




## PAPER

## Local and global stimuli in reinforcement learning

Danyang Jia<sup>1,2</sup>, Hao Guo<sup>1,2</sup>, Zhao Song<sup>1,2</sup>, Lei Shi<sup>3</sup>, Xinyang Deng<sup>4</sup>, Matjaž Perc<sup>5,6,7</sup>   
and Zhen Wang<sup>1,2,\*</sup><sup>1</sup> School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China<sup>2</sup> School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, People's Republic of China<sup>3</sup> School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming 650221, People's Republic of China<sup>4</sup> School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, People's Republic of China<sup>5</sup> Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, 2000 Maribor, Slovenia<sup>6</sup> Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan<sup>7</sup> Complexity Science Hub Vienna, Josefstädterstraße 39, 1080 Vienna, Austria

\* Author to whom any correspondence should be addressed.

E-mail: [zhenwang0@gmail.com](mailto:zhenwang0@gmail.com)**Keywords:** reinforcement learning, local and global stimuli, conditional cooperation, moody conditional cooperationRECEIVED  
16 June 2021REVISED  
20 July 2021ACCEPTED FOR PUBLICATION  
22 July 2021PUBLISHED  
10 August 2021

Original content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the  
title of the work, journal  
citation and DOI.



## Abstract

In efforts to resolve social dilemmas, reinforcement learning is an alternative to imitation and exploration in evolutionary game theory. While imitation and exploration rely on the performance of neighbors, in reinforcement learning individuals alter their strategies based on their own performance in the past. For example, according to the Bush–Mosteller model of reinforcement learning, an individual's strategy choice is driven by whether the received payoff satisfies a preset aspiration or not. Stimuli also play a key role in reinforcement learning in that they can determine whether a strategy should be kept or not. Here we use the Monte Carlo method to study pattern formation and phase transitions towards cooperation in social dilemmas that are driven by reinforcement learning. We distinguish local and global players according to the source of the stimulus they experience. While global players receive their stimuli from the whole neighborhood, local players focus solely on individual performance. We show that global players play a decisive role in ensuring cooperation, while local players fail in this regard, although both types of players show properties of 'moody cooperators'. In particular, global players evoke stronger conditional cooperation in their neighborhoods based on direct reciprocity, which is rooted in the emerging spatial patterns and stronger interfaces around cooperative clusters.

## 1. Introduction

Cooperative behavior is prevalent in nature and our society, even in the situations where free-riding is profitable [1–4]. Evolutionary game theory, based on two-player games (e.g. prisoner's dilemma games), multi-player games (e.g. public goods games), and their variants, serves as an efficient theoretical framework for exploring the emergence and maintenance of cooperative behavior [5–11]. To date, a large number of theoretical mechanisms have been proven to reveal the cooperative behavior of groups, such as self-organization [12], random walk within an appropriate range of temperatures [13], heuristics selection [14], higher-order interactions [15], in-group favoritism [16], social exclusion [17], assortativity [18], sentiment contagion [19], strategy equilibrium [20], risk perception [21], role specialization [22], cyclic dominance [23] and so on. Moreover, some pioneers have used statistical physics to reveal that adopting a strategy that punishes defectors while rewards cooperators can gain an evolutionary advantage [24, 25]. Most of these mechanisms revolve imitation (i.e. random imitation, imitation of the best, etc) [5, 26–30], however it is undeniable that some recent works have shown that in repeated dilemma games, individuals tend, based on limited information, to adopt simple and effective behavior patterns instead of adopting cautious and complex behavior patterns based on various information. For example, individuals exhibit

adaptive learning behavior, tit-for-tat (TFT), win-stay-lose-shift (WSLS), and so on, just based on limited historical information [31–35].

Individuals' adaptive behavior based on experience usually consists of two aspects. On the one hand, an individual's future strategy follows specific action rules. For example, always cooperating, always defection, TFT, tit-for-two-tats (TF2T), generous tit-for-tat (GTFT), WSLS, grim cooperate, and extortioner have been identified as representative action rules in repeated prisoner's dilemma games [36–39]. Experimental studies also suggest that participants' decision-making behavior can be characterized as noisy TFT, and it is the dominant strategy in a pairwise interactive environment [40]. In addition, the behavior patterns behind the demise of the commons across different cultures have also been studied [41].

On the other hand, humans and many species are capable of complex cognition, many of the cognitive skills have been considered as mechanisms for promoting the evolution of cooperation, such as learning [42], theory of mind [43], intent recognition [44, 45], intelligence [46], emotion [47, 48], etc. Here we focus on learning ability, and typically individuals use learning theory based on reinforcement learning to adjust their future decisions. Macy and Flache [49] used the traditional Bush–Mosteller (BM) stochastic learning model [50] for binary selection, and called it BM model of reinforcement learning. This model consists of two parts. At first, a player chooses an action based on the probability of cooperation and obtains the corresponding benefit. The player calculates her stimulus measured by whether the income satisfies aspiration. Second, driven by the reinforcement learning algorithm, the player updates the tendency of cooperation based on the current action and stimulus.

Following Macy and Flache's study, reinforcement learning mechanisms have attracted the attention of many scholars [51]. At present, for a fixed aspiration level, some researches have shown that BM players can cooperate with each other when payoff satisfies the aspiration [52–54]. In addition to changing actions, individuals can also adjust aspiration level, irrespective of BM reinforcement learning model or other reinforcement learning models [55]. In short, the principle of reinforcement learning is that individuals form two cognitive mechanisms, namely approach and avoidance, from experiential information. Approach means that payoff satisfies aspiration, so an individual's probability of repeating her previous action increases. Conversely, avoidance means that payoff is lower than aspirations, then individual's probability of repeating previous action decreases.

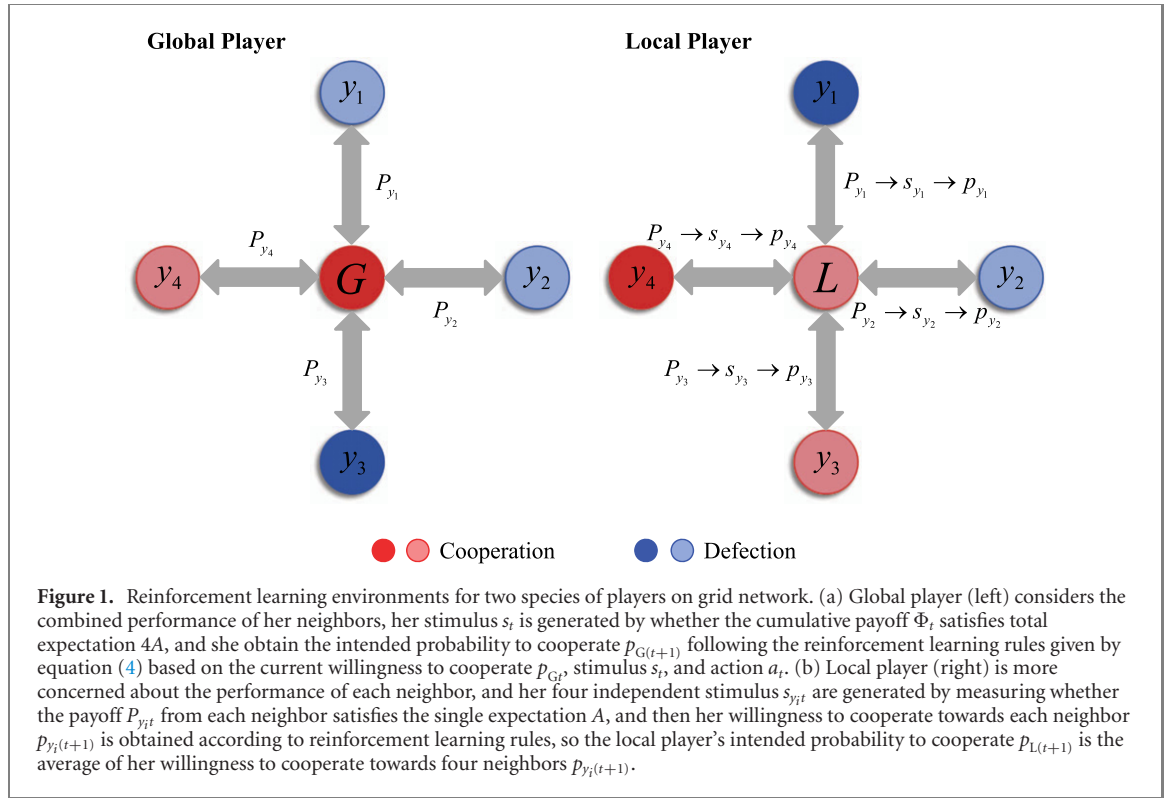
Considering that in the reinforcement learning mechanism, stimulus is measured by the individual's satisfaction with the payoff, and is used as an important indicator to drive the individual to adjust her action probability, it is essential to emphasize that each player receives payoffs with different sources. On the one hand, individuals get the corresponding payoff when interacting with each neighbor, and on the other hand, they receive cumulative payoffs. In view of this, it is natural to assume players' stimulus with various sources, like the payoff. Here, we divide agents into global players and local players, depending on their sources of stimulation. With such a framework, the global player's stimulus means the focal player's overall satisfaction with all neighbors in the neighborhood, which is measured by the difference between cumulative payoff and total expectation. The local player's stimulus means the focal player's satisfaction with a specific neighbor, which is determined by the difference between payoff from the neighbor and local expectation. In this article we focus on the performance of two types of players with different sources of stimulus under reinforcement learning rule. Simulation results show that global players play a leading role in promoting cooperation, and the probability of cooperation in the steady state follows two separated states, that is high cooperation and low cooperation. While the probability of cooperation of local players follows a normal distribution and assists the global players to achieve a high level of cooperation.

## 2. Methods

Players follow the reinforcement learning rule on the grid with  $100 \times 100$  nodes with periodic boundary conditions. All players decide whether to cooperate or defect according to an intended probability and play the prisoner's dilemma game with their four neighbors. If a pair of players choose to cooperate, they both gain the reward  $R = 1$ . If they both choose to defect, then they both get the punishment  $P = 0$ . If one chooses to cooperate and the other chooses to defect, the former gets the sucker's payoff  $S = 0$ , and the latter obtains the temptation  $T = b(b > 1)$ , respectively. Therefore, the cumulative benefit  $\Phi$  of focal agent reads as:

$$\Phi = \sum_{i=1}^4 P_{y_i}, \quad (1)$$

where  $P_{y_i}$  is the payoff that focal player gets from her neighbor  $y_i$  (figure 1). Since individual satisfaction with the current payoff will cause the fluctuation of individual emotion, stimulus,  $s_t$ , is measured as a function of the difference between payoff and expectation, according to different sources of stimulation.



Players in the network are divided into two categories (figure 1): global players and local players (which are randomly distributed on the network), and the proportion of global players is  $u$ . In particular, given an aspiration level  $A$ , global players care about the comprehensive performance of the entire neighborhood, while local players care more about the performance of each neighbor. Thus, for a global player, her stimulus,  $s_t$ , is measured by the difference between cumulative payoff  $\Phi_t$  and total expectation  $4A$ . While for a local player, she is faced with four independent stimulus from each neighbor,  $s_{y_{it}} (i = 1, 2, 3, 4)$ , measured by the difference between payoff  $P_{y_{it}}$  and single expectation  $A$  (figure 1). The details are as follows [51, 55]:

$$s_t = \tanh[\beta(\Phi_t - 4A)], \quad (2)$$

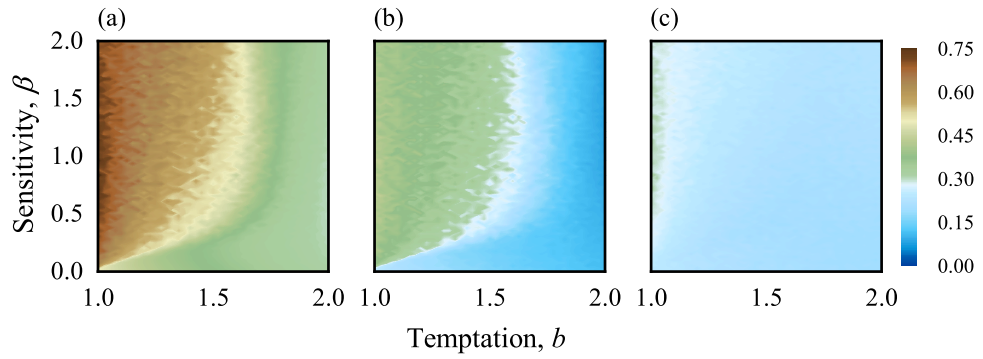
$$s_{y_{it}} = \tanh[\beta(P_{y_{it}} - A)], \quad (3)$$

where the aspiration level  $A$  is fixed at 0.5. Parameter  $\beta$  measures the sensitivity of the stimulus to the difference between payoff and expectation.

Further, players' current stimulus and action affect their intended probability to cooperate. Therefore, they update the tendency to cooperate,  $p_t$ , according to the reinforcement learning rule based on the current intended probability to cooperate  $p_t$ , stimulus  $s_t$ , and action  $a_t$ , BM model [51, 55]:

$$p_{t+1} = \begin{cases} p_t + (1 - p_t)s_t, & (a_t = C, s_t \geq 0) \\ p_t + p_t s_t, & (a_t = C, s_t < 0) \\ p_t - p_t s_t, & (a_t = D, s_t \geq 0) \\ p_t - (1 - p_t)s_t, & (a_t = D, s_t < 0). \end{cases} \quad (4)$$

Specifically, for global players, they obtain the intended probability to cooperate  $p_{G(t+1)}$  at round  $t + 1$  following the reinforcement learning rule based on the intended probability to cooperate  $p_{Gt}$ , stimulus  $s_t$ , and action  $a_t$  at round  $t$ . However, it is worth noting that local players are sensitive to the performance of each neighbor and can clearly feel the stimulus,  $s_{y_{it}}$ , from each neighbor, thereby generating an intended probability to cooperate for each neighbor,  $p_{y_{it}}$  (figure 1). Under such a situation, they first update the tendency to cooperate with each neighbor based on the reinforcement learning rules, then their tendency to cooperate  $p_{L(t+1)}$  is measured by the average of intended probability to cooperate for each neighbor,



**Figure 2.** Greater sensitivity to stimulus can trigger cooperative behavior of group. The probability to cooperate in the steady state as a function of the sensitivity  $\beta$  and the temptation to defect  $b$ . From left to right, the cooperative behavior in three situations: the total population, among the sub-population of global players, and the sub-population of local players are shown respectively. (a) Greater sensitivity to stimulus promote the group to reach a high level of cooperation. (b) Global players are more affected by the temptation to defect, and their cooperative behavior is dominant when  $b$  is small. (c) Local players are less affected by the temptation to defect, and their cooperative behavior is dominant when  $b$  is high. All results are obtained for  $A = 0.5$ ,  $u = 0.5$ .

$$p_{L(t+1)} = \frac{1}{4} \sum_{i=1}^4 p_{y_i(t+1)}. \quad (5)$$

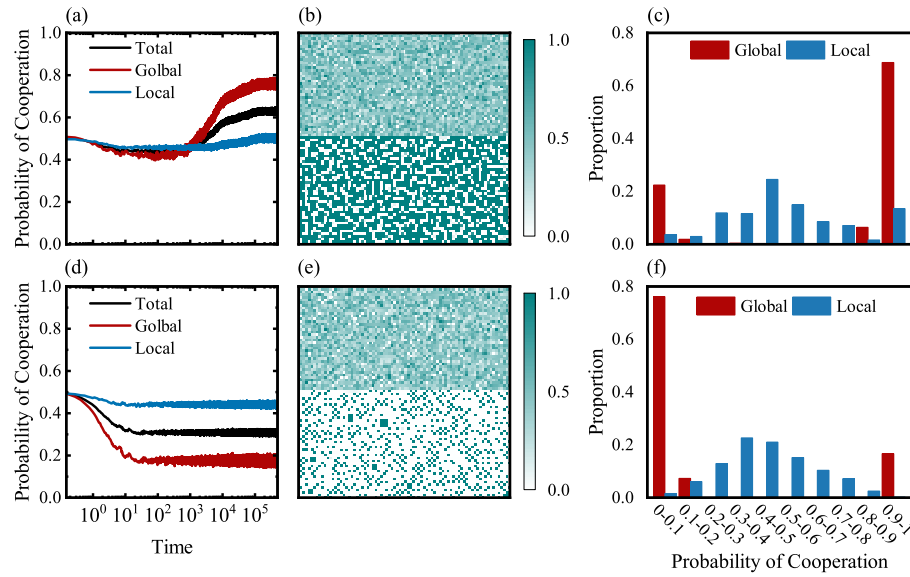
In the beginning, each global player is randomly assigned the intended probability to cooperate, and each local player is randomly assigned the vector of intended probability to cooperate, one intended probability to cooperate for each of the neighbors.

Players were selected once on average to update their intended probability to cooperate in each time step. For a full reinforcement learning run, we observed the probability of cooperation on the lattice with size  $L = 100$  over 800 000 time steps, of which the last 10 000 has up to a stable state.

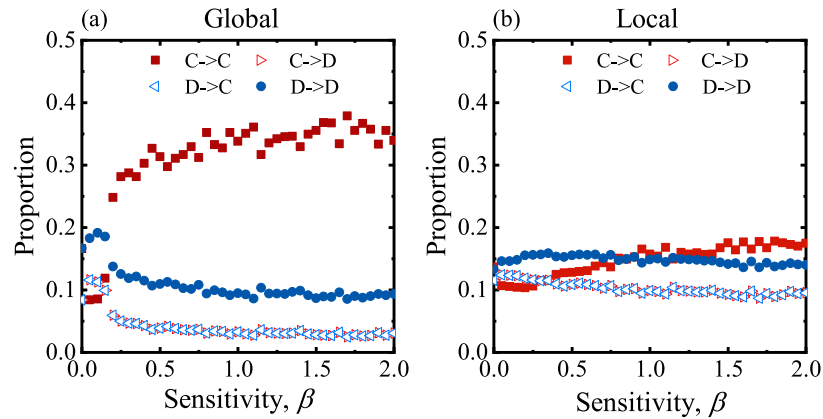
### 3. Results

To explore the cooperative behavior of two species of players with different sources of stimulus under reinforcement learning rules, we focused on how sensitivity  $\beta$  and the temptation to defect  $b$  affect the cooperative behavior (figure 2). The results show that the difference between payoff and aspiration brings player stimulus, and the player's greater sensitivity to stimulus increases the tendency to cooperate. However, the effect of sensitivity is limited, and when  $\beta$  exceeds a certain threshold, especially in the case where  $\beta > 1$ , cooperative behavior does not continue to increase with  $\beta$ . In addition to this, cooperative behavior gradually decreases with the temptation to defect  $b$  (figure 2(a)). However, the responses of the two species of players to changes in  $b$  are quite different. Global players' cooperation gradually decreases with  $b$  from high to low levels (figure 2(b)), whereas local players are more tenacious and the effect of changes in  $\beta$  and  $b$  on their cooperation behavior is minimal, with their cooperative behavior remaining stable in the range of 0.2 to 0.3 (figure 2(c)).

The dynamic process of the players' cooperation probability and its distribution in the steady state for different temptation to defect  $b$  is shown in figure 3. The results show that the evolutionary trend of global player's cooperation probability is determined by the temptation to defect. In contrast, local players' cooperation probabilities do not fluctuate significantly with external factors, either in terms of dynamic processes or changes in  $b$  values (figures 3(a) and (d)). Thus, the trend in the probability of cooperation of the global players determines the trend of the group. In order to clearly show the distribution of the cooperation probabilities of the two species of players in a steady state, local players and global players are fixed in the upper and lower parts of the grid network, respectively, and the initial cooperation probabilities of all players are given randomly. In particular, it is confirmed that fixing players' position as shown above does not affect individual decision-making. Furthermore, the results show that in a steady state, the cooperation probability of local players presents a chaotic state, while the cooperation probability of global players clearly shows two separate cooperation levels, namely high cooperation and low cooperation (figures 3(b) and (e)). With a small  $b$  value, the high cooperation is dominant (figure 3(b)), while at a large  $b$  value, the low cooperation is dominant (figure 3(e)). Further, the histograms of the probability of cooperation for the two types of players are given, identifying from a quantitative perspective that the



**Figure 3.** The distribution of individual cooperation probability of the two species of players shows great differences. The individual cooperation probability of global players is polarized (close to 1 or close to 0), while the individual cooperation probability of local players approximately obey a normal distribution with 0.5 as the center. (a) and (d) Time evolution of cooperation probability for total population, among the sub-population of global players, and the sub-population of local players. (b) and (e) Snapshots of the distribution of individual cooperation probability in steady state. Global players and local players are placed in the lower and upper halves of the grid, respectively. (c) and (f) The distribution of individual cooperation probability for two species of players. In panels (a)–(c),  $b = 1.2$ , and panels (d)–(f),  $b = 2.0$ , all results are obtained for  $A = 0.5$ ,  $\beta = 1.4$ ,  $u = 0.5$ .



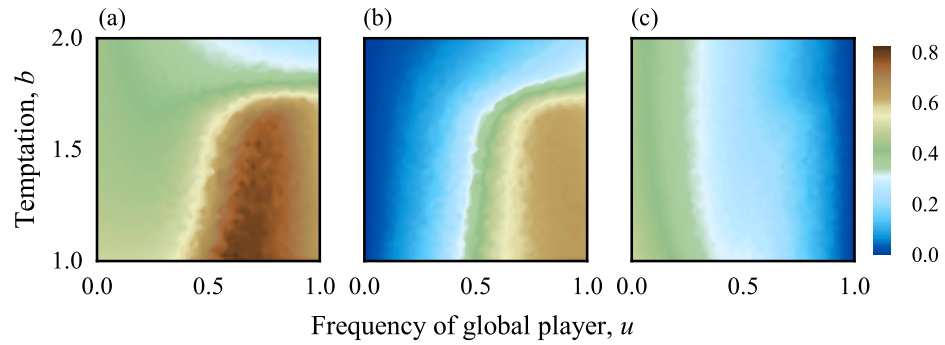
**Figure 4.** The transition probability of strategy for global players and local players as a function of the sensitivity  $\beta$ . Sensitivity  $\beta$  can enhance individual cooperative behavior. (a) The global player's cooperative strategy is significantly reinforcement as  $\beta$  increases. (b) Sensitivity has few effect on local players in strengthening cooperative strategy. All results are obtained for  $A = 0.5$ ,  $b = 1.2$ ,  $u = 0.5$ .

cooperation probability of global players has a two-level distribution, while the local player's probability of cooperation approximately follows a normal distribution with a mean of 0.5 (figures 3(c) and (f)).

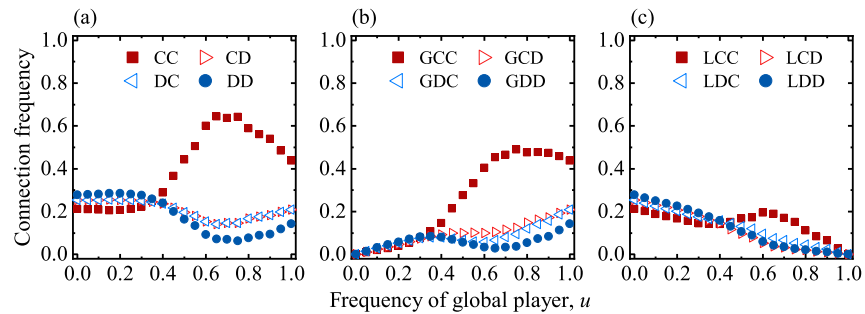
Given the large differences between global and local players, we explored the change in the probability of strategy shift with sensitivity  $\beta$  for the two types of players separately under the reinforcement learning rule (figure 4). The proportion of  $C \rightarrow C$  gradually increases, implying that global players' cooperative behavior is reinforced. The decline in the proportion of  $D \rightarrow D$  means that the defection strategy is less likely to be repeated, and the proportion of cooperation and defection substituted for each other is always consistent and small. Thus in the case of small  $b$ , increasing  $\beta$  promotes global players to gradually converge towards cooperation and avoid defection, driven by reinforcement learning rules (figure 4(a)). The trend in strategy shifts for local players is similar to that of global players, but with small fluctuations (figure 4(b)). Thus larger sensitivities  $\beta$  are more likely to motivate cooperative behavior in global players.

Then we investigated the influence of the proportion of global players in the network on cooperative behavior (figure 5). The results show that the appropriate mixing ratio of the two types of players in the





**Figure 5.** Moderating the frequency of global players in a group can significantly increase the level of cooperation. The probability to cooperate in the steady state as a function of the density of global players  $u$ , and the temptation to defect  $b$ . Panels (a)–(c) respectively show the cooperative behavior in three situations: the total population, among the sub-population of global players, and the sub-population of local players. (a) Population reaches a high level of cooperation when the proportion of global players is about 70%. (b) The cooperation probability of global players gradually strengthens as their density increases. (c) The cooperation probability of local player gradually weakens as the density of global players increases. All results are obtained for  $A = 0.5$ ,  $\beta = 1.4$ .

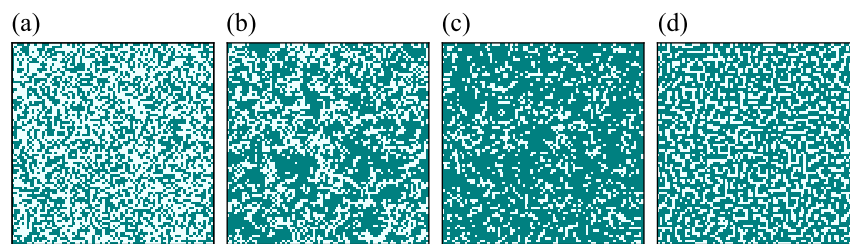


**Figure 6.** The density of global players determines the predominance of cooperation. Panels (a)–(c) respectively give the frequency of connection as a function of the proportion of global players,  $u$ , in the following three cases: pairwise interactions without regard to the focal player, pairwise interactions centered on the global player, and pairwise interactions centered on the local player. (a) Pairwise interactions in the total population between pairs of cooperating (CC) and pairs of defecting (DD) agents, as well as between cooperating–defecting pairs (CD) and vice versa (DC), clearly shows that CC interactions are dominant when the percentage of global players is over 40%. (b) CC interactions that centered on the global players (GCC) are the primary source of group cooperation. (c) The CC interactions that centered on the local players (LCC) do not continue to deteriorate with density of global players, but rise briefly. Here we take  $A = 0.5$ ,  $\beta = 1.4$ ,  $b = 1.2$ .

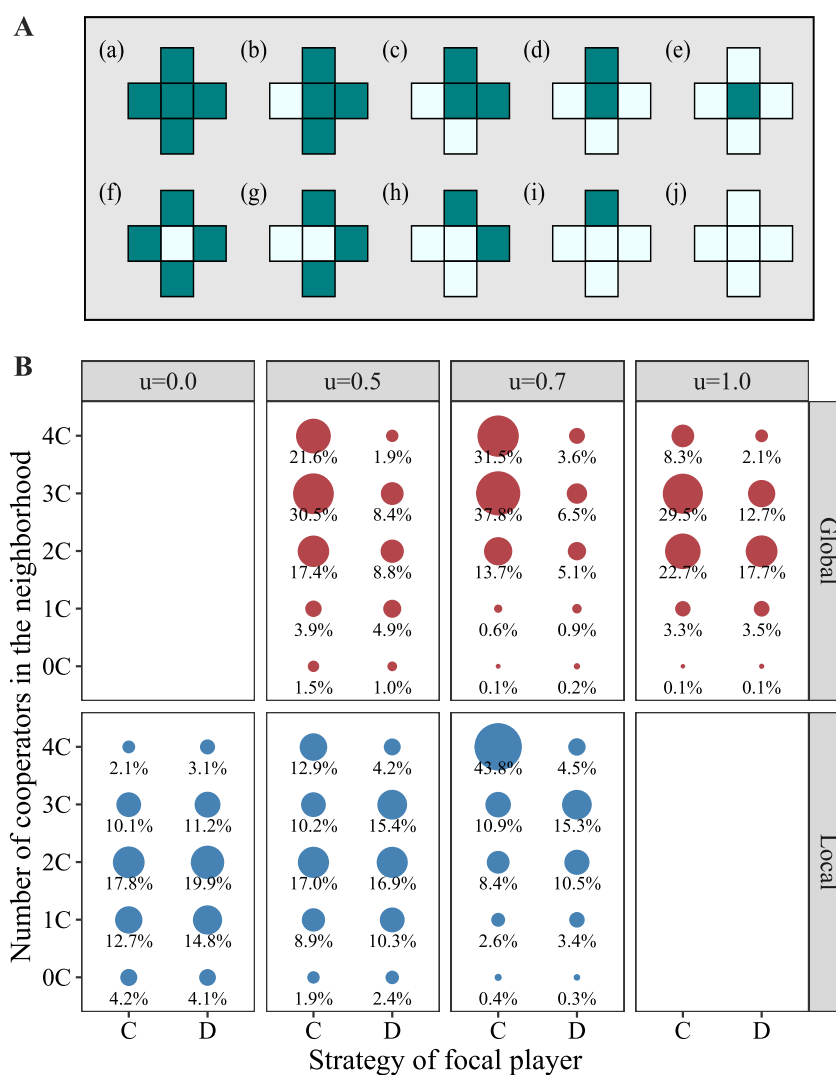
network can enable the group to achieve high cooperation (figure 5(a)). The cooperative behavior of global players increases with the value of  $u$  (figure 5(b)), while the opposite is true for local players (figure 5(c)). It is worth noting that despite the global player's cooperative behavior is dominant, group cooperation is not the best situation when the network is full of global players, but rather the network reaches its highest level of cooperation when the proportion of global players is around 0.65 (figure 5(a)).

Pairwise interactions of strategies at steady state are analyzed. The results show that when  $u$  is smaller than 0.4, overall pairwise interactions are not significantly different from each other (figure 6(a)), while pairwise interactions starting with global players gradually increase (figure 6(b)) and those starting with local players gradually decrease (figure 6(c)). As the proportion of global players increases further, CC interactions explode rapidly (figure 6(a)), especially for global players (figure 6(b)). At the same time, we are surprised to find a reversal of the decreasing trend in CC interactions for local players, achieving a brief increase (figure 6(c)). When the proportion of global players exceeds 0.7, the CC interaction gradually declines, but it always prevails. Therefore, when the density of global player is 0.7, group cooperation reaches the highest level.

In order to more intuitively observe the results of strategy evolution under reinforcement learning rules, a snapshot of the distribution of strategies in steady state is given for different  $u$  (figure 7). When the proportion of local players is large, the cooperation strategies hardly form clusters and they are distributed in scattered dots or bands (figure 7(a)). When the number of global players gradually increases, cooperative clusters are formed in the network, especially when  $u = 0.7$  (figure 7(c)). However, in the case of all global players, the cooperation strategy does not form larger cooperative clusters as expected, but instead shows a maze distribution (figure 7(d)).

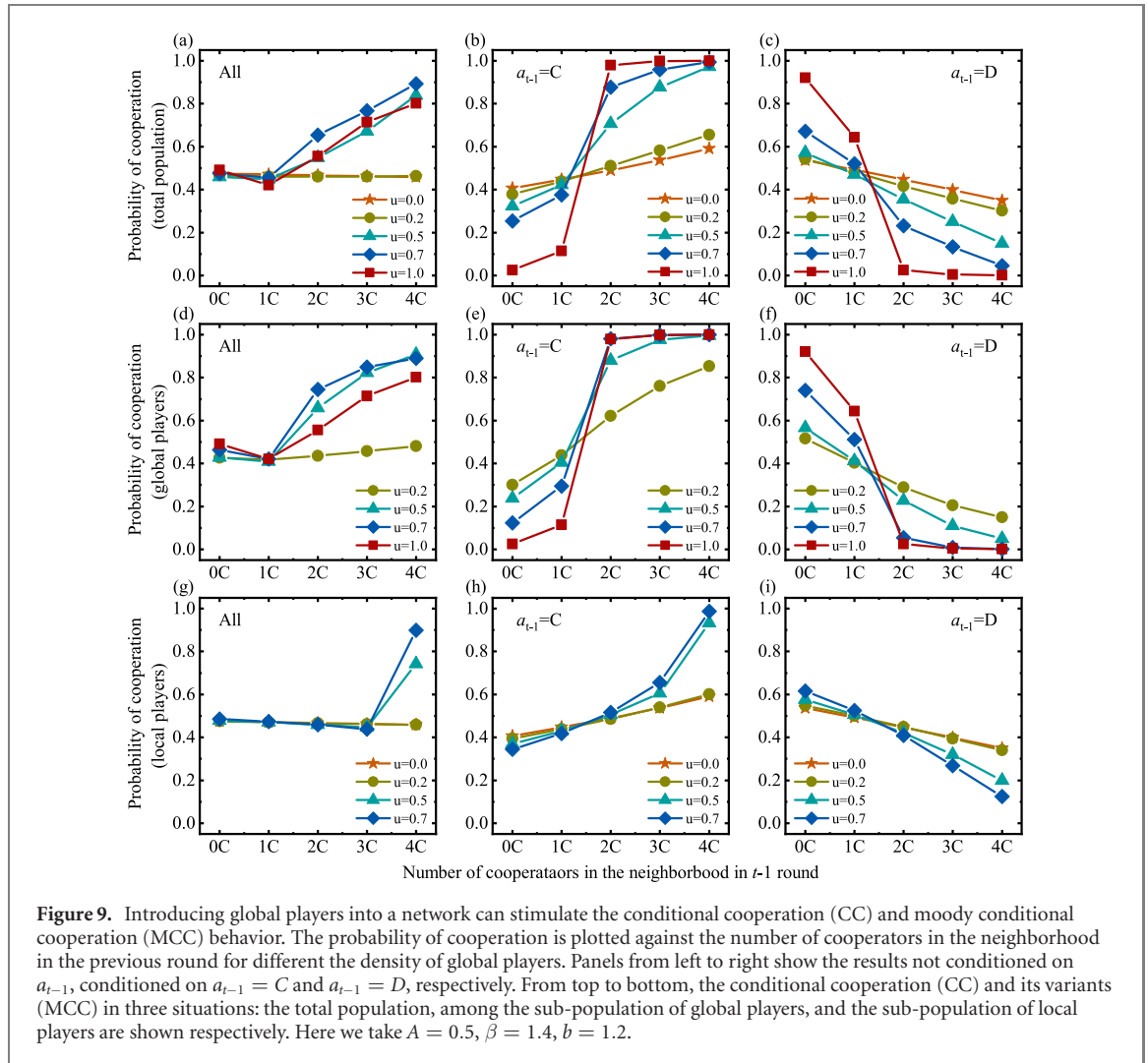


**Figure 7.** The mixed groups of two species of players are conducive to form cooperative clusters. Snapshots of the distribution of strategies in steady state for different density of global players, in panels (a)–(d),  $u$  is equal to 0, 0.5, 0.7, and 1.0, respectively. Dark cyan indicates cooperation, and light cyan indicates defection. All results are obtained for  $A = 0.5$ ,  $\beta = 1.4$ ,  $b = 1.2$ .



**Figure 8.** Panel (A) the basic structure of the individual neighborhood. Dark cyan indicates cooperation, and light cyan indicates defection. Panel (B) the distribution of strategies of focal players under the number of cooperators in the neighborhood for different density of global players. Here we take  $A = 0.5$ ,  $\beta = 1.4$ ,  $b = 1.2$ .

Then we analyzed the distribution of the basic structures of the two players forming the clusters (figure 8). It shows that when there are all local players, the neighborhoods where players are located are mainly  $c$ ,  $d$ ,  $h$ ,  $i$  basic structures, so that there are hardly any cooperative clusters of large size. With the introduction of global players, the proportion of basic structures  $a$ ,  $b$  increases significantly, especially the proportion of local players with basic structure  $a$  is as high as 43.8%, which provides the necessary conditions for the network to form larger cooperative clusters. However, when  $u = 1$ , the basic structure  $a$ ,  $b$  significantly decreases, and the basic structure  $c$ ,  $g$ ,  $h$  increases, resulting in a maze distribution of

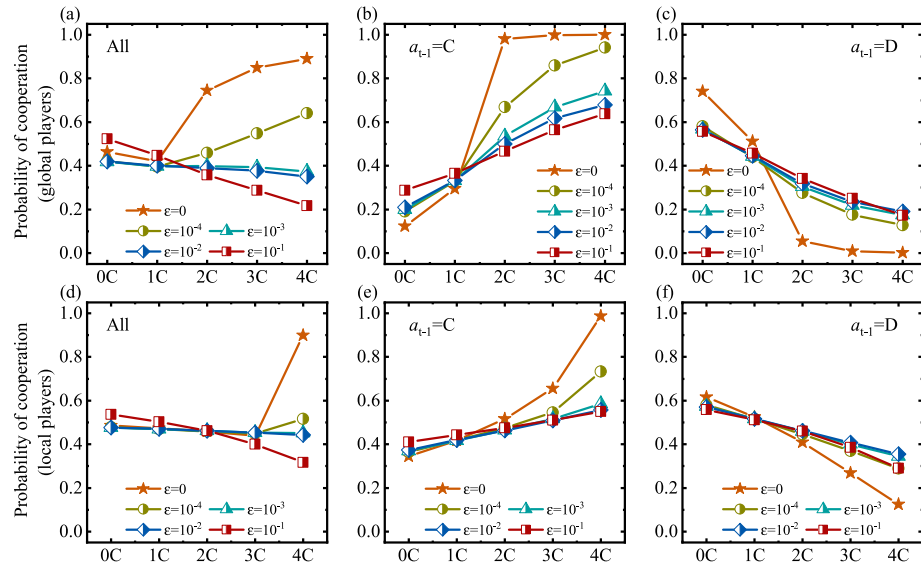


strategies. Therefore, although global players play a leading role in promoting cooperation, only cooperation with local players can enable the group to achieve a high level of cooperation.

Previous researches have shown that reinforcement learning reflects features of direct reciprocity, so we analyze the conditional cooperation and moody conditional cooperation [51, 56] with different proportions of global players (figure 9). It proves that pure local players ( $u = 0$ ) do not exhibit significant conditional cooperation (figures 9(a) and (g)), but show the characteristics of moody conditional cooperation (figures 9(b), (c), (h) and (i)). The existence of global players can trigger individual conditional cooperation behavior (figure 9(a)). However, the conditional cooperation patterns of the two types of players show great differences, that is, global players' cooperative tendency increase with the number of cooperators in their neighborhood (figure 9(d)), and local players' cooperative tendency only rises significantly when their neighbors are all cooperators (figure 9(g)). In addition, regardless of the mixed ratio of the two types of players in the population, players show moody conditional cooperation, that is, individuals who chose to cooperate in the previous round are more likely to cooperate, and vice versa. The more global players in the population, the more obvious the tendency for emotional conditional cooperation (figures 9(b) and (c)). Furthermore, results once again confirmed the leading role of global players, that is, the trend of conditional cooperation and moody conditional cooperation of global players determines the trend of the group.

Finally, in repeated prisoner's dilemma game, considering players may implement wrong decisions during the interactions [57], we give the impact of noise (errors),  $\varepsilon$ , on the outcome of the interactions (figure 10). Thus, the actual probability to cooperate at round  $t + 1$  is measured by  $\tilde{p}_{t+1} \equiv p_{t+1}(1 - \varepsilon) + (1 - p_{t+1})\varepsilon$ . Results show that noise influences cooperative behavior. In the case of high noise, the tendency of conditional cooperation changes (figures 10(a) and (d)), but both types of players still present stable moody conditional cooperation (figures 10(b), (c), (e) and (f)).





**Figure 10.** Both types of players show properties of ‘moody cooperators’, despite noise interference with cooperative behavior. The probability of cooperation is plotted against the number of cooperators in the neighborhood in the previous round for different error rates,  $\epsilon$ . Panels from left to right show the results not conditioned on  $a_{t-1}$ , conditioned on  $a_{t-1} = C$  and  $a_{t-1} = D$ , respectively. The conditional cooperation (CC) and its variants (MCC) for global players (top) and local players (bottom) are shown respectively. Here we take  $A = 0.5$ ,  $\beta = 1.4$ ,  $b = 1.2$ ,  $u = 0.7$ .

#### 4. Conclusion

Under the BM model of reinforcement learning, we classified players into two types, global players and local players, depending on the source of stimulus they perceive. By changing the mixed ratio of the two types of players in the group to study how players with different stimulus affect individual cooperative behavior. How players with different stimulus influence individual cooperative behavior was investigated by varying the mixing ratio of the two types of players in the group. Research shows that global players play a dominant role in facilitating cooperation, and their behavior largely determines the trend of the entire group. But it does not mean that the network can achieve a high level of cooperation when all players are global players. The fact is that network reaches high cooperation when there is a low density of local players in the population. This is due to the significant differences in the reciprocity patterns of the two types of players.

Looking at the group as a whole, our results reconfirm previous research [51, 56] that conditional cooperation and moody conditional cooperation reveal behavior patterns that individuals generally follow in repeated dilemma games. Most importantly, we also find differences in direct reciprocity among individuals with different sources of stimulus. The probability of global player cooperation increases with the number of cooperators in the neighborhood, exhibiting conditional cooperation. In contrast, no conditional cooperation features are observed when the population is full of local players. The introduction of global players can stimulate individual conditional cooperation behavior, and the situation where neighbors are all collaborators can significantly increase the probability of local player cooperation. All individuals show the characteristics of moody conditional cooperation even in the presence of noise. In particular, global players are sensitive to the number of cooperators in the neighbors in the previous round, while local players seem to be more cautious or strict, as significant increases in the probability of cooperation of local players occurred when more than half of the neighbors are cooperators.

In addition, the distributions of the individual cooperation probabilities of the two types of players in the steady state show large differences, with the global player’s cooperation probability showing a two-level distribution (close to 0 or close to 1), while the local player’s cooperation probability approximately follows a normal distribution.

Our model is a simple variant of the BM model of reinforcement learning, but it has obtained interesting results, and provides new insights for readers to understand cooperative behavior in repeated dilemma games.

Adaptive behavior about adjusting strategies is considered to be highly cognitive and complex, and it is not difficult to understand the high costs involved in implementing learning strategies. Currently, cognitive costs have been analyzed in exploring the effects of trust-based strategies [58], intention recognition [59], evolutionary cycles in finite populations [60], finite automata [61], etc on the evolution of cooperation in

repeated prisoner's dilemma games. In future work we will focus on the impact of cognitive costs on the evolution of cooperation in the framework of reinforcement learning.

## Acknowledgment

We acknowledge support from National Natural Science Foundation for Distinguished Young Scholars (Grants No. 62025602), National Natural Science Foundation of China (Grants No. U1803263, 11931015, 81961138010), National Key R & D Program of China (2018YFB1403501), National Key R & D Program of China (2019YFB2102304), Fok Ying-Tong Education Foundation, China (171105), Key Technology Research and Development Program of Science and Technology-Scientific and Technological Innovation Team of Shaanxi Province (Grant No. 2020TD-013), Key Area R & D Program of Guangdong Province (No. 2019B010137004) to ZW. MP was supported by the Slovenian Research Agency (Grant Nos. P1-0403, J1-2457, and J1-9112). LS was supported by a key project (No. 11931015) of the National Natural Science Foundation of China (NNSFC) and NNSFC project No. 11671348.

## Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

## ORCID iDs

Matjaž Perc  <https://orcid.org/0000-0002-3087-541X>

## References

- [1] Lehmann L and Keller L 2006 The evolution of cooperation and altruism—a general framework and a classification of models *J. Evol. Biol.* **19** 1365–76
- [2] West S A, Griffin A S and Gardner A 2007 Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection *J. Evol. Biol.* **20** 415–32
- [3] Xia C, Gracia-Lázaro C, Moreno Y and Moreno Y 2020 Transition from reciprocal cooperation to persistent behaviour in social dilemmas at the end of adolescence *Chaos* **30** 063122
- [4] Gutiérrez-Roig M, Gracia-Lázaro C, Perelló J, Moreno Y and Sánchez A 2014 Effect of memory, intolerance, and second-order reputation on cooperation *Nat. Commun.* **5** 4362
- [5] Nowak M A and May R M 1992 Evolutionary games and spatial chaos *Nature* **359** 826–9
- [6] Ohtsuki H, Hauert C, Lieberman E and Nowak M A 2006 A simple rule for the evolution of cooperation on graphs and social networks *Nature* **441** 502–5
- [7] Szabó G and Fáth G 2007 Evolutionary games on graphs *Phys. Rep.* **446** 97–216
- [8] Szolnoki A, Perc M and Szabó G 2009 Phase diagrams for three-strategy evolutionary prisoner's dilemma games on regular graphs *Phys. Rev. E* **80** 056104
- [9] Rand D G and Nowak M A 2011 The evolution of antisocial punishment in optional public goods games *Nat. Commun.* **2** 434
- [10] Szolnoki A, Szabó G and Perc M 2011 Phase diagrams for the spatial public goods game with pool punishment *Phys. Rev. E* **83** 036101
- [11] Lee S, Holme P and Wu Z X 2011 Emergent Hierarchical structures in multiadaptive games *Phys. Rev. Lett.* **106** 028702
- [12] Szolnoki A, Szabó G and Perc M 2012 Self-organization of punishment in structured populations *New J. Phys.* **14** 043013
- [13] Javarone M A 2016 Statistical physics of the spatial prisoner's dilemma with memory-aware agents *Eur. Phys. J. B* **89** 42
- [14] Cardoso F M, Gracia-Lázaro C and Moreno Y 2020 Dynamics of heuristics selection for cooperative behaviour *New J. Phys.* **22** 123037
- [15] Alvarez-Rodriguez U, Battiston F, De Arruda G F, Moreno Y, Perc M and Latora V 2021 Evolutionary dynamics of higher-order interactions in social networks *Nat. Hum. Behav.* **5** 586–95
- [16] Fu F, Tarnita C E, Christakis N A, Wang L, Rand D G and Nowak M A 2012 Evolution of in-group favoritism *Sci. Rep.* **2** 460
- [17] Li K, Cong R, Wu T and Wang L 2015 Social exclusion in finite populations *Phys. Rev. E* **91** 042810
- [18] Duh M, Gosak M, Slavinec M and Perc M 2019 Assortativity provides a narrow margin for enhanced cooperation on multilayer networks *New J. Phys.* **21** 123016
- [19] Li K, Wei Z and Cong R 2019 Sentiment contagion dilutes prisoner's dilemmas on social networks *Europhys. Lett.* **128** 38002
- [20] Amaral M A and Javarone M A 2020 Strategy equilibrium in dilemma games with off-diagonal payoff perturbations *Phys. Rev. E* **101** 062309
- [21] Amaral M A and Javarone M A 2020 Heterogeneity in evolutionary games: an analysis of the risk perception *Proc. R. Soc. A* **476** 20200116
- [22] Jia D, Wang X, Song Z, Romić I, Li X, Jusup M and Wang Z 2020 Evolutionary dynamics drives role specialization in a community of players *J. R. Soc. Interface* **17** 20200174
- [23] Guo H, Song Z, Geček S, Li X, Jusup M, Perc M, Moreno Y, Boccaletti S and Wang Z 2020 A novel route to cyclic dominance in voluntary social dilemmas *J. R. Soc. Interface* **17** 20190789
- [24] Binder K and Hermann D K 1988 *Monte Carlo Simulations in Statistical Physics* (Berlin: Springer)
- [25] Liggett T M 1985 *Interacting Particle Systems* (Berlin: Springer)
- [26] Schlag K H 1998 Why imitate, and if so, how? A bounded rational approach to multi-armed bandits *J. Econ. Theory* **78** 130–56

- [27] Schlag K H 1999 Which one should I imitate? *J. Math. Econ.* **31** 493–522
- [28] Nowak M A, Bonhoeffer S and May R M 1994 Spatial games and the maintenance of cooperation *Proc. Natl Acad. Sci.* **91** 4877–81
- [29] Nowak M A and Sigmund K 2004 Evolutionary dynamics of biological games *Science* **303** 793–9
- [30] Artiges E, Gracia-Lazaro C, Floria L M and Moreno Y 2019 Replicator population dynamics of group interactions: broken symmetry, thresholds for metastability, and macroscopic behavior *Phys. Rev. E* **100** 052307
- [31] Milinski M 1987 Tit for tat in sticklebacks and the evolution of cooperation *Nature* **325** 433–5
- [32] Nowak M A and Sigmund K 2004 Tit for tat in heterogeneous populations *Nature* **355** 250–3
- [33] Santos F P, Santos F C and Pacheco J M 2018 Social norm complexity and past reputations in the evolution of cooperation *Nature* **555** 242–5
- [34] Amaral M A, Wardil L, Perc M and Silva J 2016 Stochastic win-stay-lose-shift strategy with dynamic aspirations in evolutionary social dilemmas *Nature* **94** 032317
- [35] Deng X, Zhang Z, Deng Y, Liu Q and Chang S 2016 Self-adaptive win-stay-lose-shift reference selection mechanism promotes cooperation on a square lattice *Appl. Math. Comput.* **284** 322–31
- [36] Axelrod R 1984 *The Evolution of Cooperation* (New York: Basic Books)
- [37] Kraines D and Kraines V 1993 Learning to cooperate with Pavlov an adaptive strategy for the iterated prisoner's dilemma with noise *Theory Decis* **35** 107–50
- [38] Nowak M and Sigmund K 1993 A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game *Nature* **364** 56–8
- [39] Hilbe C, Chatterjee K and Nowak M A 2018 Partners and rivals in direct reciprocity *Nat. Hum. Behav.* **2** 469–77
- [40] Wang Z, Jusup M, Shi L, Lee J-H, Iwasa Y and Boccaletti S 2018 Exploiting a cognitive bias promotes cooperation in social dilemma experiments *Nat. Commun.* **9** 2954
- [41] Jusup M, Maciel-Cardoso F, Gracia-Lazaro C, Liu C, Wang Z and Moreno Y 2020 Behavioural patterns behind the demise of the commons across different cultures *R. Soc. Open Sci.* **7** 201026
- [42] Buoni L, Babuška R and Schutter B D 2010 Multi-agent reinforcement learning: an overview *Innovations in Multi-Agent Systems and Applications-1* vol 310 (Berlin: Springer) pp 183–221
- [43] Devaine M, Hollard G and Daunizeau J 2014 Theory of mind: did evolution fool us? *PLoS One* **9** e87619
- [44] Han T A, Santos F C, Lenaerts T and Pereira L M 2015 Synergy between intention recognition and commitments in cooperation dilemmas *Sci. Rep.* **5** 9312
- [45] Han The Anh T A, Moniz Pereira L and Santos F C 2011 Intention recognition promotes the emergence of cooperation *Adapt. Behav.* **19** 264–79
- [46] McNally L, Brown S P and Jackson A L 2012 Cooperation and the evolution of intelligence *Proc. R. Soc. B.* **279** 3027–34
- [47] Pereira L M, Lenaerts T, Martinez-Vaquero L A and Han T A 2017 Social manifestation of guilt leads to stable cooperation in multi-agent systems *The 16th Int. Conf. Autonomous Agents and Multiagent Systems (AAMAS)* pp 1422–30
- [48] de Melo C M, Terada K and Santos F C 2021 Emotion expressions shape human social norms and reputations *Iscience* **24** 102141
- [49] Macy M W and Flache A 2002 Learning dynamics in social dilemmas *Proc. Natl Acad. Sci.* **99** 7229–36
- [50] Bush R R and Mosteller F 1955 *Stochastic Models for Learning* (New York: Wiley)
- [51] Ezaki T, Horita Y, Takezawa M and Masuda N 2016 Reinforcement learning explains conditional cooperation and its moody cousin *PLoS Comput. Biol.* **12** e1005034
- [52] Macy M W 1991 Learning to cooperate: stochastic and tacit collusion in social exchange *Am. J. Sociol.* **97** 808–43
- [53] Izquierdo L R, Izquierdo S S, Gotts N M and Polhill J G 2007 Transient and asymptotic dynamics of reinforcement learning in games *Games Econ. Behav.* **61** 259–76
- [54] Izquierdo S S, Izquierdo L R and Gotts N M 2008 Reinforcement learning dynamics in social dilemmas *J. Artif. Soc. Soc. Simul.* **11** 1
- [55] Masuda N and Nakamura M 2011 Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated prisoner's dilemma *J. Theor. Biol.* **278** 55–62
- [56] Horita Y, Takezawa M, Inukai K, Kita T and Masuda N 2017 Reinforcement learning accounts for moody conditional cooperation behavior: experimental results *Sci. Rep.* **7** 39275
- [57] Sigmund K 2010 *The Calculus of Selfishness* (Princeton, NJ: Princeton University Press)
- [58] Han T A, Perret C and Powers S T 2021 When to (or not to) trust intelligent machines: insights from an evolutionary game theory analysis of trust in repeated games *Cogn. Syst. Res.* **68** 111–24
- [59] Han T A, Pereira L M and Santos F C 2012 Corpus-based intention recognition in cooperation dilemmas *Artificial Life* **18** 365–83
- [60] Imhof L A, Fudenberg D and Nowak M A 2005 Evolutionary cycles of cooperation and defection *Proc. Natl Acad. Sci.* **102** 10797–800
- [61] Ho T-H 1996 Finite automata play repeated prisoner's dilemma with information processing costs *J. Econ. Dyn. Control* **20** 173–207