





# Memory-driven Q-learning model for cooperation in snowdrift game with dynamic behavioral types

Xiang Li <sup>a, , 1</sup>, Bin Pi <sup>b, , 1</sup>, Liang-Jian Deng <sup>b, , 1</sup>, Qin Li <sup>c, , \*</sup>

<sup>a</sup> Yingcai Honors College, University of Electronic Science and Technology of China, Chengdu, 611731, China

<sup>b</sup> School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu, 611731, China

<sup>c</sup> Business College, Southwest University, Chongqing, 402460, China

## ARTICLE INFO

### Keywords:

Evolutionary game theory  
Q-learning  
Multi-agent decision-making  
Cooperative optimization  
Population dynamics

## ABSTRACT

Evolutionary games on complex networks provide an effective framework for studying the emergence of cooperation, where Q-learning, a mechanism that enables agents to make strategy decisions based on experiential feedback, has garnered significant attention in this field. Previous studies predominantly rely on the assumption that individual behavioral types remain static over time. However, during prolonged game interactions, agents struggle to maintain fixed behavioral types. For instance, a profiteer might transition to a conformist based on accumulated experience. Therefore, we propose a Q-learning-based dynamic behavioral type adaptation model, specifically defining the action space of Q-learning as a selection between profiteer and conformist types. Furthermore, we integrate a memory mechanism into Q-learning reward calculations to reflect agents' trade-offs between historical experiences and immediate payoffs. Experimental results demonstrate that evolutionary games ultimately reach either full cooperation or full defection states depending on the cost-to-benefit ratio in the snowdrift game. By analyzing the evolution of four agent types, including cooperative profiteers, cooperative conformists, defective profiteers, and defective conformists, we delineate the phases of these evolutionary trajectories and explain the underlying mechanisms through our newly proposed feedback-driven  $\Delta Q$ . Additionally, by investigating memory-driven Q-learning itself, we illustrate its advantage in promoting cooperation among other mechanisms and reveal that prioritizing immediate rewards significantly enhances cooperation. Our study pioneers the integration of dynamic behavioral type adaptation with memory-driven Q-learning, offering an innovative model for cooperative optimization in real-world multi-agent systems.

## 1. Introduction

Game theory, as a scientific discipline that investigates how rational individuals make optimal decisions during interactions [1], provides a tripartite research framework encompassing player, strategy set, and payoff [2]. Within this framework, the emergence of cooperation has been extensively studied. Researchers have proposed numerous game models including the prisoner's dilemma (PD) [3–5], snowdrift game (SDG) [6–8], stag hunt game (SHG) [9,10], and harmony game (HG), serving as classic contexts for cooperation research. They are varied by distinct payoff matrices.

\* Corresponding author.

E-mail address: [qinli1022@gmail.com](mailto:qinli1022@gmail.com) (Q. Li).

<sup>1</sup> Xiang Li and Bin Pi have equal contributions.

<https://doi.org/10.1016/j.apm.2025.116313>

Received 4 April 2025; Received in revised form 20 June 2025; Accepted 14 July 2025

Available online 18 July 2025

0307-904X/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Early game theory research assumed fully rational participants [11], with pairwise interactions converging to a Nash equilibrium specific to each model [12]. This approach, however, often results in social dilemmas, where individual rationality undermines collective welfare. To address these limitations, networked evolutionary games adopt a population-based perspective, modeling individuals who interact repeatedly with multiple agents and accumulate payoffs over time. This framework shifts the analytical focus from static Nash equilibrium to dynamic processes governed by natural selection. Complex networks typically serve as platforms for such studies, with nodes representing agents and edges denoting interactions [13]. Common network structures, including square lattice network [14,15], small-world network [16–18], and scale-free network [19–21] have been shown to effectively reflect real-world interactions [22,23]. These topologies provide robust platforms for investigating cooperation dynamics in evolutionary game theory.

The general process of networked evolutionary games proceeds as follows: at each time step, agents accumulate payoffs through interactions with their neighbors based on the payoff matrix and subsequently update their game strategies according to specified update rules. Researchers have explored many realistic behavioral mechanisms into the strategy updating and evolution process to enhance cooperation, including memory effects [24–26], reputation systems [27,28], compassion-driven interactions [29], and moral preferences [30]. These studies typically assume a uniform and fixed behavioral type for all agents. However, behavioral heterogeneity is prevalent in real-world social systems. Szolnoki and Perc made a significant contribution by investigating the dynamics between two distinct behavioral types, conformists and profiteers, demonstrating that conformity strengthens network reciprocity [31]. Although these models provide valuable insights, research on dynamic role-switching between conformist and profiteer behaviors remains limited. While previous work has modeled such kind of transitions using Markov processes [32], it neglects agents' behavioral preferences for utility maximization in their decision-making.

To bridge this gap, we propose a Q-learning framework to guide agents in selecting their behavioral types. Each agent maintains a unique Q-table, shaped throughout their interactions, which captures experience-based trade-offs in choosing between conformist and profiteer types. Notably, Q-learning is commonly employed in evolutionary games to guide strategy selection, with action spaces typically defined as cooperation and defection, while state spaces are defined as the number of neighboring cooperators [33,34]. There are also studies that redefine the action space to model social engagement, such as by incorporating strategic abstention in dynamic environments [35].

In this study, we redefine the Q-learning action space to encompass profiteer and conformist roles rather than traditional cooperation or defection strategy. At each game iteration, agents first determine their strategy using role-specific update rules and then select their behavioral types for the next round through an  $\epsilon$ -greedy exploration mechanism within the Q-learning framework [36]. Moreover, unlike conventional Q-learning, which relies on immediate payoffs as rewards, our model incorporates memory effects by including payoffs from previous time steps in the reward calculation, reflecting agents' ability to recall past interactions. This is distinct from the previous work that introduced memory effects into individuals' payoff calculation [26,32].

The main contributions of our paper are summarized as follows:

- *We pioneer the application of Q-learning as a basis for individual behavioral type transitions.* By redefining the action space of traditional Q-learning from game strategies to behavioral types, we capture the real-world phenomenon where individuals weigh prior experiences when adopting different behavioral types. Through experiments, we demonstrate the superiority of this mechanism in promoting cooperation.
- *We classify and provide a detailed, phased explanation of the evolutionary processes under the proposed mechanism.* Our findings reveal that most evolutionary trajectories converge to either a full cooperation (FC) or full defection (FD) state. We introduce a feedback-based  $\Delta Q$  to interpret these dynamics and identify that the emergence of cooperative conformist agents in the final stage of FC evolution results from stochastic exploration, a finding applicable across various network structures.
- *We incorporate memory effects into the reward calculation of Q-learning.* Our experiments further establish the critical importance of prioritizing immediate rewards over future or historical payoffs in achieving rapid cooperation, which can provide a valuable perspective for cooperative optimization.

The remainder of this study is organized as follows. In Section 2, we introduce the game model and payoff calculation mechanism, outline the strategy update rules for profiteers and conformists, and elaborate on how agents achieve behavioral type adaptation through Q-learning. Section 3 showcases our experimental results, where we sequentially investigate the evolution of cooperation, the evolution of agent types, and the properties of memory-driven Q-learning. We reveal the advantages of reinforcement learning among other typical mechanisms in promoting cooperation. In addition, we verify our theoretical analysis on various network structures. In Section 4, we summarize the conclusion, significance, and prospects of this paper.

## 2. Model

In this section, we first present the game model and payoff calculation mechanism underlying our model. Subsequently, we introduce two distinct strategy update rules for different types of agents: profiteers employing the Fermi rule and conformists following the majority of neighbors. Then, we elaborate on the Q-learning-based type transition mechanism where agents dynamically switch between profiteers and conformists, explicitly explaining how individual memory factors are incorporated into the reward design of the Q-learning model.

### 2.1. Game model and payoff calculation

The two-agent two-strategy game model is generally defined by  $R$  (reward, the payoff for cooperator when two cooperators interact),  $S$  (sucker's payoff, the payoff for cooperator when a cooperator interacts with a defector),  $T$  (temptation, the payoff for defector when a defector interacts with a cooperator), and  $P$  (punishment, the payoff for defector when two defectors interact) as:

$$\begin{matrix} & C & D \\ \begin{matrix} C \\ D \end{matrix} & \begin{pmatrix} R & S \\ T & P \end{pmatrix} \end{matrix}, \quad (1)$$

where  $C$  and  $D$  denote cooperation and defection, respectively. Among all the models, SDG is widely studied, characterized by the payoff relationship  $T > R > S > P$  [37–39]. In this paper, we simplify the setting by assigning the reward ( $R$ ) to 1 for mutual cooperation and the punishment ( $P$ ) to 0 for mutual defection. The corresponding payoff matrix  $A$  is given by:

$$A = \begin{pmatrix} 1 & 1-r \\ 1+r & 0 \end{pmatrix}, \quad (2)$$

where  $0 < r < 1$  is a variable parameter indicating the cost-to-benefit ratio.

Within every discrete time step of the evolutionary game dynamics, each agent synchronously engages in pairwise interactions with all neighboring agents to aggregate their total payoff. Therefore, the payoff obtained by agent  $i$  at time step  $t$  can be expressed as:

$$\Pi_i^t = \sum_{j \in \Omega_i} \xi_j^T A \xi_j, \quad (3)$$

where  $\xi_x = [1, 0]^T$  denotes the cooperative strategy while  $\xi_x = [0, 1]^T$  represents the defective strategy by agent  $x$ .  $\Omega_i$  is the neighbor set of agent  $i$ .

### 2.2. Strategy evolution

In this study, we investigate two types of agents: profiteers and conformists. Next, we will analyze how these two types of agents update their game strategies. For profiteers, they use the classical Fermi rule to update strategies. This is a widely used update rule [31,32,36]. In each round, profiteer  $i$  randomly selects a neighbor  $j$ , and the probability of adopting strategy  $\xi_j$  is:

$$P_{\text{profiteer}}(\xi_i \leftarrow \xi_j) = \frac{1}{1 + \exp \left[ \left( \Pi_i^t - \Pi_j^t \right) / \kappa \right]}. \quad (4)$$

Here,  $\kappa > 0$  represents the noise coefficient, which models the degree of irrationality in an agent's decision-making. Consider a scenario where the payoff of neighbor  $j$  exceeds that of profiteer  $i$ . As  $\kappa \rightarrow 0$ , agent  $i$  behaves fully rationally, adopting  $j$ 's strategy. Conversely, as  $\kappa \rightarrow \infty$ , agent  $i$  disregards the payoff advantage of agent  $j$  relative to itself and instead chooses whether to adopt  $j$ 's strategy with equal probability.

Whereas for conformists, their strategy selection prioritizes majority preferences within their local neighborhood. Let  $N_i^{\xi_j}$  denote the number of neighbors adopting strategy  $\xi_j$  in conformist  $i$ 's neighbor set  $\Omega_i$ , and  $k_h = |\Omega_i|/2$  represents half of the agent's degree. The probability of conformist  $i$  adopting strategy  $\xi_j$  is expressed as:

$$P_{\text{conformist}}(\xi_i \leftarrow \xi_j) = \frac{1}{1 + \exp \left[ \left( k_h - N_i^{\xi_j} \right) / \kappa \right]}. \quad (5)$$

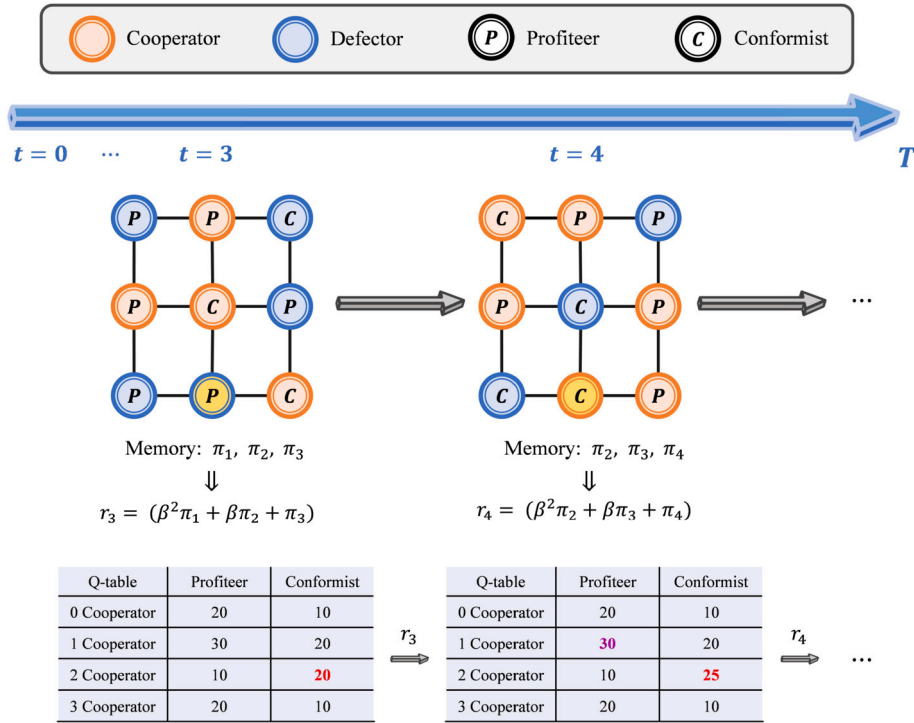
The strategy adopted by conformists in the subsequent round is independent of their own current strategy; they tend to choose strategies that account for a higher proportion among their neighbors. For example, if more than half of conformist  $i$ 's neighbors adopt strategy  $\xi_j$ , such that  $k_h < N_i^{\xi_j}/2$ , the probability that  $i$  adopts  $\xi_j$  will exceed  $1/2$ .

### 2.3. Memory-driven Q-learning for behavioral type adaptation

In real life, individuals do not always maintain a single spirit state, but rather choose behavioral tendencies based on their personal experiences. Q-learning serves as an effective tool to embody this adaptive decision-making process. In our model, agents' behavioral types are not fixed but dynamically switch between profiteers and conformists through the Q-learning mechanism.

In the evolution process, each agent maintains a two-dimensional Q-table. The state space is defined as  $S = \{0, 1, 2, \dots, |N_i^C|\}$ , with  $|N_i^C|$  representing the number of cooperators in the agent's neighbor set  $\Omega_i$ , and the action space is  $\mathcal{A} = \{\text{profiteer}, \text{conformist}\}$ . At each time step, agents update their Q-tables according to the following rule:

$$Q_{t+1}(s, a) = (1 - \alpha)Q_t(s, a) + \alpha[r_t(s, a) + \gamma \max_{a' \in \mathcal{A}} Q_t(s', a')], \quad (6)$$



**Fig. 1. An illustration of the model.** The schematic illustrates the evolutionary game process between profiteers and conformists within a  $3 \times 3$  lattice network across two time steps ( $t = 3$  and  $t = 4$ ), highlighting how the yellow-marked agent adapts its type and selects strategies based on the memory-driven Q-learning model.

where  $s \in S$  and  $a \in \mathcal{A}$  represent the current state and action, respectively, while  $s' \in S$  denotes the subsequent state after taking action  $a$ . Besides,  $\alpha$  quantifies the learning rate that controls the update speed of Q-values, and  $\gamma$  serves as the discount factor that balances short-term versus long-term rewards: smaller  $\gamma$  values incentivize agents to prioritize immediate rewards, whereas larger values encourage strategic focus on cumulative future gains.

Generally,  $r_t(s, a)$  is defined by the immediate payoff obtained after taking an action, which means  $r_t(s, a) = \Pi^t$ . However, since agents inevitably incorporate influences from historical interactions during decision-making, we redefine the Q-learning reward through a memory-weighted aggregation of past payoffs:

$$r_t = \sum_{i=0}^{M-1} \beta^i \Pi^{t-i}. \quad (7)$$

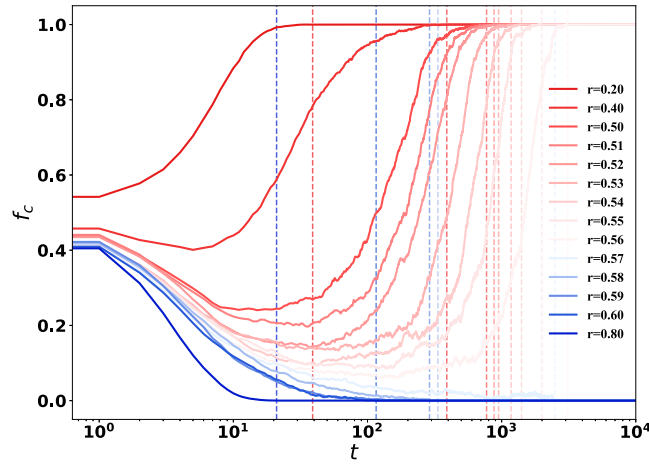
Here,  $M$  denotes the memory length capturing the temporal span of historical payoff, while  $0 \leq \beta \leq 1$  quantifies the memory strength. Larger  $M$  and  $\beta$  will result in a stronger memory effect. When  $t < M$ , we stipulate  $M = t$ .

Above, we establish a memory-driven Q-learning model for dynamic type adaptation. At each time step, agents adhere to an  $\epsilon$ -greedy exploration strategy: with probability  $\epsilon$ , they randomly explore behavioral types (profiteer or conformist), and with probability  $1 - \epsilon$ , they exploit the current optimal action by selecting the type (profiteer or conformist) that maximizes the Q-value for the observed state. The Q-table is then updated by Eq. (6) integrating the memory-augmented reward  $r_t$  through Eq. (7), thereby enabling agents to adaptively balance historical experiences with real-time environmental feedback.

Fig. 1 illustrates the evolutionary process of our model, in which we set the memory length  $M = 3$ . Focusing on the yellow-marked agent who acts as a defective profiteer and obtains a payoff  $\pi_3$  at time step  $t = 3$ . We first calculate the Q-learning reward using Eq. (7) based on the current payoff and the previous two rounds' payoffs, then apply the  $\epsilon$ -greedy exploration strategy (choosing exploitation in the diagram). Given the state of having two cooperators among the agent's neighbors, we select the conformist type with the higher Q-value (marked in red in the left Q-table) for the next round and update its Q-table via Eq. (6). Finally, the agent adopts cooperation as its strategy by Eq. (4) since its current type is profiteer. Consequently, the agent becomes a cooperative conformist at  $t = 4$ . It can be foreseen that if the agent chooses to exploit at  $t = 4$ , it will adopt the behavioral type corresponding to the Q-value marked in purple.

### 3. Simulation results and discussions

In this section, we present the experimental results and corresponding analysis. We conduct extensive simulations to study the effects of the memory-driven Q-learning model for behavioral type adaptation on cooperation evolution and agent's type dynamics.



**Fig. 2. The evolution of cooperation frequency in relation to time under different cost-to-benefit ratios.** Simulations are performed under 14 distinct values of  $r$  until the population stabilizes in either a full cooperation (FC) state or a full defection (FD) state, where a vertical dashed line corresponding to the color is marked. Parameters for the memory-driven Q-learning model include the learning rate  $\alpha = 0.8$ , discount factor  $\gamma = 0.5$ , exploration probability  $\epsilon = 0.1$ , memory length  $M = 5$ , and memory strength  $\beta = 0.2$ . Curves converging to FC are depicted in red gradients (from dark red at  $r = 0.50$  to light pink at  $r = 0.56$ ), while those reaching FD are shown in blue gradients (from light blue at  $r = 0.57$  to navy blue at  $r = 0.80$ ). We can deduce that the critical threshold determining the population's ultimate state satisfies  $0.56 < r_0 < 0.57$ .

The four individual types include: cooperative profiteers (CP), cooperative conformists (CC), defective profiteers (DP), and defective conformists (DC).

First, we analyze the relationship between cooperation frequency and cost-to-benefit ratio, capturing snapshots of agent type distributions at key time steps. Results reveal that populations exclusively stabilize in either a full cooperation (FC) state or a full defection (FD) state under varying cost-to-benefit ratio environments. To explain this, we track the proportion dynamics of all four agent types in both scenarios and introduce the feedback-based  $\Delta Q$  to quantify behavioral preferences, which is defined as:

$$\Delta Q_{t+1}^X = \frac{\sum_X [Q_{t+1}^X(s, a) - Q_t^X(s, a)]}{N_{t+1}^X}, \quad (8)$$

where  $X$  denotes CP, CC, DP or DC,  $N_{t+1}^X$  is the number of  $X$  at time step  $t + 1$ . Previous studies introduced a  $\Delta Q$ , defined as the difference in Q-values between two actions in the current state, which reflects agents' preferences for action selection at a specific time step [36]. However, our newly proposed feedback-based  $\Delta Q$  evaluates temporal behavioral feedback:  $\Delta Q > 0$  implies a positive outcome, increasing the likelihood of repeating the behavior in similar future states.

Next, we mainly study the FC state. By adjusting the exploration probability  $\epsilon$  in Q-learning, we uncover an evolutionary shift from conformists to profiteers. Moreover, we explore the joint impact of  $\epsilon$  and the learning rate  $\alpha$  on the evolution of agent types. In addition, we explore how memory-driven Q-learning parameters affect the time to achieve cooperation. In particular, we compare our proposed Q-learning type transition mechanism with the other three mechanisms to illustrate its advantages in promoting the evolution of cooperation.

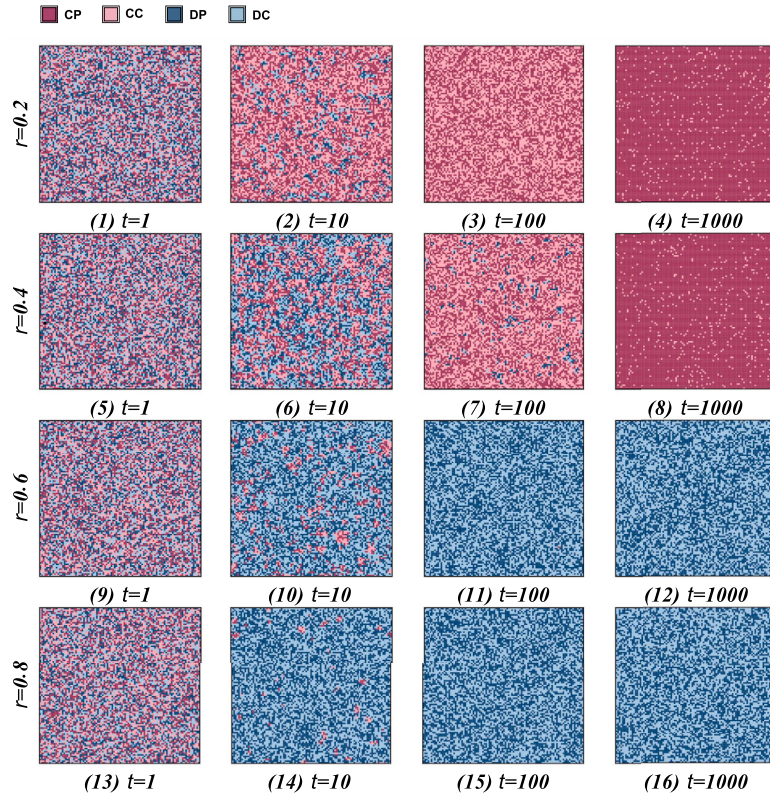
The above simulations are performed on a  $100 \times 100$  regular square lattice (SL) with periodic boundary conditions and von Neumann neighborhood. We also perform simulations on the triangular and hexagonal lattices in the last subsection, proving the scalability of the conclusion on the square lattice. The noise factor during strategy updates is fixed at  $\kappa = 0.1$ . The agent types in the initial network are randomly assigned, resulting in the number of each four types being roughly the same at the beginning.

### 3.1. Evolution of cooperation

The game model determines the environment of the evolutionary game and thereby influences cooperation levels, typically measured by the cooperation frequency  $f_c$ , which represents the proportion of cooperative agents at a given time step. We first investigate the evolution of  $f_c$  across different game environments. Initially, we test five typical cost-to-benefit ratios:  $r = 0.2, 0.4, 0.5, 0.6, 0.8$ , then refine the analysis between  $r = 0.5$  and  $r = 0.6$  with a step size of 0.01.

From a conceptual perspective, as  $r$  increases, the payoffs for defection grow while the costs of cooperation rise, leading to an expected decline in cooperation fraction ( $f_c$ ). Fig. 2 illustrates the evolution of  $f_c$  under the aforementioned cost-to-benefit ratio conditions. In particular, in contrast to mechanisms proposed in prior studies [26,32,40], where the stabilized cooperation proportion transitions gradually between 0 and 1 with changes in  $r$ , our mechanism exhibits abrupt transitions: for  $r \leq 0.56$ ,  $f_c$  ultimately stabilizes at 1, achieving a full cooperation (FC) state, though increasing  $r$  leads to greater fluctuations and longer convergence times to FC. Conversely, for  $r \geq 0.57$ ,  $f_c$  stabilizes at 0, resulting in a full defection (FD) state, with shorter convergence times as  $r$  increases. These results demonstrate that as  $r$  increases, populations become less likely to attain FC and more prone to FD, with a critical threshold  $0.56 < r_0 < 0.57$  dictating the final state.





**Fig. 3.** The network's snapshots at four characteristic time points under four cost-to-benefit ratios. The parameters of memory-driven Q-learning remain identical to those in Fig. 2. Initially, agent types are randomly assigned in the network, with colored blocks representing different agent types: cooperative profiteers (CP), cooperative conformists (CC), defective profiteers (DP), and defective conformists (DC). When  $r = 0.2$  and  $r = 0.4$ , the network ultimately reaches full cooperation (FC) states; when  $r = 0.6$  and  $r = 0.8$ , the network ultimately reaches full defection (FD) states.

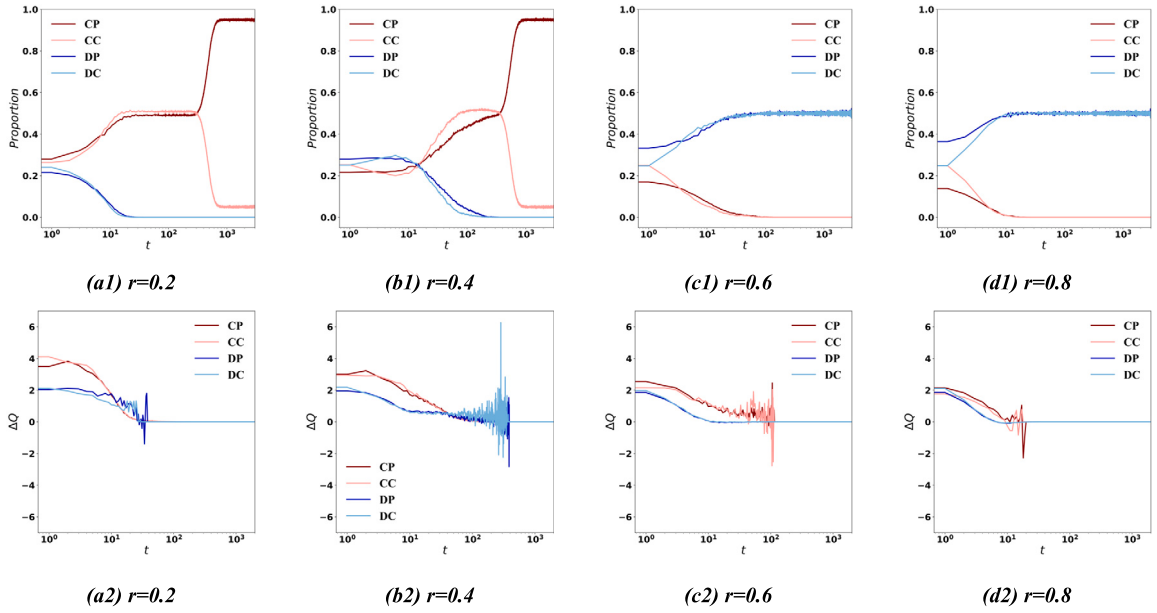
To explore the stages of cooperation evolution and facilitate subsequent analysis of type transitions, we capture network snapshots at four representative time steps ( $t = 1, 10, 100, 1000$ ) under different cost-to-benefit ratios. These snapshots preliminarily reveal the evolutionary patterns of four agent types in the network, as illustrated in Fig. 3. The network at  $t = 1$  displays a random distribution of four agent types. Evolutionary games with  $r = 0.2$  and  $r = 0.4$  ultimately reach the FC state, while those with  $r = 0.6$  and  $r = 0.8$  converge to the FD state. In the former cases, cooperators resist defector invasion through cluster formation. As shown in Figs. 3 (3) and (7), FC is achieved by  $t = 100$  at  $r = 0.2$ , whereas residual defectors persist at  $r = 0.4$ , which means increasing  $r$  progressively slows down cooperators' expansion efficiency. It is worth noting that conformists at  $t = 1000$  become sparsely distributed without observable clustering compared to  $t = 100$ , indicating their weak competitiveness against profiteers in FC states. We conjecture that conformist emergence may stem from agents' exploration behaviors in  $\epsilon$ -greedy mechanism, a proposition to be verified in next subsection. Thereby, the FC state formation process occurs through two distinct phases: initial coalition building between CP and CC against DP and DC, followed by competition between CP and CC.

Contrastingly, cooperators in the latter two cases with  $r = 0.6$  and  $r = 0.8$  maintain clustered aggregation but fail to withstand defector invasion, with faster defector dominance at a higher  $r$  value. Unlike FC scenarios, the agent exhibits no preference between DP and DC types in the FD state. As demonstrated in Figs. 3 (12) and (16), the two types maintain a comparable population ratio until  $t = 1000$ . Therefore, FD state evolution comprises a single consolidation phase, where DP and DC collectively attack CP and CC agents.

### 3.2. Evolution of type

Having investigated the evolutionary process of cooperation, we now focus on the transition dynamics of agent types. Under four representative cost-to-benefit ratios, we examine the temporal evolution of population proportions for four agent types (CP, CC, DP, DC). Using Eq. (8), we introduce feedback-based  $\Delta Q$  to explain type transitions from the perspective of behavioral feedback.

As shown in Fig. 4 (a1), when  $r = 0.2$ , cooperative agents initially exhibit a slight advantage over defective ones. Then the proportion of cooperators increases while defectors decline, widening the gap. Although DP persists slightly longer than DC, defectors become extinct before  $t = 100$ , marking the network's transition to the FC state. Concurrently, CP and CC dominate the network. Later on, after maintaining comparable proportions temporarily, CP rapidly grows and stabilizes at approximately 0.9, while CC declines sharply to around 0.1, aligning with the two-phase dynamics observed in Figs. 3 (1)~(8). This phenomenon of a sudden jump is also



**Fig. 4.** The temporal evolution of population proportions for different agent types and the corresponding  $\Delta Q$  values. Parameters for the memory-driven Q-learning remain identical to those in Fig. 2. We carry out a 3000-step experiment on the premise of ensuring the final stability. (a1)–(d1) illustrate the proportion dynamics of four agent types: CP, CC, DP, and DC. (a2)–(d2) depict the temporal evolution of  $\Delta Q$  calculated using Eq. (8).

observed in other evolutionary processes that converge to the FC state. However, it does not manifest in evolutionary processes that converge to the FD state. At  $r = 0.4$  demonstrated in (b1), the evolutionary pattern retains the two-phase structure but delays defector extinction to later stages. DP demonstrates significantly greater resilience than DC. Subsequently, CP surges while CC diminishes, mirroring the  $r = 0.2$  process, though CC briefly holds an early advantage over CP. For  $r = 0.6$  displayed in (c1), defectors immediately dominate and amplify their advantage, driving cooperators to extinction near  $t = 100$  as the network reaches the FD state. Later on, DP and DC proportions remain balanced throughout, consistent with the single-phase deduction. At  $r = 0.8$  shown in (d1), the evolution resembles  $r = 0.6$  but accelerates cooperator extinction, achieving the FD state by  $t = 10$ .

Figs. 4 (a2)–(d2) illustrate the evolutionary dynamics of four types of feedback-based  $\Delta Q$  under corresponding cost-to-benefit ratios. Observations reveal that fluctuations in feedback-based  $\Delta Q$  are strongly correlated with agents' extinction. When a specific type of individual exhibits significant  $\Delta Q$  volatility, its population approaches extinction. Although fluctuations in  $\Delta Q$  inevitably relate to declining population sizes, this phenomenon critically undermines the collective payoff levels of remaining individuals of the same type. Consequently, such agents enter a vicious cycle of evolutionary decline, ultimately leading to inevitable extinction. This mechanism serves as the primary driver for networks transitioning into either the FC state or FD state. Furthermore, as  $r$  increases, the initial evolutionary advantage of cooperation over defection diminishes and the fluctuating population type in later evolutionary stages shifts from defectors to cooperators, thereby influencing the network's final equilibrium state.

Previous experiments indicate that in evolutionary games reaching the FC state, the second phase involves the transition from CC to CP. We hypothesize that the CC agents appearing after this phase result from stochastic exploration. Consequently, the theoretically stable proportion of CC after the second phase should be:

$$P_{CC}^* = \epsilon \cdot p, \quad (9)$$

where  $p$  denotes the probability of selecting conformists during exploration. In random exploration,  $p = 1/2$ , yielding  $P_{CC}^* = \epsilon/2$ . As depicted in Fig. 5 (b), when  $\epsilon = 0.2$ , the network first undergoes the first phase (cooperators resisting defectors), followed by the second phase (CP resisting CC). Ultimately, the CC proportion stabilizes near 0.1. Apart from the case of 0-exploration, the phenomenon of a sudden jump persists, and its occurrence time appears to be independent of  $\epsilon$ . Subfigures (c)–(f) in Fig. 5 differ from (b) only in the final stable CC and CP proportions. Since the CC proportions in all five experiments stabilize after 1000 steps, we compute their averaged proportions from steps 1000~2000 and compare them with theoretical values to derive relative errors  $e$  in Table 1, where  $e = |\frac{P_{CC} - P_{CC}^*}{P_{CC}^*}|$ . We observe that the experimental results closely align with theoretical predictions, with errors across all five experimental groups remaining within 0.1%. Besides, through Figs. 5 (b)–(f), we can obtain that the proportion of stabilized CC coincides with the horizontal theoretical line derived from Eq. (9). Therefore, for evolutionary games converging to the FC state, we conclude that the only reason for the emergence of CC in the second phase is the random exploration of agents.

Notably, Fig. 5 (a) ( $\epsilon = 0.0$ ) exhibits a distinct single-phase evolution. Unlike scenarios with moderate exploration where CP and CC populations fluctuate, in the 0-exploration case, all four agent types maintain constant counts after stabilization. Under this condition, the network cannot achieve an absolute FC state; a small fraction of defectors persists, and the cooperation proportion

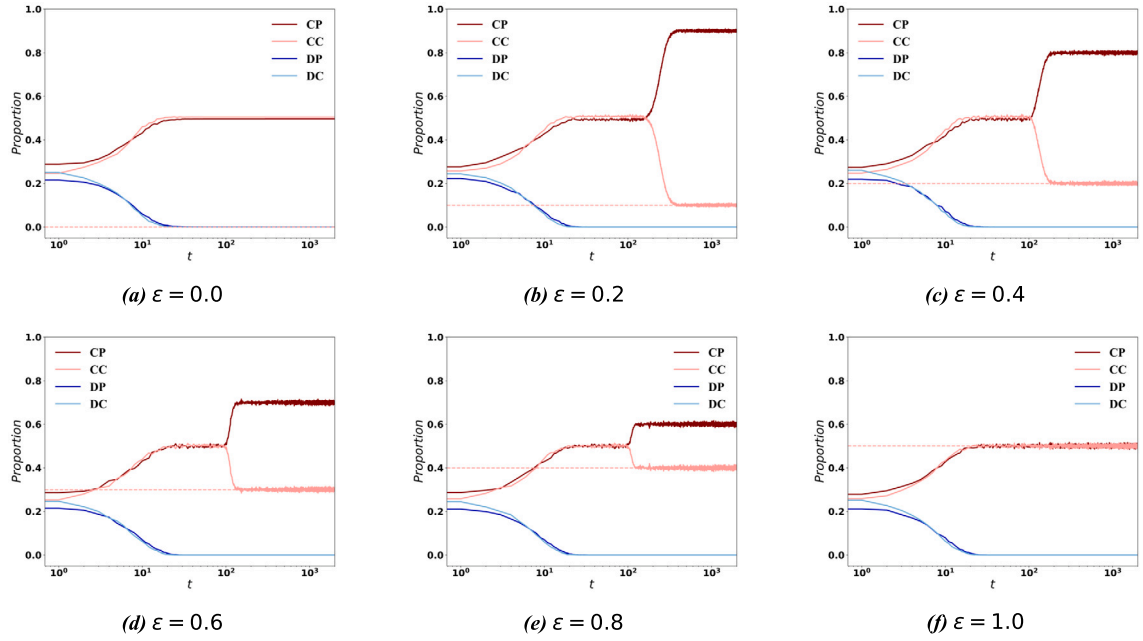


Fig. 5. The evolution of population proportions for four agent types under varying exploration probabilities. Parameters for the memory-driven Q-learning remain consistent with Fig. 2 except for the exploration probability  $\epsilon$ . The cost-to-benefit ratio is fixed to  $r = 0.2$ . Simulations are run for 2000 steps to ensure the stabilization of proportions. Six subplots respectively present results for  $\epsilon = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ . The horizontal dashed line in the figure indicates the proportion of CC that is theoretically stabilized.

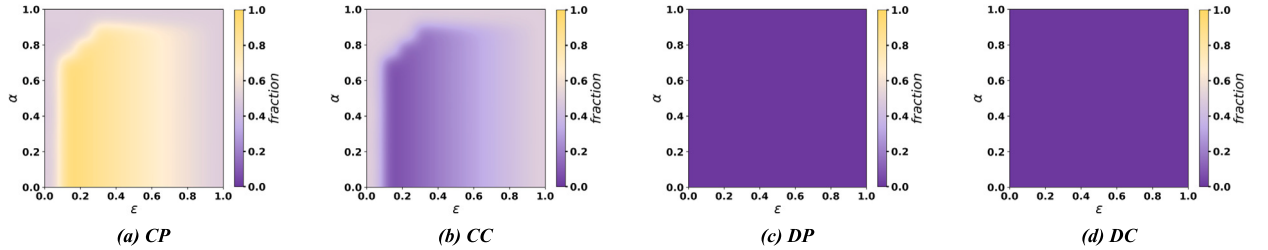


Fig. 6. Heatmaps depicting the stabilized proportions of four agent types as functions of exploration probability  $\epsilon$  and learning rate  $\alpha$  during evolutionary dynamics. The Q-learning parameters, except for  $\epsilon$  and  $\alpha$ , are consistent with those in Fig. 2. The cost-to-benefit ratio is set to  $r = 0.2$  to investigate evolutionary processes converging to the full cooperation state. Experiments are conducted over 3000 steps to ensure stabilization, with the average proportions calculated from the final 500 steps. Subplots (a), (b), (c), and (d) represent the proportion heatmaps for cooperative profiteers (CP), cooperative conformists (CC), defective profiteers (DP), and defective conformists (DC), respectively.

Table 1

The experimental-theoretical error in the stabilized CC proportion.

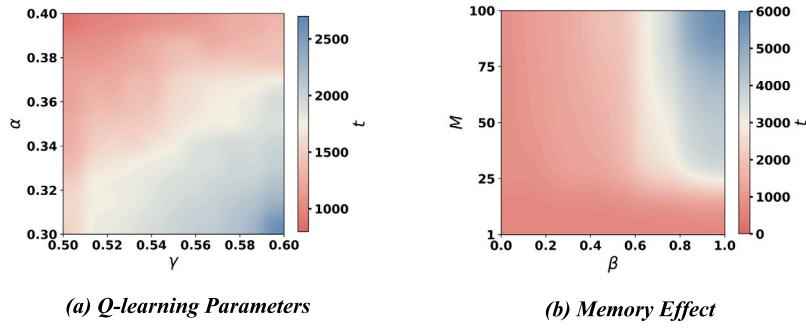
$P_{CC}$  represents the averaged CC proportion over the last 1000 steps for the five experiments in Figs. 5 (b)~(f).  $P_{CC}^*$  denotes the theoretical  $P_{CC}$  value calculated via Eq. (9). The relative error  $e = \left| \frac{P_{CC} - P_{CC}^*}{P_{CC}^*} \right|$ .

	$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$	$\epsilon = 1.0$
$P_{CC}$	0.10002	0.20015	0.30021	0.39983	0.49984
$P_{CC}^*$	0.10000	0.20000	0.30000	0.40000	0.50000
$e(\%)$	0.02	0.075	0.07	0.0425	0.032

remains constant, which is very close to 1 but never reaches 1. This phenomenon underscores the necessity of moderate exploration in Q-learning.

Furthermore, by incorporating learning rate  $\alpha$ , we get heatmaps illustrating the stabilized proportions of four agent types, as shown in Fig. 6. Given the relatively straightforward character of FD evolutionary process, our study continues to focus on FC evolution. Subplots (c) and (d) indicate that DP and DC ultimately become extinct across all parameter settings since the cost-to-benefit ratio is set to  $r = 0.2$ . Subplots (a) and (b) reveal that as  $\epsilon$  increases from 0.1, the proportion of CP gradually decreases while that of CC increases, with their stabilized proportions tending to equality when  $\epsilon = 1.0$ , consistent with the findings above. The light purple





**Fig. 7.** Heatmaps illustrating the time  $t$  required to reach the full cooperation state. (a) Time  $t$  as a function of learning rate  $\alpha \in [0.30, 0.40]$  and discount factor  $\gamma \in [0.50, 0.60]$  with fixed parameters  $M = 5, \beta = 0.2, \epsilon = 0.1$  and  $r = 0.5$ . (b) Time  $t$  as a function of memory length  $M$  and memory strength  $\beta$  with fixed parameters  $\alpha = 0.6, \gamma = 0.5, \epsilon = 0.1$  and  $r = 0.5$ . Warmer hues indicate shorter times  $t$ , reflecting enhanced promotion of cooperation.

vertical column corresponding to  $\epsilon = 0$  reflects the special case discussed in the previous simulation. In addition, as  $\alpha$  increases from 0, it has a negligible impact on the stabilized proportions of CP and CC initially. However, when  $\alpha \geq 0.7$ , an increase in  $\alpha$  suppresses the spread of CP agents and promotes the spread of CC agents under low exploration probabilities, leading to equality of their stabilized proportions, which becomes evident across all exploration probabilities when  $\alpha = 1.0$ .

### 3.3. Effect of memory-driven Q-learning on cooperation

In this subsection, we independently explore the effects of adjusting parameters in memory-driven Q-learning on cooperation, including the learning rate  $\alpha$ , discount factor  $\gamma$ , memory length  $M$ , and memory strength  $\beta$ . The focus of this subsection is on the evolutionary processes that converge to the FC state, with the impact on cooperation specifically quantified by the time required to achieve the FC state, rather than the steady-state cooperation level. Furthermore, we evaluate our proposed mechanism against three distinct settings, highlighting the advantages of employing Q-learning for agents' type transitions in promoting cooperation.

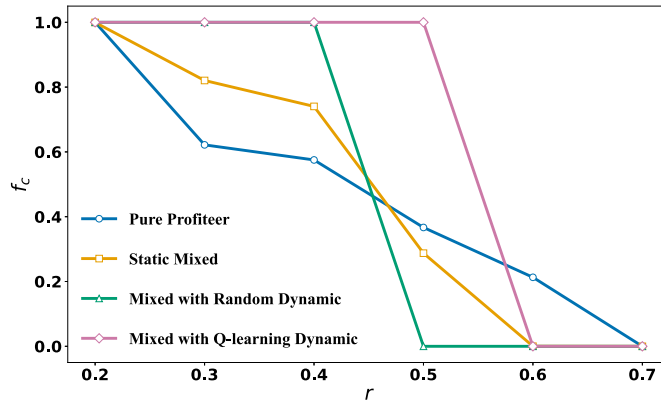
Fig. 7 (a) shows a heatmap illustrating the impact of the Q-learning parameters (learning rate  $\alpha$  and discount factor  $\gamma$ ) on the time  $t$  required to reach the FC state. For a fixed  $\alpha$ , increasing  $\gamma$  leads to a gradual increase in the time required to reach the FC state, indicating that prioritizing long-term rewards hinders cooperation. This suggests that cooperation is promoted when agents focus more on immediate benefits. Conversely, for a fixed  $\gamma$ , increasing  $\alpha$  reduces the time to reach the FC state, which means a higher learning rate promotes cooperation.

Fig. 7 (b) illustrates the relationship between the time required to reach the FC state and the memory parameters, namely memory length  $M$  and memory strength  $\beta$ . When  $M = 1$ , as derived from Eq. (7), the value of  $\beta$  does not affect the Q-learning reward, resulting in a uniform color at the corresponding position in the heatmap. Overall, the heatmap transitions from reddish tones in the lower-left corner to bluish tones in the upper-right corner, indicating that increases in both memory parameters slow down the emergence of cooperation. This experiment demonstrates that agents should prioritize current interests as the primary basis for future decision-making. Excessive focus on historical benefits hinders cooperative outcomes, a conclusion in line with the impact of  $\gamma$ , collectively underscoring the critical role of present-oriented evaluation in cooperation promotion.

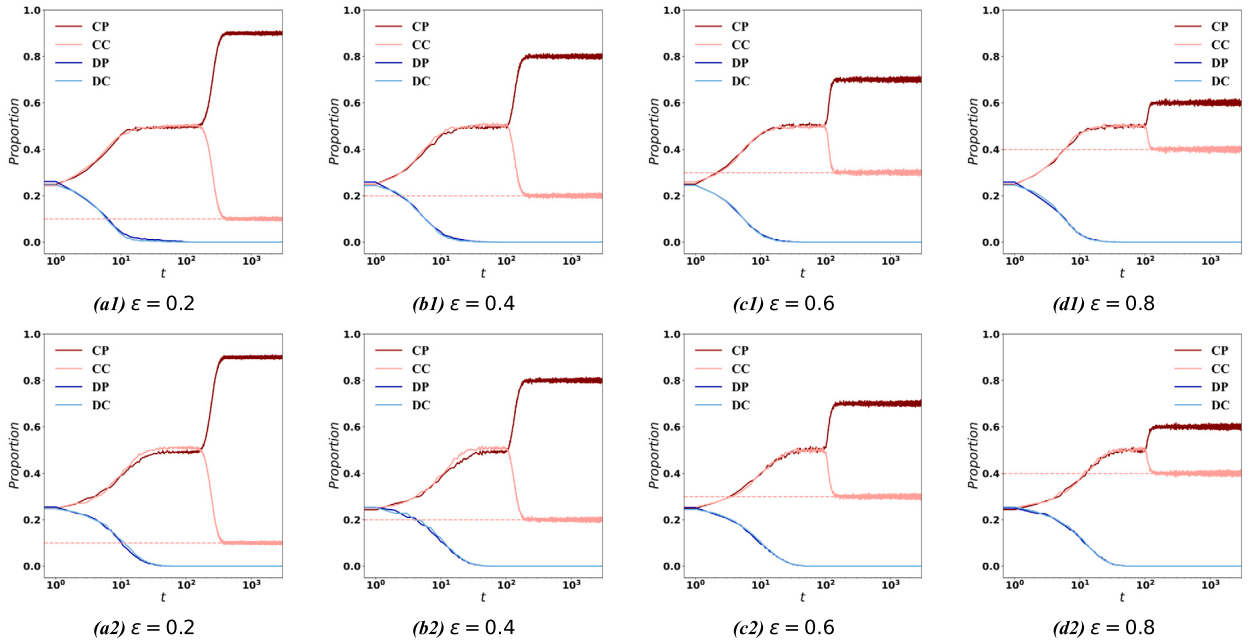
We consider it essential to compare Q-learning with other mechanisms to confirm its effectiveness. To this end, we use three distinct settings, as shown in Fig. 8, which illustrates the relationship between the stabilized cooperation proportion and the parameter  $r$ . The blue line shows a population composed entirely of profiteers; the yellow line denotes a population with an equal mix of profiteers and conformists, where each individual's strategy remains fixed; the green line indicates a scenario where each individual randomly selects to be a profiteer or conformist with equal probability at each time step; and the purple line corresponds to our proposed mechanism, wherein individuals dynamically adjust their behavioral type via Q-learning. Notably, we do not conduct simulations with a purely conformist population, as cooperation levels in this setting are independent of individual payoffs (unrelated to  $r$ ). The results demonstrate that our proposed mechanism sustains the FC state across the range  $r \leq 0.5$ , whereas cooperation levels in other settings exhibit varying degrees of decline. For  $r \geq 0.6$ , all settings except the pure profiteer case transition to the FD state, with the pure profiteer setting also entering the FD state at  $r = 0.7$ . Overall, our proposed mechanism exhibits superior performance in promoting cooperation.

### 3.4. Exploration of other network structures

Considering all simulations above are conducted on the regular square lattice, where each agent has a degree of four. In this subsection, we investigate the effect of other network structures on the evolution results and whether the conclusion derived from Eq. (9) remains valid. We take the triangular and hexagonal lattices as examples, focusing on the proportion dynamics of four types of agents. As shown in Fig. 9, the top four panels depict the results for the triangular lattice, where all agents have a degree of six, while the bottom four panels represent the results for the hexagonal lattice, where all agents have a degree of three. Both sets of results closely resemble those obtained on the square lattice, as presented in Fig. 5. Under cost-to-benefit ratio  $r = 0.2$ , the evolutionary dynamics still unfold in two phases: in the first phase, cooperation rapidly dominates, leading to the extinction of defectors. In the



**Fig. 8.** A line chart illustrating the stabilized cooperation frequency as a function of the cost-to-benefit ratio  $r$  with four different settings. When  $r < 0.2$ , all scenarios stabilize in the FC state, whereas when  $r \geq 0.7$ , all scenarios stabilize in the FD state. The blue line represents a scenario where all individuals are fixed as profiteers. The yellow line depicts a population with a fixed composition of half profiteers and half conformists, with each individual's behavioral type remaining unchanged. The green line indicates a population where, at each time step, individuals randomly choose to be either a profiteer or a conformist with equal probability. The purple line corresponds to our proposed mechanism, where individuals use Q-learning to determine whether to act as a profiteer or a conformist in the next time step, with parameters consistent with those in Fig. 2. All experiments are conducted over 5000 steps, with the average over the final 1000 steps taken as the result.



**Fig. 9.** Population dynamics of four agent types under different exploration probabilities on triangular and hexagonal lattices. The cost-to-benefit ratio is fixed at  $r = 0.2$ . The memory-driven Q-learning parameters are consistent with Fig. 2, except for the exploration probability  $\epsilon$ . Simulations are conducted over 3000 steps to achieve stabilized proportions. Subplots (a1), (b1), (c1), and (d1) depict results for the triangular lattice, while subplots (a2), (b2), (c2), and (d2) show results for the hexagonal lattice. The horizontal dashed line indicates the theoretically stabilized proportion of CC.

second phase, the evolutionary trajectory is from CC to CP. The phenomenon of sudden jump still exists on these networks, with the final stable value of CP decreasing and that of CC increasing as  $\epsilon$  grows. Table 2 compares the stabilized proportion of CC with the theoretical values calculated employing Eq. (9). It can be seen that the errors are all within 0.25%. Therefore, we conclude that the finding expressed by Eq. (9) holds true for both triangular and hexagonal lattices.

#### 4. Conclusions and outlooks

In this study, we investigate the evolutionary game between profiteers and conformists. By defining the action space of Q-learning as these two agent types, we introduce a Q-learning model for dynamic type adaptation. Moreover, we incorporate the memory mechanism into the reward calculation of Q-learning. At each time step, agents first select cooperation or defection based on their type-specific strategy update rules. Subsequently, they employ Q-learning to determine their behavioral type for the next time step.

Table 2

Comparison of experimental and theoretical stabilized CC proportions on triangular and hexagonal lattices.  $P_{CC}$  represents the mean CC proportion over the final 500 steps for the experiments shown in Fig. 9.  $P_{CC}^*$  denotes the theoretical value calculated using Eq. (9). The relative error is defined as  $e = |\frac{P_{CC} - P_{CC}^*}{P_{CC}^*}|$ .

		$\epsilon = 0.2$	$\epsilon = 0.4$	$\epsilon = 0.6$	$\epsilon = 0.8$
Triangular lattice	$P_{CC}$	0.10024	0.19991	0.29995	0.40009
	$P_{CC}^*$	0.10000	0.20000	0.30000	0.40000
	$e$ (%)	0.24	0.045	0.0167	0.0225
Hexagonal lattice	$P_{CC}$	0.09992	0.19970	0.29994	0.39955
	$P_{CC}^*$	0.10000	0.20000	0.30000	0.40000
	$e$ (%)	0.08	0.15	0.02	0.1125

according to their current states. Finally, they compute Q-learning rewards by integrating payoffs from the current and previous rounds to update their Q-tables.

Our simulations on cooperation evolution reveal that the system converges to a full cooperation state at lower cost-to-benefit ratios  $r$  and transitions to a full defection state at higher  $r$ . Next, we focus on the evolution of four agent types: cooperative profiteers (CP), cooperative conformists (CC), defective profiteers (DP), and defective conformists (DC). Further simulations show the evolution toward full cooperation unfolds in two phases: (i) CP and CC eliminate DP and DC until defection vanishes, followed by (ii) CP outcompeting CC, with CP demonstrating dominance. Conversely, the evolution of full defection occurs in a single phase, where DP and DC eradicate CP and CC. These extinction dynamics are explained through the proposed feedback-based  $\Delta Q$  analysis. Surprisingly, during the second phase of full cooperation evolution, CC populations stabilize at non-zero equilibrium values rather than disappearing entirely. We attribute this unexpected persistence to the stochastic exploration inherent in the  $\epsilon$ -greedy strategy, which introduces variability in type selection, allowing CC to resist complete extinction. This finding is also validated in other network topologies, including triangular and hexagonal lattices. Besides, we perform simulations about the combined influence of exploration probability  $\epsilon$  and learning rate  $\alpha$  on type evolution and observe that high learning rates suppress the spread of CP and promote the spread of CC under low exploration probabilities. Additionally, we systematically study the impact of the parameters, including learning rate, discount factor, memory length, and memory strength within our memory-driven Q-learning framework, on the evolution of cooperation. The results demonstrate that prioritizing immediate rewards over historical or future payoffs significantly boosts cooperative outcomes, offering a practical insight for designing adaptive systems. Furthermore, comparative simulations with other mechanisms demonstrate the advantage of our Q-learning approach in promoting the emergence and evolution of cooperation.

Overall, our model offers several advantages. It shifts empirically driven behavioral transitions from mere game strategy selection to the level of behavioral types, capturing the instability of individual behavioral type preferences observed in real-world scenarios. The discussion on the memory mechanism demonstrates the effectiveness of prioritizing immediate payoffs in promoting cooperation, which is both intriguing and thought-provoking. However, there are some aspects of this paper that could be improved. For example, in the strategy update process of profiteers, they can select the agent with the highest payoff from neighbors for payoff comparison, rather than randomly selecting a neighbor. Besides, extending the snowdrift game, which is the focus of this paper, to the generalized pairwise game is also a worthwhile aspect to investigate. Moreover, the conclusion regarding the role of immediate payoffs in enhancing cooperation remains at the level of qualitative simulations. Providing a theoretical proof would significantly advance the analysis of multi-agent systems in practical applications.

## CRedit authorship contribution statement

**Xiang Li:** Visualization, Methodology, Formal analysis, Investigation, Writing – original draft. **Bin Pi:** Validation, Formal analysis, Conceptualization, Writing – original draft, Data curation. **Liang-Jian Deng:** Writing – review & editing, Supervision, Funding acquisition. **Qin Li:** Writing – review & editing, Supervision, Methodology, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work is supported in part by the National Nature Science Foundation of China (NSFC) under Grant No. 12271083, and in part by the Natural Science Foundation of Sichuan Province under Grant No. 2022NSFSC0501.

## Data availability

Data will be made available on request.

## References

- [1] E.N. Barron, *Game Theory: An Introduction*, John Wiley & Sons, Hoboken, NJ, 2024.
- [2] Q. Li, et al., Open data in the digital economy: an evolutionary game theory perspective, *IEEE Trans. Comput. Soc. Syst.* 11 (3) (2023) 3780–3791.
- [3] Z. Zeng, et al., Complex network modeling with power-law activating patterns and its evolutionary dynamics, *IEEE Trans. Syst. Man Cybern. Syst.* 55 (4) (2025) 2546–2559.
- [4] Z. Xu, et al., Memory-based spatial evolutionary prisoner's dilemma, *Chaos Solitons Fractals* 178 (2024) 114353.
- [5] M. Feng, et al., Information dynamics in evolving networks based on the birth-death process: random drift and natural selection perspective, *IEEE Trans. Syst. Man Cybern. Syst.* 54 (8) (2024) 5123–5136.
- [6] R. Sugden, et al., *The Economics of Rights, Co-Operation and Welfare*, Springer, 2004.
- [7] Z.-W. Ding, et al., Emergence of anti-coordinated patterns in snowdrift game by reinforcement learning, *Chaos Solitons Fractals* 184 (2024) 114971.
- [8] X. Xiong, et al., Adaptive payoff-driven interaction in networked snowdrift games, *Chaos Solitons Fractals* 185 (2024) 115187.
- [9] B. Skyrms, *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press, 2004.
- [10] M. Bello, et al., Intuition and deliberation in the stag hunt game, *Sci. Rep.* 9 (1) (2019) 14833.
- [11] C. Herfeld, Revisiting the criticisms of rational choice theories, *Philos. Compass* 17 (1) (2022) e12774.
- [12] M. Ye, et al., Distributed Nash equilibrium seeking in games with partial decision information: a survey, *Proc. IEEE* 111 (2) (2023) 140–157.
- [13] Z. Zeng, et al., Bursty switching dynamics promotes the collapse of network topologies, *Proc. R. Soc. A* 481 (2310) (2025) 20240936.
- [14] L.S. Flores, et al., Cooperation in regular lattices, *Chaos Solitons Fractals* 164 (2022) 112744.
- [15] Q. Jian, et al., Impact of reputation assortment on tag-mediated altruistic behaviors in the spatial lattice, *Appl. Math. Comput.* 396 (2021) 125928.
- [16] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* 393 (6684) (1998) 440–442.
- [17] H. Wang, et al., Epidemic dynamics on higher-dimensional small world networks, *Appl. Math. Comput.* 421 (2022) 126911.
- [18] C. Liu, et al., Evolution of strategies in evolution games on small-world networks and applications, *Chaos Solitons Fractals* 189 (2024) 115676.
- [19] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, *Science* 286 (5439) (1999) 509–512.
- [20] Y. Yao, et al., Protection and improvement of indirect identity cognition on the spatial evolution of cooperation, *Phys. A* 620 (2023) 128791.
- [21] J. Wang, et al., Strategically positioning non-competitive individuals can rescue cooperation in scale-free networks, *Europhys. Lett.* 143 (4) (2023) 41002.
- [22] S. Milgram, The small world problem, *Psychol. Today* 2 (1) (1967) 60–67.
- [23] R. Albert, Scale-free networks in cell biology, *J. Cell Sci.* 118 (21) (2005) 4947–4957.
- [24] A.J. Stewart, J.B. Plotkin, Small groups and long memories promote cooperation, *Sci. Rep.* 6 (1) (2016) 26889.
- [25] Z. Danku, et al., Knowing the past improves cooperation in the future, *Sci. Rep.* 9 (1) (2019) 262.
- [26] B. Pi, et al., An evolutionary game with conformists and profiteers regarding the memory mechanism, *Phys. A* 597 (2022) 127297.
- [27] M. Milinski, et al., Reputation helps solve the 'tragedy of the commons', *Nature* 415 (6870) (2002) 424–426.
- [28] M. Feng, et al., An evolutionary game with the game transitions based on the Markov process, *IEEE Trans. Syst. Man Cybern. Syst.* 54 (1) (2023) 609–621.
- [29] Y. Li, et al., Effects of compassion on the evolution of cooperation in spatial social dilemmas, *Appl. Math. Comput.* 320 (2018) 437–443.
- [30] H. Wei, et al., Moral preferences co-evolve with cooperation in networked populations, *IEEE Trans. Evol. Comput.* (2024).
- [31] A. Szolnoki, M. Perc, Conformity enhances network reciprocity in evolutionary social dilemmas, *J. R. Soc. Interface* 12 (103) (2015) 20141299.
- [32] B. Pi, et al., A memory-based spatial evolutionary game with the dynamic interaction between learners and profiteers, *Chaos* 34 (6) (2024).
- [33] 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.
- [34] M. Wunder, et al., Classes of multiagent q-learning dynamics with epsilon-greedy exploration, in: *Proc. 27th Int. Conf. Mach. Learn. (ICML-10)*, 2010, pp. 1167–1174.
- [35] Z. Yang, et al., Interaction state q-learning promotes cooperation in the spatial prisoner's dilemma game, *Appl. Math. Comput.* 463 (2024) 128364.
- [36] H. Guo, et al., Effect of state transition triggered by reinforcement learning in evolutionary prisoner's dilemma game, *Neurocomputing* 511 (2022) 187–197.
- [37] H. Liang, et al., Analysis and shifting of stochastically stable equilibria for evolutionary snowdrift games, *Syst. Control Lett.* 85 (2015) 16–22.
- [38] W. Ye, S. Fan, Evolutionary snowdrift game with rational selection based on radical evaluation, *Appl. Math. Comput.* 294 (2017) 310–317.
- [39] J. Pu, et al., Effects of time cost on the evolution of cooperation in snowdrift game, *Chaos Solitons Fractals* 125 (2019) 146–151.
- [40] J. Han, R. Wang, Complex interactions promote the frequency of cooperation in snowdrift game, *Phys. A* 609 (2023) 128386.