mathcal{S}, \mathcal{A}, \mathbf{P}, r)$, whereby $\mathcal{S}$ and $\mathcal{A}$ are the state space and action space, respectively. $\mathbf{P}$ represents the transition probabilities, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ assigns the agent's reward for adopting action a at state s.

In our model, at each time step t, the state s_i ($s_i \in \mathcal{S}$) of agent i is determined by the number of cooperators in its neighborhood N_i last round. That is, the state space $\mathcal{S} = \{0, 1, ...|N_i|\}$. Compared with the stateless Q-learning, the state s contains the information about the focal agent's neighbors' strategies. Each agent can choose which action to play at next time step from the action space $\mathcal{A} = \{C, D\}$, where C is represented by column vector $(1, 0)^T$ and D is $(0, 1)^T$. Then at time step $t+1$, given the joint action of agent i and its neighbors, the straightforward payoff of agent i can be calculated as:

$$P_i = \sum_{j \in N_i} a_i^T A a_j \tag{2}$$

The neighborhood reward information sharing mechanism is introduced in the reward shaping process, which enables the agent to take into account the payoff information of its neighbors when learning strategies. An information sharing strength parameter is defined as α_r to control the influence of information about neighbors' payoffs in the focal agent's learning process. Specifically, the reward of agent i for choosing action a_i in state s at time t can be calculated according to the following formula:

$$r_t^i(s, a_i) = (1 - \alpha_r) \times P_i + \alpha_r \times \bar{P}_{N_i} \tag{3}$$

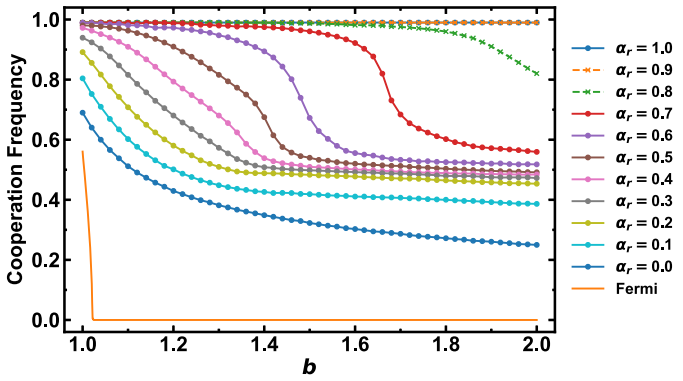

Fig. 1. The cooperation frequency obtained from simulation as a function of the temptation to defect b ($1 \le b \le 2$) for different values of information sharing strength parameter α_r. The orange line at the bottom represents the cooperation frequency for different b values when the agents are driven by the classic Fermi function.

$$\bar{P}_{N_i} = \frac{1}{|N_i|} \sum_{j \in N_i} P_j \tag{4}$$

where $0 \le \alpha_r \le 1$. Therefore, the reward r of agent i comprises the reward information of both itself and its neighbors. Then the Q-table will be updated according to the following formula:

$$Q_{t+1}(s,a) = (1-\alpha) \times Q_t(s,a) + \alpha \left[r_t(s,a) + \gamma \min_{a' \in A} Q_t(s',a') \right] \tag{5}$$

where $0 \le \alpha \le 1$ denotes the learning rate, $0 \le \gamma \le 1$ denotes the discount factor which represents how much the agents care about the future rewards, and s' is the next state.

After updating Q-table at each time step, agents choose the action to play in the next round according to the current state. Specifically, the agents choose the action with the highest Q-value in their current states with probability $1-\epsilon$, and randomly select one action with probability ϵ.

We compared the Monte Carlo simulation results under different sets of learning parameters, and finally set $\alpha = 0.1$, $\gamma = 0.9$, and $\epsilon = 0.02$. For more details about how the values of these parameters are chosen, readers are referred to the Appendix.

The Monte Carlo simulations are conducted on a regular square lattice network with a size of $L = 100$. We take the average of the last 5000 of the 2×10^5 total simulation steps with 50 independent runs to evaluate the steady states for a fixed set of parameter values.

3. Simulation results

First, we show the cooperation frequency as a function of the temptation to defect b under different values of information sharing strength parameter α_r in Fig. 1. It can be seen that the frequency of cooperation is effectively enhanced when the agents update strategies using the Q-learning method without neighboring reward information sharing ($\alpha_r = 0$) compared to the traditional Fermi's rule, where f_C declines to 0 steeply as b increases. We believe that the cooperation is promoted because of the random explorations and intrinsic fluctuations in the classic Q-learning method when $\alpha_r = 0$ (Readers are referred to the Appendix for more details). It can be seen that the cooperation frequency increases monotonically until almost 1 as α_r increases, no matter the value of b. However, the defectors still dominate for the majority of values of b. When the agents learn strategies with the neighboring reward information taken into account, the cooperation frequency is continuously significantly promoted as α_r increases. It's worth noting that f_C rises slightly around 0.5 when α_r increases from 0.2 to 0.6 for $1.5 \le b \le 2$. One can find that the f_C drops rapidly at first, then decreases marginally as b increases from 1 to 2 when $0 \le \alpha_r \le 0.5$. When $0.6 \le \alpha_r \le 0.7$, f_C drops rapidly when b is moderate and stays stable when b is relatively small or large. The cooperation frequency maintains close to 1 until b increases to more than 1.8 when $\alpha_r = 0.8$. The values of f_C are close to 1 when $0.9 \le \alpha_r \le 1$, regardless of the value of b. In brief, Q-learning with neighboring reward information sharing can promote cooperation.

To reveal the impact of different α_r values on the cooperation level, we begin by investigating how the cooperation frequency evolves over time for different α_r values. Fig. 2 provides the proportion of cooperators as a function of time for different values of α_r with fixed $b = 1.6$. One can observe that f_C decreases rapidly in the first few generations when $0 \le \alpha_r \le 0.3$. After that, f_C gradually rises back to a stable value which is lower than the initial cooperation frequency f_0. But the trends of f_C over time are contrary when $0.3 < \alpha_r \le 0.6$, it rises at first and falls into a stable value higher than f_0 subsequently. When $\alpha_r \ge 0.7$, it can be clearly seen that f_C steadily increases to a high value. Specifically, the stationary f_C is close to 0.9 with slight fluctuations when $\alpha_r = 0.7$. And it's almost 1 at the stable state when $0.8 \le \alpha_r \le 1.0$. Hence, we can conclude that the agents efficiently acquire cooperative strategies when the sharing of neighborhood reward information has a significant contribution ($alpha > 0.6$). Nevertheless, in other cases, the behavior of the agents is more complex. It gradually converges after experiencing fluctuations, maintaining a considerable proportion of cooperators.

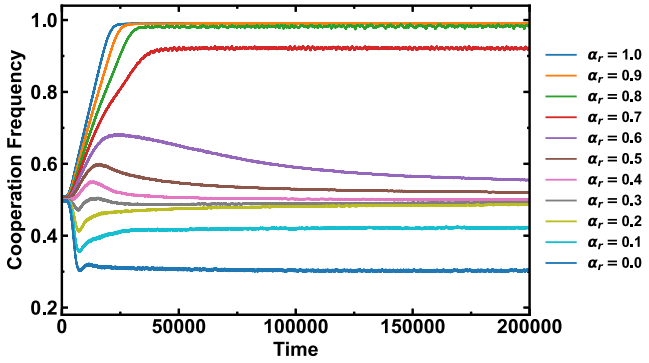

Fig. 2. The changes of the cooperation frequency over time for temptation to defect $b = 1.6$ and different values of neighboring reward information sharing strength α_r. Each curve in this panel is obtained by averaging the time series of 20 independent runs for the corresponding value of α_r.

Next, for a clearer view of the underlying mechanisms leading to the phenomenon observed above, we inspect the characteristic snapshots of strategies for different α_r values in different stages of evolution. It can be seen from Fig. 3(a)/(e)/(i)/(m)/(q) that for the first few steps, there is no obvious difference in the distribution of strategies for different α_r values. From Fig. 3(a)/(b)/(c)/(d), one can observe that the cooperators are invaded by defectors when $\alpha_r = 0$. And the surviving cooperators are scattered across the network, most of which form small and compact clusters to resist the defectors. When $0.2 \leq \alpha_r \leq 0.6$, the chessboard-like structure gradually forms in the characteristic snapshots. It is apparent in Fig. 3(h)/(l)/(p) that almost all agents are in the chessboard-like structure. Namely, nearly all the focal agents play the contrary action against their four neighbor, which makes the snapshots of strategies look like a chessboard. The corresponding partial enlarged patterns of the chessboard inside the yellow box in Fig. 3(h)/(l)/(p) are shown in Fig. 3(u)/(v)/(w). Besides, there are some line-like and dot-like gatherings of cooperators or defectors at the stable states. As can be seen from Fig. 3(t), cooperators dominate on the network when $\alpha_r = 0.8$, while sporadic agents explore defection.

To further investigate how the chessboard pattern forms and influences the evolution of cooperation in turn, we compute the difference between the rewards for cooperation and defection using the toy model shown in Fig. 4. The blue node at the center represents the focal agent. We represent the set of agents shown in Fig. 4, excluding the focal agent i and its four immediate neighbors, as M_i. According to Eqs. (3) and (4), the reward of the focal agent is determined by its own payoff and the average payoff of its four neighbors in the blue dashed box. Taking the upper neighbor of the center agent as an example, its payoff is affected by the strategies of the agents in the light blue dashed box. Therefore, the reward of the focal agent can be obtained considering all the strategies of agents in the toy model. We can derive the difference between the rewards for cooperation and defection as follows:

$$r_C - r_D = \alpha_r \left[\frac{3}{4} n_C (b-1) + b \right] + n_C (1-b) \tag{6}$$

where n_C denotes the number of cooperative neighbors of the focal agent. As shown in Fig. 3, there is global chessboard pattern when $0.2 \leq \alpha \leq 0.6$, which illustrates that the focal agent can obtain higher rewards by cooperating (defecting) if all its neighbors are defectors (cooperators). That is to say, $r_C - r_D$ is always positive when $n_C = 0$ for cooperators and negative when $n_C = 4$ for defectors in the chessboard. According to Eq. (6), the values of $r_C - r_D$ are 0.32, 0.64, and 0.96 when $n_C = 0$ and $b = 1.6$ for $\alpha = 0.2$, 0.4, and 0.6, respectively. The values of $r_C - r_D$ are −1.72, −1.04, and −0.36 when $n_C = 4$ and $b = 1.6$ for $\alpha = 0.2$, 0.4, and 0.6, respectively. This can explain why the chessboard structure emerges during the evolution for the appropriate value of α_r.

Combining the changes in the reward difference between different strategies during evolution and the formation of chessboard pattern together provides a more profound comprehension of the trends in cooperation frequency over time. The stable cooperation frequencies for different values of b and α_r are shown in Fig. 5(a), where the five curves from bottom to top correspond to the values of α_r and b satisfy that $r_C - r_D = 0$ when $n_C = 0, 1, ..., 4$, respectively. Given $b = 1.6$, the values of α_r on the curves are 0, 0.29, 0.48, 0.61, and 0.70, respectively. According to Eq. (6), the reward for defection is always higher for $n_C > 0$ when $\alpha_r < 0.29$, yet lower for $n_C = 0$ no matter the values of α_r. Take the case of $\alpha_r = 0.2$ as an example, the time series of the cooperation frequency f_C and chessboard ratio f_B are plotted in Fig. 6(a). For the majority of the agents, n_C is 1, 2, or 3 for the first few steps after initialization. The agents are inclined to defect when $n_C > 0$ after learning the knowledge about rewards for different states and actions, which leads to the decrease of f_C. Consequently, the number of cooperative neighbors n_C for some defectors drops to 0, forcing them into cooperating to obtain higher rewards. Hence, the stable chessboard subpattern that the cooperators surrounded by defectors arises. As the chessboard pattern gradually forms and expends to global, a further decrease in f_C is prevented. And the cooperation frequency eventually rises back to a value close to 0.5. The chessboard pattern can exist stably because the agents with it have no motivation to change their strategies except through random exploration with a probability of ϵ. Fig. 6(b) shows the time series of the f_C and f_B when $\alpha_r = 0.6$, where f_C rises first and then drops. When $\alpha_r = 0.6$, cooperation takes advantage for $n_C < 4$ according to Eq. (6), so the agents tend to cooperate at first. As a consequence, the value of n_C reaches 4 for some agents with the increase of f_C, which tempts the agents to defect for higher rewards. Therefore, the chessboard pattern forms and expands, leading the f_C to drop into a value slightly higher than 0.5. It can be concluded that when the value of α_r satisfies that cooperation and defection obtains higher rewards when $n_C = 0$ and $n_C = 4$, respectively, the chessboard structure will emerge and expand in the network, and the stationary cooperation frequency will approach 0.5 (Half of the agents are cooperators).

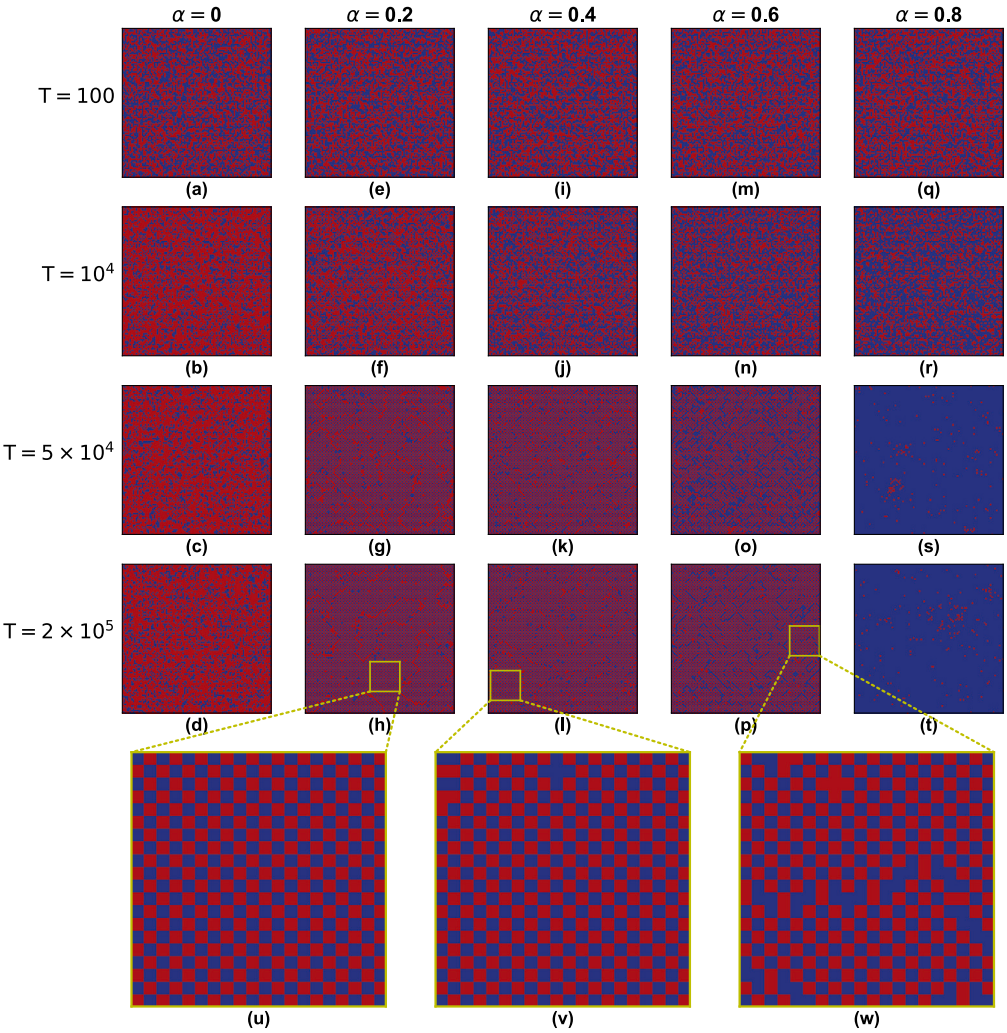

Fig. 3. The evolutionary characteristic snapshots of strategies over time for difference values of α_r. Cooperators and defectors are colored by blue and red, respectively. Panels (a)-(t), The columns are ordered from left to right as follows: $\alpha_r = 0$, $\alpha_r = 0.2$, $\alpha_r = 0.4$, $\alpha_r = 0.6$, and $\alpha_r = 0.8$, and the rows are ordered from top to bottom as: $T = 100$, $T = 10^4$, $T = 5 \times 10^4$, and $T = 2 \times 10^5$. Panels (u), (v), and (w) are the partial enlarged patterns of the area in yellow box shown in subplot (h), (l), and (p), respectively. The depicted results in all panels were obtained by employing $b = 1.6$ on a regular lattice with a size of 100×100.

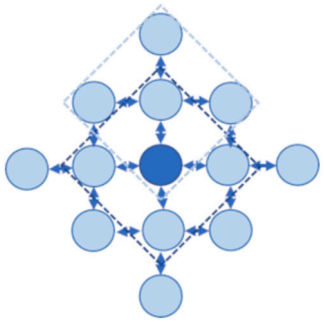

Fig. 4. The panel shows the toy model used to calculate the rewards for the central agent (blue). The reward for the focal agent is the weighted sum of its own payoff and the average of its neighbors' payoff. The focal agent acquires its payoff by gaming with its four neighbors in the blue dashed box. The payoff of the central agent's four neighbors can be obtained by means of the same way. Hence, the reward for the focal agent can be determined only considering the strategies of all the agents shown in this subpattern.

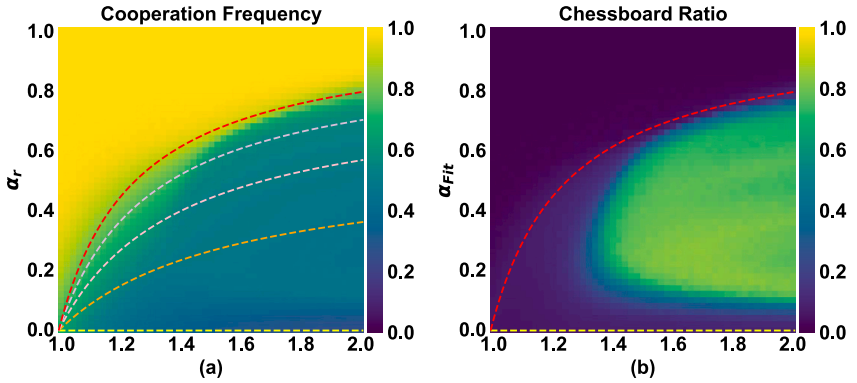

Fig. 5. Panel (a) depicts the chessboard ratio f_B in $b-\alpha_r$ panel. The five lines from bottom to top: values of b and α_r that meet the condition $r_C - r_D = 0$ when $n_C = 0$, 1, 2, 3, and 4, respectively. Panel (b) represents the stable cooperation frequency f_C for different values of b and α_r. Herein, we define the chessboard ratio f_B as the proportion of the agents taking contrary strategy against all its four neighbors. The yellow and red dashed line correspond to the values of b and α_r satisfy that $r_C - r_D = 0$ when $n_C = 4$ and $n_C = 0$, respectively.

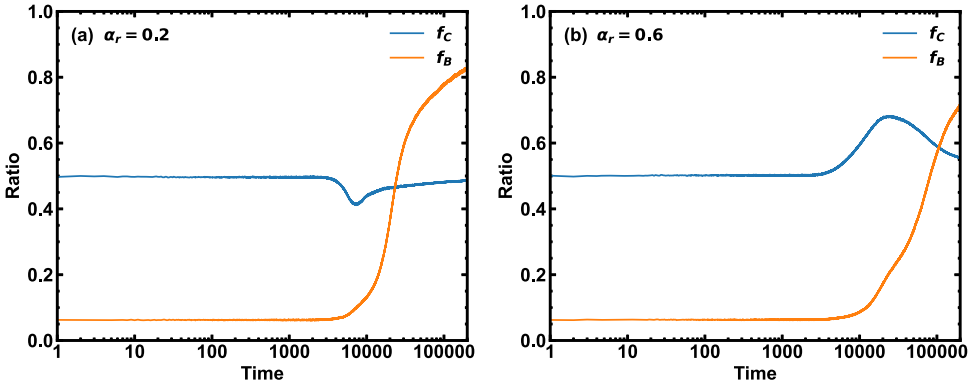

Fig. 6. Panel (a) and (b) demonstrate the temporal changes in the cooperation frequency f_C and the chessboard ratio f_B over time for $\alpha_r = 0.2$ and $\alpha_r = 0.6$, respectively. The value of b is set at a constant 1.6.

As shown in Fig. 5(a), the values of stable cooperation frequencies f_C are almost 1 when the parameter points (b, α_r) are above the red dashed line. This is because $r_C - r_D$ is always positive for those values of α_r and b, which lead the agents throughout the network to cooperate. Then we show the stable chessboard ratio for different values of b and α_r in Fig. 5(b). The yellow and red dashed lines indicate parameter points (b, α_r) satisfy that $r_C - r_D = 0$ when $n_C = 4$ and $n_C = 0$. It can be seen that the parameter combinations of b and α_r in the area between the two lines lead to higher chessboard ratios. However, there is barely any chessboard structure in the area above the red dashed line. The value of α_r and b above (below) each line in Fig. 5(a) and (b) make $r_C - r_D > 0$ ($r_C - r_D < 0$) for the corresponding n_C. Therefore, theoretically, the chessboard pattern can form and exist stably only for the values of α_r and b located between the lines in Fig. 5(b), which is consistent with the Monte Carlo simulation results.

We further verify the analysis of the formation reason of the chessboard pattern through statistical analysis and the results obtained from the simulation. Due to the randomness and fluctuation of the number of cooperative neighbors, we consider the expectation of the focal agent's reward computed using the subpattern shown in Fig. 4. The configuration of the agents in M_i is designated according to global cooperation frequencies. Table 1 shows the expectation computed under different values of n_C and f_C. When there are no cooperators around the focal agent ($n_C = 0$), the cooperative action obtains a higher reward under different values of f_C, making the cooperator that defectors surround tend to insist on cooperating. In contrast, defectors surrounded by cooperators ($n_C = 4$) will persist in defecting. Therefore, the chessboard pattern can form gradually and exist stably during fluctuating variation or when f_C declines. Fig. 7 shows that the size relation of the Q-value for cooperation and defection obtained from practical training matches the computed expectation result.

Next, we turn to the validation of the analysis on promoting cooperation, the change in the difference between rewards for cooperation and defection with neighboring reward information sharing can be derived using the following formula:

Table 1
The reward expectations for different actions of the focal agent in different states. The results were obtained using the subpattern shown in Fig. 4. The number of cooperators in N_i is n_C. The strategies of the agents in M_i are designated according to the cooperation frequency.

Cooperation Frequency	n_C	0	1	2	3	4
0	C	**0.113**	0.902	1.69	2.478	3.267
	D	0	**1.36**	**2.72**	**4.08**	**5.44**
0.2	C	**0.181**	0.963	1.744	2.525	3.307
	D	0.068	1.421	2.774	4.127	5.48
0.4	C	**0.249**	1.024	1.798	2.572	3.347
	D	0.136	**1.482**	**2.828**	**4.174**	**5.52**
0.6	C	**0.317**	1.085	1.852	2.619	3.387
	D	0.204	**1.543**	**2.882**	**4.221**	**5.56**
0.8	C	**0.385**	1.146	1.906	2.666	3.427
	D	0.272	**1.604**	**2.936**	**4.268**	**5.6**
1	C	**0.453**	1.207	1.96	2.713	3.467
	D	0.34	**1.665**	**2.99**	**4.315**	**5.64**

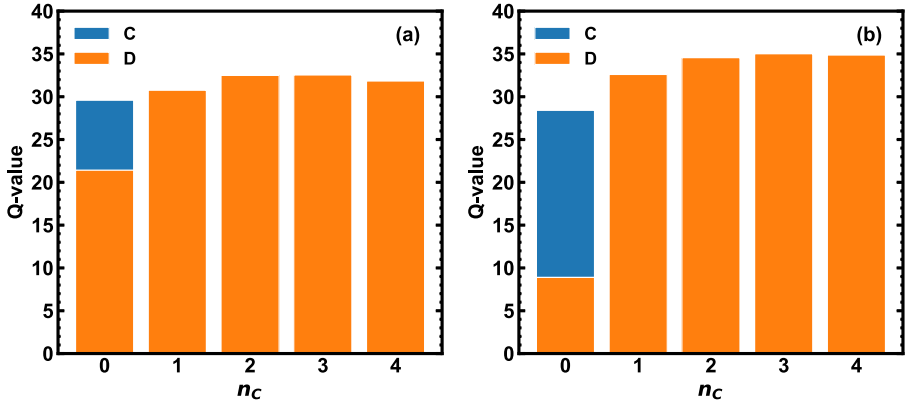

Fig. 7. The panels show the typical Q-table results. The five bars from left to right in each panel depict the Q-value for both actions at state 0, 1, 2, 3, and 4, respectively. The blue (orange) bars represent the Q-value for cooperation (defection). The parameters are set as $b = 1.6$ and $\alpha_r = 0.2$.

$$
\begin{aligned}
&\Delta(r_t(s,C) - r_t(s,D)) \\
&\quad = \left[r_t(s,C) - r_t(s,D)\right] - \left[P_t(s,C) - P_t(s,D)\right] \\
&\quad = \left(1-\alpha_r\right)\left[\left(P_C + \alpha_r \bar{P}_{N_C}\right) - \left(P_D + \alpha_r \bar{P}_{N_D}\right)\right] - \left(P_C - P_D\right) \\
&\quad = -\alpha_r\left(P_C - P_D\right) + \alpha_r\left(\bar{P}_{N_C} - \bar{P}_{N_D}\right) \\
&\quad = \alpha_r n_C(b-1) + \alpha_r\left(\frac{n_C + (4-n_C)b}{4} - 0\right) \\
&\quad = \alpha_r\left[0.75(b-1)n_C + b\right] \geq 0
\end{aligned}
\tag{7}
$$

Eq. (7) indicates that the advantage of cooperation over defection is promoted after taking neighbors' payoff into account. Table 2 presents the proportion of the agents with a higher Q-value for defection at state n_C, but with a higher one for cooperation at state n_C after considering neighbors' payoff when $\alpha_r = 0.2$ and $b = 1.6$. It is evident that a significant number of defectors become cooperators after the neighboring reward information sharing mechanism is introduced, resulting in an increase in cooperation frequencies. This is consistent with our previous inference.

The influence of α_r on cooperation frequency can be qualitatively predicted using the minimum subpattern shown in Fig. 4. Taking into account the spatial structure and reward function, the central agent updates its strategy using the Q-learning method, while the strategies of other agents are designated according to f_C. Pair configurations $p_{x,x'}$ indicate the probability of finding an individual playing strategy x accompanied by a neighbor playing s', where $x, x' \in \mathcal{A} = \{C, D\}$ in the PDG model used. Furthermore,

Table 2
The proportion of agents with a higher Q-value for defection at state n_C originally, but with a higher Q-value for cooperation after considering neighbors' payoffs. The parameters are fixed as $b = 1.6$ and $\alpha_r = 0.2$.

	n_C	0	1	2	3	4
$Q_r(s,C) > Q_r(s,D)$, $Q_P(s,C) < Q_P(s,D)$	$\alpha_r = 0.2$	0.441	0.348	0.295	0.173	0.183
	$\alpha_r = 0.4$	0.572	0.426	0.321	0.207	0.183
	$\alpha_r = 0.6$	0.545	0.523	0.442	0.340	0.223
	$\alpha_r = 0.8$	0.411	0.418	0.608	0.664	0.753
	$\alpha_r = 1.0$	0.403	0.419	0.603	0.723	0.757

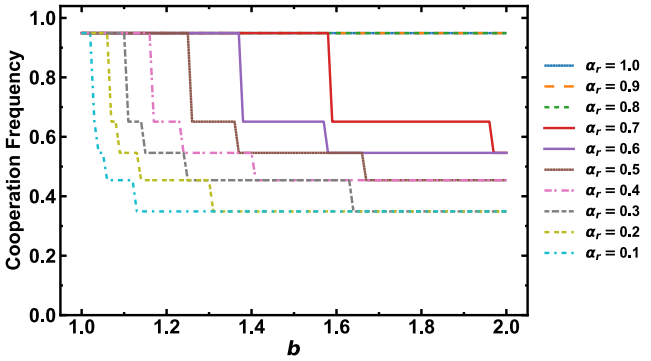

Fig. 8. Cooperation frequency obtained from theoretical analysis using the minimum subpattern as a function of b ($1 \le b \le 2$) under different values of α_r.

if agent i switches from cooperation to defection, the probabilities $p_{C,D}$ and $p_{C,C}$ decrease, whereas the probabilities $p_{D,D}$ and $p_{D,C}$ increase. These changes can be represented using the following formula:

$$\begin{aligned}\dot{p}_{C,C} = &\sum_{i\in N_i}\sum_{i\in M_i} n_C(N_i)\prod_{j\in N_i}\prod_{k\in M_i} p_{D,s_j}p_{D,s_k} \\ &\times f\left(r_C(N_i,M_i) - r_D(N_i,M_i)\right) \\ &- \sum_{i\in N_i}\sum_{i\in M_i} n_C(N_i)\prod_{j\in N_i}\prod_{k\in M_i} p_{C,s_j}p_{C,s_k} \\ &\times f\left(r_D(N_i,M_i) - r_C(N_i,M_i)\right)\end{aligned} \tag{8}$$

$$\begin{aligned}\dot{p}_{C,D} = &\sum_{i\in N_i}\sum_{i\in M_i} \left(4 - 2n_C(N_i)\right)\prod_{j\in N_i}\prod_{k\in M_i} p_{D,s_j}p_{D,s_k} \\ &\times f\left(r_C(N_i,M_i) - r_D(N_i,M_i)\right) \\ &- \sum_{i\in N_i}\sum_{i\in M_i} \left(4 - 2n_C(N_i)\right)\prod_{j\in N_i}\prod_{k\in M_i} p_{C,s_j}p_{C,s_k} \\ &\times f\left(r_D(N_i,M_i) - r_C(N_i,M_i)\right)\end{aligned} \tag{9}$$

$$f(r_C - r_D) = \begin{cases} 1 - \frac{\epsilon}{2}, & \text{if } r_C > r_D \\ \frac{\epsilon}{2}, & \text{else} \end{cases} \tag{10}$$

where $r_C(N_i, M_i)$ ($r_D(N_i, M_i)$) denotes the reward of the focal agent for cooperation (defection) given the strategy configurations of agents in N_i and M_i. Eqs. (8) and (9) are sufficient because of the symmetry ($p_{C,D} = p_{D,C}$) and $p_{C,C} + p_{C,D} + p_{D,C} + p_{D,D} = 1$. It can be seen from Fig. 8 that the cooperation frequency in the stationary state obtained from the theoretical analysis matches the simulation results. (Note that the discontinuities in Fig. 8 occur because the rewards for cooperation and defection are equal for these specific values of b.)

4. Conclusion

In summary, we investigated the evolution of cooperation on regular lattice networks where the Q-learning agents make decisions on the information about their neighbors' strategies and payoffs. Specifically, the state is represented by the number of cooperative neighbors, and the reward function combines the information on the neighbors' payoff by summing up the focal agent's payoff and its neighbors' average payoff with weight $1 - \alpha_r$ and α_r ($0 \le \alpha_r \le 1$), respectively. From the Monte Carlo simulation results, we found that taking the information on payoff into account can significantly promote cooperation. At the same

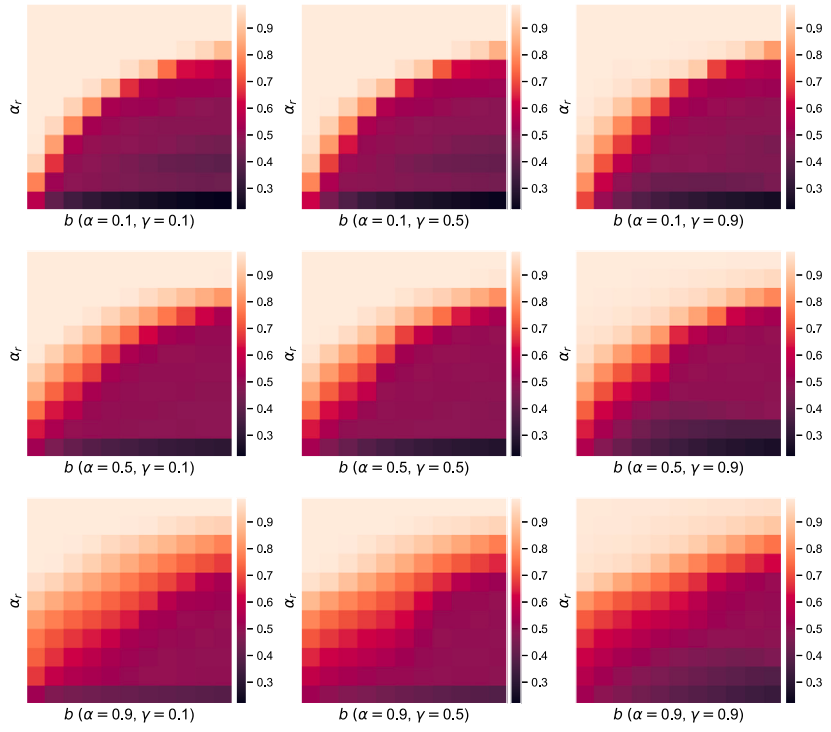

Fig. A.1. The cooperation frequency in $b - \alpha_r$ panel for different combinations of learning parameters α and γ.

time, for the moderate value of α_r, there is an obvious chessboard structure in the characteristic snapshot of the strategies at steady states. We further analyzed the reasons for the increase in cooperation frequency and the formation of the chessboard-like pattern. After that, we explored the relationship between the change in cooperation frequency and the variation in the proportion of the chessboard pattern during evolution. Finally, we show that theoretical analysis of cooperation frequencies based on the pair approximation method, which are consistent with the simulation results. We hope this work will inspire approaches for revealing how social dilemmas can be addressed by sharing more information in games among reinforcement learning agents.

Data availability

No data was used for the research described in the article.

Acknowledgements

This paper is supported by the National Key R&D Program of China under Grant 2022ZD0119600, National Natural Science Foundation of China under Grant 61961146005.

Appendix A

We conducted experiments on different parameter combinations to choose appropriate values for α and γ. Each α and γ value was selected from 0.1, 0.5, and 0.9. The cooperation frequencies in the $b - \alpha_r$ plane for different parameter combinations are shown in Fig. A.1. We ultimately selected $\alpha = 0.1$ and $\gamma = 0.9$. With these parameters, the proportion of dominant cooperation under different b and α_r values and the average cooperation frequency are higher.

There exists randomness and fluctuation to some extent during the training process. Firstly, the agent will randomly choose actions with a probability of ϵ. Besides, the next state after the agent after selecting an action is related to the strategies of other agents, the state transition and the cumulative value of the Q-table may cause the Q-values of C and D to fluctuate continuously (as shown in Fig. A.2). This effect is particularly evident for individuals situated on the edge of the checkerboard-shaped structure. However, for individuals in a stable chessboard structure, the fluctuations in the Q-table are usually smaller, as shown in Fig. A.3.

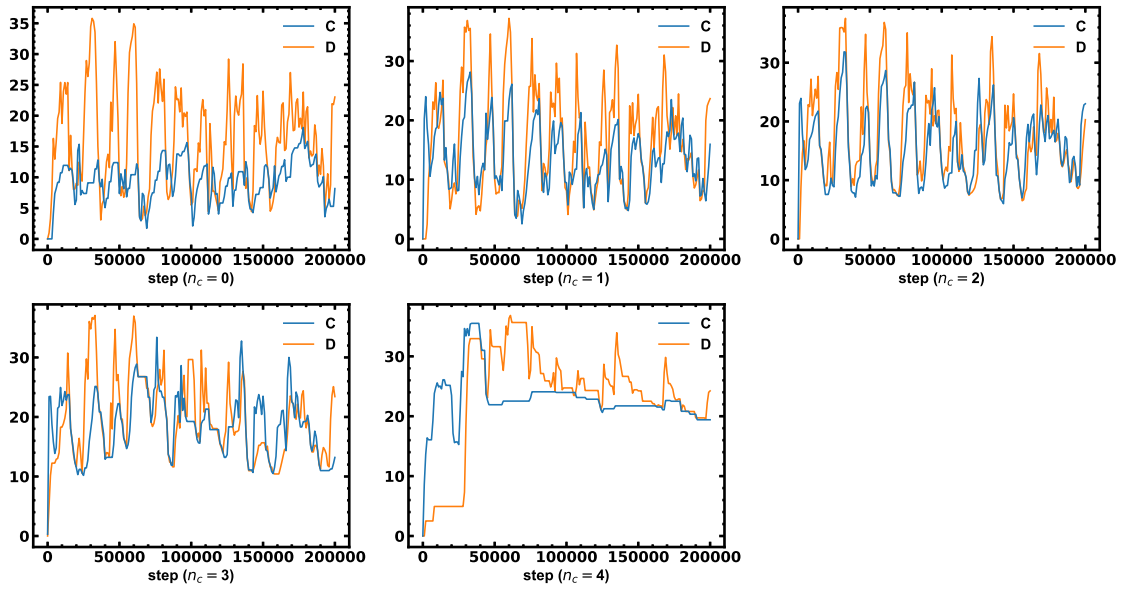

Fig. A.2. The fluctuation in the Q-table. This Q-table was obtained during the training process of one typical agent on the edge of the chessboard structure, showing the Q-values of two actions in different states.

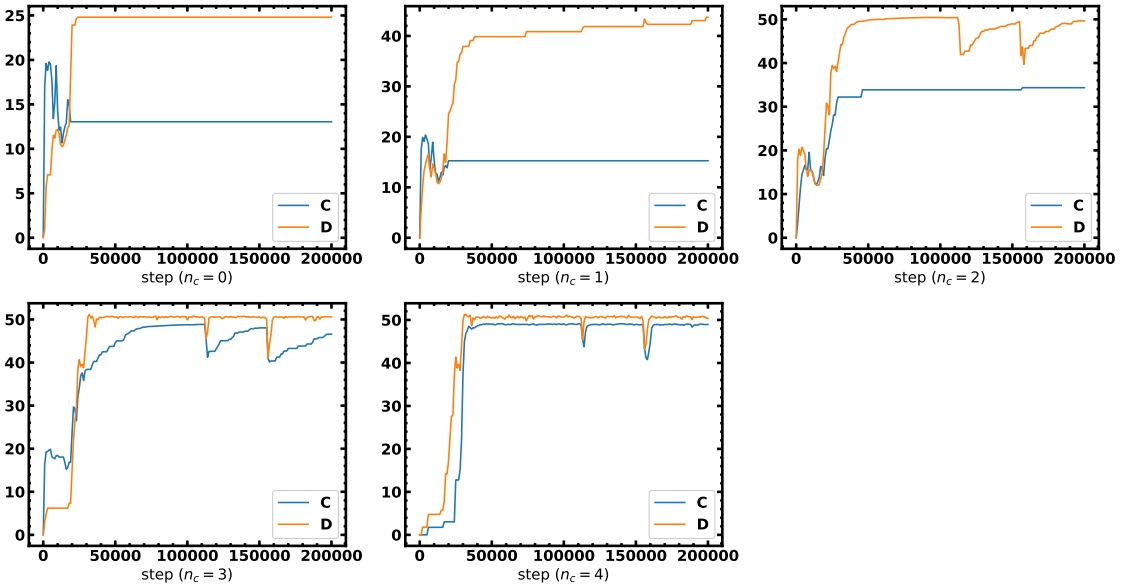

Fig. A.3. The Q-table obtained during the training process of one typical agent located in the relatively stable regions of the chessboard structure. It shows the Q-values of each action in different states.

References

[1] J. Von Neumann, O. Morgenstern, Theory of games and economic behavior, J. Philos. 42 (20) (1945) 550, https://doi.org/10.2307/2019327.
[2] R. Axelrod, W.D. Hamilton, The evolution of cooperation, Science 211 (4489) (1981) 1390–1396, https://doi.org/10.1126/science.7466396.
[3] J.M. Smith, Evolution and the Theory of Games, Cambridge University Press, Cambridge, 1982.
[4] J. Hofbauer, K. Sigmund, Evolutionary game dynamics, Bull. Am. Math. Soc. 40 (4) (2003) 479–519, https://doi.org/10.1090/S0273-0979-03-00988-1.
[5] M.A. Nowak, Evolutionary Dynamics: Exploring the Equations of Life, Harvard University Press, 2006.
[6] A. Rapoport, A.M. Chammah, C.J. Orwant, Prisoner's Dilemma: A Study in Conflict and Cooperation, vol. 165, University of Michigan Press, 1965.
[7] R. Axelrod, Effective choice in the prisoner's dilemma, J. Confl. Resolut. 24 (1) (1980) 3–25, https://doi.org/10.1177/002200278002400101.
[8] M. Perc, A. Szolnoki, Social diversity and promotion of cooperation in the spatial prisoner's dilemma game, Phys. Rev. E 77 (1) (2008) 011904, https://doi.org/10.1103/PhysRevE.77.011904.
[9] M.A. Nowak, R.M. May, Evolutionary games and spatial chaos, Nature 359 (6398) (1992) 826–829, https://doi.org/10.1038/359826a0.
[10] G. Szabó, C. Tőke, Evolutionary prisoner's dilemma game on a square lattice, Phys. Rev. E 58 (1) (1998) 69–73, https://doi.org/10.1103/PhysRevE.58.69.

[11] C. Hauert, M. Doebeli, Spatial structure often inhibits the evolution of cooperation in the snowdrift game, Nature 428 (6983) (2004) 643–646, https://doi.org/10.1038/nature02360.
[12] G. Szabó, J. Vukov, A. Szolnoki, Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices, Phys. Rev. E 72 (4) (2005) 047107, https://doi.org/10.1103/PhysRevE.72.047107.
[13] G. Szabó, G. Fáth, Evolutionary games on graphs, Phys. Rep. 446 (4) (2007) 97–216, https://doi.org/10.1016/j.physrep.2007.04.004.
[14] F. Fu, M.A. Nowak, C. Hauert, Invasion and expansion of cooperators in lattice populations: prisoner's dilemma vs. snowdrift games, J. Theor. Biol. 266 (3) (2010) 358–366, https://doi.org/10.1016/j.jtbi.2010.06.042.
[15] X. Meng, C. Xia, Z. Gao, L. Wang, S. Sun, Spatial prisoner's dilemma games with increasing neighborhood size and individual diversity on two interdependent lattices, Phys. Lett. A 379 (8) (2015) 767–773, https://doi.org/10.1016/j.physleta.2014.12.051.
[16] F.C. Santos, J.M. Pacheco, Scale-free networks provide a unifying framework for the emergence of cooperation, Phys. Rev. Lett. 95 (9) (2005) 098104, https://doi.org/10.1103/PhysRevLett.95.098104.
[17] F.C. Santos, J.M. Pacheco, T. Lenaerts, Evolutionary dynamics of social dilemmas in structured heterogeneous populations, Proc. Natl. Acad. Sci. 103 (9) (2006) 3490–3494, https://doi.org/10.1073/pnas.0508201103.
[18] Y. Chen, H. Lin, C. Wu, Evolution of prisoner's dilemma strategies on scale-free networks, Phys. A, Stat. Mech. Appl. 385 (1) (2007) 379–384, https://doi.org/10.1016/j.physa.2007.06.008.
[19] Z. Wu, J. Guan, X. Xu, Y. Wang, Evolutionary prisoner's dilemma game on Barabási–Albert scale-free networks, Phys. A, Stat. Mech. Appl. 379 (2) (2007) 672–680, https://doi.org/10.1016/j.physa.2007.02.085.
[20] H. Wang, Y. Sun, L. Zheng, W. Du, Y. Li, The public goods game on scale-free networks with heterogeneous investment, Phys. A, Stat. Mech. Appl. 509 (2018) 396–404, https://doi.org/10.1016/j.physa.2018.06.033.
[21] T. Cimpeanu, A. Di Stefano, C. Perret, T.A. Han, Social diversity reduces the complexity and cost of fostering fairness, Chaos Solitons Fractals 167 (2023) 113051, https://doi.org/10.1016/j.chaos.2022.113051.
[22] W. Wang, J. Ren, G. Chen, B. Wang, Memory-based snowdrift game on networks, Phys. Rev. E 74 (5) (2006) 056113, https://doi.org/10.1103/PhysRevE.74.056113.
[23] W. Du, H. Zheng, M. Hu, Evolutionary prisoner's dilemma game on weighted scale-free networks, Phys. A, Stat. Mech. Appl. 387 (14) (2008) 3796–3800, https://doi.org/10.1016/j.physa.2008.02.036.
[24] S. Qin, Y. Chen, X. Zhao, J. Shi, Effect of memory on the prisoner's dilemma game in a square lattice, Phys. Rev. E 78 (4) (2008) 041129, https://doi.org/10.1103/PhysRevE.78.041129.
[25] X. Cao, W. Du, Z. Rong, The evolutionary public goods game on scale-free networks with heterogeneous investment, Phys. A, Stat. Mech. Appl. 389 (6) (2010) 1273–1280, https://doi.org/10.1016/j.physa.2009.11.044.
[26] J. Wang, L.N. Liu, E.Z. Dong, L. Wang, An improved fitness evaluation mechanism with memory in spatial prisoner's dilemma game on regular lattices, Commun. Theor. Phys. 59 (3) (2013) 257–262, https://doi.org/10.1088/0253-6102/59/3/02.
[27] T.A. Han, L.M. Pereira, F.C. Santos, T. Lenaerts, Good agreements make good friends, Sci. Rep. 3 (1) (2013) 2695, https://doi.org/10.1038/srep02695.
[28] J. Li, C. Zhang, Q. Sun, Z. Chen, J. Zhang, Changing the intensity of interaction based on individual behavior in the iterated prisoner's dilemma game, IEEE Trans. Evol. Comput. 21 (4) (2017) 506–517, https://doi.org/10.1109/TEVC.2016.2628385.
[29] M.A. Javarone, D. Marinazzo, Evolutionary dynamics of group formation, PLoS ONE 12 (11) (2017) e0187960, https://doi.org/10.1371/journal.pone.0187960.
[30] Y. Li, J. Zhang, M. Perc, Effects of compassion on the evolution of cooperation in spatial social dilemmas, Appl. Math. Comput. 320 (2018) 437–443, https://doi.org/10.1016/j.amc.2017.10.002.
[31] A. Szolnoki, X. Chen, Environmental feedback drives cooperation in spatial social dilemmas, Europhys. Lett. 120 (5) (2018) 58001, https://doi.org/10.1209/0295-5075/120/58001.
[32] M.A. Amaral, M.A. Javarone, Heterogeneous update mechanisms in evolutionary games: mixing innovative and imitative dynamics, Phys. Rev. E 97 (4) (2018) 042305, https://doi.org/10.1103/PhysRevE.97.042305.
[33] Y. Li, H. Wang, W. Du, M. Perc, X. Cao, J. Zhang, Resonance-like cooperation due to transaction costs in the prisoner's dilemma game, Phys. A, Stat. Mech. Appl. 521 (2019) 248–257, https://doi.org/10.1016/j.physa.2019.01.088.
[34] M.A. Amaral, M.A. Javarone, Strategy equilibrium in dilemma games with off-diagonal payoff perturbations, Phys. Rev. E 101 (6) (2020) 062309, https://doi.org/10.1103/PhysRevE.101.062309.
[35] A. Kumar, V. Capraro, M. Perc, The evolution of trust and trustworthiness, J. R. Soc. Interface 17 (169) (2020) 20200491, https://doi.org/10.1098/rsif.2020.0491.
[36] M.H. Duong, T.A. Han, Cost efficiency of institutional incentives for promoting cooperation in finite populations, Proc. R. Soc. A, Math. Phys. Eng. Sci. 477 (2254) (2021) 20210568, https://doi.org/10.1098/rspa.2021.0568.
[37] A. Szolnoki, X. Chen, Cooperation and competition between pair and multi-player social games in spatial populations, Sci. Rep. 11 (1) (2021) 12101, https://doi.org/10.1038/s41598-021-91532-5.
[38] S. Wang, X. Chen, Z. Xiao, A. Szolnoki, Decentralized incentives for general well-being in networked public goods game, Appl. Math. Comput. 431 (2022) 127308, https://doi.org/10.1016/j.amc.2022.127308.
[39] H. Lee, C. Cleveland, A. Szolnoki, Mercenary punishment in structured populations, Appl. Math. Comput. 417 (2022) 126797, https://doi.org/10.1016/j.amc.2021.126797.
[40] C. Wang, A. Szolnoki, Inertia in spatial public goods games under weak selection, Appl. Math. Comput. 449 (2023) 127941, https://doi.org/10.1016/j.amc.2023.127941.
[41] T. Cimpeanu, F.C. Santos, T.A. Han, Does spending more always ensure higher cooperation? An analysis of institutional incentives on heterogeneous networks, Dyn. Games Appl. (2023), https://doi.org/10.1007/s13235-023-00502-1.
[42] W.D. Smart, L.P. Kaelbling, Effective reinforcement learning for mobile robots, in: Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292), vol. 4, 2002, pp. 3404–3410.
[43] P. Kormushev, S. Calinon, D.G. Caldwell, Robot motor skill coordination with EM-based reinforcement learning, in: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2010, pp. 3232–3237.
[44] S. Gu, E. Holly, T. Lillicrap, S. Levine, Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, in: 2017 IEEE International Conference on Robotics and Automation (ICRA), IEEE Press, Singapore, Singapore, 2017, pp. 3389–3396.
[45] X. Zhao, L. Xia, L. Zhang, Z. Ding, D. Yin, J. Tang, Deep reinforcement learning for page-wise recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 95–103.
[46] P. Wei, S. Xia, R. Chen, J. Qian, C. Li, X. Jiang, A deep-reinforcement-learning-based recommender system for occupant-driven energy optimization in commercial buildings, IEEE Int. Things J. 7 (7) (2020) 6402–6413, https://doi.org/10.1109/JIOT.2020.2974848.
[47] L. Huang, M. Fu, F. Li, H. Qu, Y. Liu, W. Chen, A deep reinforcement learning based long-term recommender system, Knowl.-Based Syst. 213 (2021) 106706, https://doi.org/10.1016/j.knosys.2020.106706.
[48] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489, https://doi.org/10.1038/nature16961.
[49] S. Shalev Shwartz, S. Shammah, A. Shashua Safe Multi-Agent, Reinforcement learning for autonomous driving, arXiv:1610.03295, https://doi.org/10.48550/arXiv.1610.03295, Oct. 2016.

[50] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of Go without human knowledge, Nature 550 (7676) (2017), https://doi.org/10.1038/nature24270.
[51] P. Andras, L. Esterle, M. Guckert, T.A. Han, P.R. Lewis, K. Milanovic, T. Payne, C. Perret, J. Pitt, S.T. Powers, N. Urquhart, S. Wells, Trusting intelligent machines: deepening trust within socio-technical systems, IEEE Technol. Soc. Mag. 37 (4) (2018) 76–83, https://doi.org/10.1109/MTS.2018.2876107.
[52] O. Vinyals, I. Babuschkin, W.M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D.H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al., Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354, https://doi.org/10.1038/s41586-019-1724-z.
[53] N.C. Luong, D.T. Hoang, S. Gong, D. Niyato, P. Wang, Y.C. Liang, D.I. Kim, Applications of deep reinforcement learning in communications and networking: a survey, IEEE Commun. Surv. Tutor. 21 (4) (2019) 3133–3174, https://doi.org/10.1109/COMST.2019.2916583.
[54] W.-b. Liu, X.-j. Wang, Dynamic decision model in evolutionary games based on reinforcement learning, Syst. Eng. - Theory Pract. 29 (3) (2009) 28–33, https://doi.org/10.1016/S1874-8651(10)60008-7.
[55] Z. Wang, A. Murks, W. Du, Z. Rong, M. Perc, Coveting thy neighbors fitness as a means to resolve social dilemmas, J. Theor. Biol. 277 (1) (2011) 19–26, https://doi.org/10.1016/j.jtbi.2011.02.016.
[56] A. Kianercy, A. Galstyan, Dynamics of Boltzmann Q learning in two-player two-action games, Phys. Rev. E, Stat. Nonlinear Soft Matter Phys. 85 (2012) 041145, https://doi.org/10.1103/PhysRevE.85.041145.
[57] Z. Wang, A. Szolnoki, M. Perc, Self-organization towards optimally interdependent networks by means of coevolution, New J. Phys. 16 (3) (2014) 033041, https://doi.org/10.1088/1367-2630/16/3/033041.
[58] T. Ezaki, Y. Horita, M. Takezawa, N. Masuda, Reinforcement learning explains conditional cooperation and its Moody cousin, PLoS Comput. Biol. 12 (7) (2016) e1005034, https://doi.org/10.1371/journal.pcbi.1005034.
[59] H. Ding, G. Zhang, S. Wang, J. Li, Z. Wang, Q-learning boosts the evolution of cooperation in structured population by involving extortion, Phys. A, Stat. Mech. Appl. 536 (2019) 122551, https://doi.org/10.1016/j.physa.2019.122551.
[60] Y. Shi, Z. Rong, Analysis of Q-learning like algorithms through evolutionary game dynamics, IEEE Trans. Circuits Syst. II, Express Briefs 69 (5) (2022) 2463–2467, https://doi.org/10.1109/TCSII.2022.3161655.
[61] Z. Song, H. Guo, D. Jia, M. Perc, X. Li, Z. Wang, Reinforcement learning facilitates an optimal interaction intensity for cooperation, Neurocomputing 513 (2022) 104–113, https://doi.org/10.1016/j.neucom.2022.09.109.
[62] L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, Z. Wang, Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning, Nonlinear Dyn. 108 (2) (2022) 1837–1845, https://doi.org/10.1007/s11071-022-07289-7.
[63] Y. Geng, Y. Liu, Y. Lu, C. Shen, L. Shi, Reinforcement learning explains various conditional cooperation, Appl. Math. Comput. 427 (2022) 127182, https://doi.org/10.1016/j.amc.2022.127182.
[64] L.P. Kaelbling, M.L. Littman, A.W. Moore, Reinforcement learning: a survey, J. Artif. Intell. Res. 4 (1996) 237–285, https://doi.org/10.1613/jair.301.
[65] M. Wiering, M. Van Otterlo (Eds.), Reinforcement Learning: State-of-the-Art, Adaptation, Learning, and Optimization, vol. 12, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
[66] T.A. Han, L.M. Pereira, F.C. Santos, Corpus-based intention recognition in cooperation dilemmas, Artif. Life 18 (4) (2012) 365–383, https://doi.org/10.1162/ARTL_a_00072.
[67] A. Di Stefano, C. Jayne, C. Angione, T.A. Han, Recognition of Behavioural Intention in Repeated Games Using Machine Learning, MIT Press, 2023.