Reinforcement Learning Can Be a Double-Edged Sword for Cooperation on Higher-Order Networks

Yan Xu, Dawei Zhao, *Member, IEEE*, Tina Perc Benko, Chengyi Xia, *Senior Member, IEEE*, and Matjaž Perc

***Abstract*— Collective cooperation is fundamental to individual survival and social development, and exploring its mechanism of emergence is of great significance. However, most existing studies related to the evolutionary dynamics on higher order networks assume that all agents within a population follow the same strategy updating rule. This assumption does not align with reality and is an oversimplification. To this end, we propose a higher order network game framework featuring a hybrid strategy updating rule. Specifically, we use scale-free random hypergraphs (SRHs) to characterize the underlying network topology of the population. Then, we categorize agents into two types: imitation learners and autonomous learners according to social learning and behaviorism theories. For imitation learners, we apply the Fermi rule to characterize their probabilistic imitation behaviors, while for autonomous learners, we adopt the reinforcement learning method to highlight their decision-making features. Through a series of simulation experiments and theoretical analyses, we find that autonomous learners have a dual impact on cooperation in groups: they inhibit cooperation at low dilemma intensities but promote cooperation at high dilemma intensities. In addition, we show that smaller group sizes are more conducive to cooperation. Our findings provide valuable insights for better understanding the impact of hybrid updating mechanisms on the evolutionary dynamics of collective cooperation in higher order networks.**

***Index Terms*— Collective cooperation, evolutionary dynamics, higher order networks, hybrid strategy updating rule, reinforcement learning.**

Received 29 July 2025; accepted 15 October 2025. Date of publication 4 November 2025; date of current version 17 December 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62573316, in part by Tianjin Municipal Natural Science Foundation under Grant 22JCZDJC00550, in part by Shandong Provincial Natural Science Foundation under Grant ZR2020YQ06, in part by Taishan Scholars Program under Grant TSQN202211210, in part by the Slovenian Research and Innovation Agency under Grant P1-0403 and Grant P5-0027, and in part by the Open Research Topics of the Key Laboratory of Computing Power Network and Information Security Ministry of Education under Grant 2024ZD026 and Grant 2023PY010. This article was recommended by Associate Editor H. Yuan. *(Corresponding authors: Dawei Zhao; Chengyi Xia; Matjaž Perc.)*

Yan Xu is with the School of Artificial Intelligence, Tiangong University, Tianjin 300387, China, also with the Department of Computer Science, Tianjin Police Institute, Tianjin 300382, China, and also with the School of Computer Science and Engineering, Tianjin University of Technology, Tianjin 300384, China (e-mail: haprnxu@163.com).

Dawei Zhao is with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250014, China, and also with Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan 250014, China (e-mail: zhaodw@sdas.org).

Tina Perc Benko is with the Faculty of Natural Sciences and Mathematics, University of Maribor, 2000 Maribor, Slovenia (e-mail: tina.perc1@um.si).

Chengyi Xia is with the School of Artificial Intelligence, Tiangong University, Tianjin 300387, China (e-mail: cyxia@tiangong.edu.cn).

Matjaž Perc is with the Faculty of Natural Sciences and Mathematics, University of Maribor, 2000 Maribor, Slovenia, also with the Community Healthcare Center Dr. Adolf Drolc Maribor, Maribor, Slovenia, also with the Department of Physics, Kyung Hee University, Seoul 02447, Republic of Korea, also with the Complexity Science Hub, Vienna, Austria, and also with the University College, Korea University, Seoul 02841, Republic of Korea (e-mail: matjaz.perc@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSMC.2025.3624366.

Digital Object Identifier 10.1109/TSMC.2025.3624366

NOMENCLATURE

$\mathcal{H}(\mathcal{X}, \mathcal{E})$	Hypergraph.
N	Number of nodes.
k	Order of a hyperedge.
$d_H(i)$	Hyperdegree of node i.
r	Dilemma intensity.
α	Proportion of autonomous learners.
η	Learning rate.
δ	Discount factor.
ϵ	Exploration rate.
κ	Noise intensity.
f_c	Cooperation rate.

I. INTRODUCTION

COOPERATION is the foundation of individual survival and even social prosperity [1], [2]. The emergence of prosocial behaviors, which benefit others at personal costs, remains an active research topic within evolutionary theory [3]. In the real world, cooperation occurs not just in dyadic interactions but more extensively in interactions among three or more agents [4], [5], [6], such as a pack of wolves hunting together. Nondyadic interactions have been shown to give rise to new collective behaviors and dynamic patterns [7], [8], [9]. Pairwise interactions, or even aggregations of multiple pairwise interactions, often fail to capture the essence of collective social behaviors. Higher-order networks, such as hypergraphs [10], [11] and simplicial complexes [4], [12], can better model multiagent interactions, providing an effective means to investigate the evolutionary dynamics of group cooperation.

In recent years, research based on complex networks has emphasized the impact of population structures [10], [13] and strategy updating rules [14], [15] on collective cooperation. Different population structures lead to distinct dynamic features for cooperative behaviors during the evolutionary process. As an example, Alvarez-Rodriguez et al. [10] examined public goods games on hypergraphs possessing diverse topological features and analyzed the dependence of group cooperation on its size through empirical data. Similarly, Civilini et al. [13] demonstrated that higher-order structures can promote cooperation, even in scenarios where pairwise interactions would lead to full defection. Moreover, Sheng et al. [16] developed a mathematical model to analyze the evolution of behavioral strategies on arbitrary higher-order topologies, revealing that higher-order interactions often enhance cooperation above and beyond purely dyadic interactions. Both numerical simulations [17], [18], [19] and theoretical analyses [16], [20] point out that higher-order structures have profound effects on system dynamics, potentially even reversing outcomes that arise under strictly dyadic interactions. Beyond evolutionary dynamics, higher-order interactions have also been demonstrated to generate novel collective phenomena in other fields, including spreading dynamics [9], [21], consensus formation [22], [23], and synchronization processes [24], [25]. In addition, strategy updating rules play a crucial role in the evolutionary outcomes within structured populations. These microscopic rules determine whether cooperation can evolve [26], [27], thereby further influencing the system's final cooperation level [28]. For example, Ohtsuki and Nowak [29] derived a replication equation under a given updating rule; they discovered that defection is favored under the pairwise-comparison updating rule, whereas cooperation can emerge when the death–birth updating rule is applied. Similarly, Zhang et al. [30] proposed a Moran-like process and compared it with the Fermi rule and replicator dynamics, showing that the Moran-like process can greatly facilitate the evolution of cooperation. From these studies, it is clear that the underlying network topologies and strategy updating rules are key factors for the emergence of cooperation. Therefore, exploring the influence of updating rules on the evolutionary dynamics within higher-order networks is significant and necessary to deeply understand the intrinsic motivation of collective cooperation.

Note that most existing works on the cooperative evolution in higher-order networks impose the assumption that all the members of a population are homogeneous. For example, Xu et al. [11] explored the impact of dynamic environments on higher-order cooperation, and Shi et al. [31] analyzed the learning dynamics of agents with higher-order interactions, etc. [32]. However, in these works, all the agents follow the same strategy updating rule. This assumption does not align with reality and represents an oversimplification. Indeed, individuals often possess diverse characteristics, backgrounds, and decision-making logics, leading to heterogeneous strategy updating rules within the population. According to Bandura's social learning theory [33] and Skinner's behaviorism theory [34], individuals can be broadly classified into two categories: imitation learners and autonomous learners. Imitation learners tend to mimic the behaviors of others, motivated by a desire for security; they believe that following successful examples or models can minimize trial-and-error costs and avoid bearing the consequences of poor decisions alone. On the contrary, autonomous learners prefer to rely on their own accumulated experiences and judgments, making decisions independently of others. To better understand the emergence mechanisms of cooperative behaviors in social dilemmas, it is necessary to develop a more realistic, hybrid strategy updating rule that integrates these two learning paradigms.

Therefore, in this study, we propose a higher-order network game model with a hybrid strategy updating rule. Specifically, we adopt scale-free random hypergraphs (SRHs) to represent the underlying network structure of the population. For imitation learners, the Fermi rule [35], [36] is chosen as the representative strategy updating mechanism. The Fermi rule effectively describes how agents make probabilistic decisions based on the performance of other peers under noisy and uncertain environments. This mechanism reflects the feature that agents struggle to imitate globally optimal decisions due to incomplete information and cognitive limitations, making it closely aligned with the behavioral patterns of imitation learners in real-world settings. On the other hand, for autonomous learners, the Q-learning method [31], [37] is selected as the representative approach, which allows agents to dynamically adjust their behaviors based on their own experience, without relying on others. It highlights the characteristics of autonomous decision-making and adaptive learning, which correspond to the key elements of autonomous learners' behavioral patterns. To summarize, the main contributions of this study are presented as follows.

1) This study proposes a new hybrid strategy updating framework based on higher-order networks to examine the evolutionary dynamics of cooperation. In contrast to prior studies that rely on pairwise interaction structures or a single updating rule, this framework adopts SRHs to model multiagent interactions, and then integrates multiple updating rules to capture the diverse decision-making patterns observed in real-world scenarios. This framework represents a certain level of generality and scalability.
2) To characterize the heterogeneity of human decision-making, this study introduces a more realistic heterogeneous learning approach by distinguishing between imitation learners and autonomous learners. For imitation learners, we use the Fermi rule to feature their probabilistic imitation behaviors, while for autonomous learners, we apply the Q-learning algorithm to describe their independent decision-making processes. By accounting for individual differences, this approach provides a feasible pathway for exploring the impact of behavioral diversity on collective cooperation.
3) A series of simulation experiments show that autonomous learners exert a dual impact on group cooperation under the hybrid updating mechanism. Specifically, when the dilemma intensity r is low, autonomous learners suppress cooperation; conversely, they promote cooperation when r is high. Our findings provide important scientific insights and practical

guidance for understanding cooperative dynamics within complex social systems.

The structure of this article is arranged as follows. Section II outlines some basic concepts related to the Q-learning algorithm and the multiplayer snowdrift game. In Section III, we describe the proposed hybrid strategy updating mechanism based on higher-order networks, including hypergraph construction and the evolutionary processes of imitation and autonomous learners. In Section IV, we present simulation results along with the corresponding analyses. Finally, Section V concludes this article by summarizing main findings and presenting potential directions for future works.

II. Related Works

A. Cooperative Evolution Based on Higher-Order Networks

Higher-order networks effectively capture multibody interactions, enabling more accurate simulations of complex interaction patterns observed in real-world systems, and have consequently garnered considerable attention in the academic community [16], [38]. For example, Alvarez-Rodriguez et al. [10] developed a hypergraph-based game model using empirical data to investigate the relationship between cooperative behaviors and group size. Wang et al. [39] further demonstrated that higher-order interactions can significantly enhance network reciprocity and promote the stability of cooperative strategies. Similar studies can also be found in [40] and [41]. In contrast to these studies, Pan et al. [42] and Chen et al. [43] explored evolutionary games on higher-order networks under heterogeneous conditions. Their findings indicate that heterogeneity in investments and multiplication factors can greatly enhance cooperation through nonuniform resource allocation—particularly under power-law distributions, where cooperative behaviors become more prominent.

In summary, although existing studies have shed light on the potential of higher-order interactions to promote cooperation, most of them adopt a single strategy update rule. This limitation makes it difficult to capture the diversity of decision-making processes observed in real-world systems, thereby reducing the generalizability and robustness of their findings.

B. Cooperative Evolution Driven by Reinforcement Learning Mechanisms

As a trial-and-error learning paradigm, reinforcement learning enables agents to iteratively refine their decision-making policies through continuous interaction with the environment, thereby maximizing long-term cumulative rewards [44], [45], [46]. Integrating reinforcement learning into game-theoretic models allows agents to adapt more effectively to complex and dynamic environments, thereby offering new insights into the emergence of cooperation and providing potential solutions to social dilemmas. Lu and Wang [47] proposed a modified strategy update rule based on reinforcement learning and demonstrated that it can significantly enhance overall cooperation levels. Song et al. [48] incorporated the Bush–Mosteller (BM) model with the iterated prisoner's dilemma, revealing the critical role of adaptive interaction strength in resolving social dilemmas. Zhang et al. [49] introduced the Q-learning algorithm into evolutionary game dynamics and discovered that reinforcement learning can lead to periodic oscillations in cooperation under specific game scenarios. Usui and Masahiko [50] theoretically demonstrated that strategies such as "win-stay, lose-shift," the grim trigger, and always-defect can form symmetric equilibria in repeated prisoner's dilemma games driven by reinforcement learning. In addition, Wang et al. [51] and Guo et al. [52] applied the self-regarding Q-learning algorithm to investigate how agent state transitions in dynamic interaction environments influence the evolution of cooperation.

It is worth noting that the above studies are all based on pairwise interaction networks. How reinforcement learning influence the evolution of group cooperation on higher-order networks remains an open question.

III. Preliminaries

A. Multiplayer Snowdrift Game

The multiplayer snowdrift game [53] is a classic paradigm for examining altruistic behavior among self-interested agents in group interaction settings, and it symbolizes a social dilemma where individual interests conflict with collective interests. Specifically, in a group of L agents, each agent independently decides whether to participate in snow shoveling. Those who choose to shovel are cooperators (C), while those who do not are defectors (D). The cost of shoveling, c, is evenly shared among cooperators. If the snowdrift is removed (i.e., there are cooperators), every agent, including defectors, receives a benefit b, with $b > c > 0$. The multiplayer snowdrift game reveals the nature of cooperation dilemmas: when individual benefits are equal, their contributions may differ. In other words, there is no direct connection between the benefits that individuals receive and the efforts they take.

The payoffs received by cooperators and defectors, respectively, within a k-agent group are as shown in the following equation:

$$\bar{P}_C(x_C) = b - \frac{c}{x_C} \quad \text{for } x_C \in [1, k]$$
$$\bar{P}_D(x_C) = \begin{cases} 0, & \text{for } x_C = 0 \\ b, & \text{for } x_C \in [1, k-1] \end{cases} \tag{1}$$

where x_C denotes the number of cooperating agents. Without loss of generality, by introducing $r = c/b$ as the cost–benefit ratio [53], the payoffs can be further expressed by the following equation:

$$P_C(x_C) = 1 - \frac{r}{x_C} \quad \text{for } x_C \in [1, k]$$
$$P_D(x_C) = \begin{cases} 0, & \text{for } x_C = 0 \\ 1, & \text{for } x_C \in [1, k-1]. \end{cases} \tag{2}$$

B. Q-Learning Algorithm

Q-learning [37], a model-free reinforcement learning algorithm, aims to discover a policy that can maximize expected rewards by iteratively updating an action–value function.

In this approach, an agent in state s_x, which belongs to the state set $\mathcal{S} = \{s_1, \ldots, s_n\}$, chooses an action a_y from the action set $\mathcal{A} = \{a_1, \ldots, a_m\}$ according to the values in its Q-table. The Q-table is defined as a Cartesian product matrix $S \times \mathcal{A} \rightarrow \mathbb{R}$ with the rows corresponding to states and the columns to actions, as outlined in the following equation:

$$\mathbf{Q}(t) = \begin{pmatrix} Q_{s_1,a_1}(t) & \cdots & Q_{s_1,a_m}(t) \\ \vdots & \ddots & \vdots \\ Q_{s_n,a_1}(t) & \cdots & Q_{s_n,a_m}(t) \end{pmatrix} \tag{3}$$

where $Q_{s_x,a_y}(t)$ denotes the estimated Q value for taking action a_y in state s_x. The element $Q_{s,a}$ in the Q-table is updated in accordance with the following equation:

$$\begin{aligned} Q_{s,a}(t+1) &= h(\mathbf{Q}(t), p(t)) \\ &= (1-\eta) Q_{s,a}(t) \\ &\quad + \eta \left[p(t) + \delta \max_{a' \in \mathcal{A}} Q_{s',a'}(t) \right] \end{aligned} \tag{4}$$

where $\eta \in (0, 1]$ denotes the learning rate, while $\delta \in [0, 1)$ represents the discount factor, which determines the importance of future rewards. $p(t)$ means the payoff that the agent receives at round t, and $\max_{a' \in \mathcal{A}} Q_{s',a'}(t)$ signifies the maximum Q value for state s'.

To avoid getting stuck in a local optimum, agents commonly use the ϵ-greedy approach for action selection, as presented in the following equation:

$$a(t) = \begin{cases} \arg\max\limits_{a' \in \mathcal{A}} Q_{s,a'}(t), & \theta > \epsilon \\ \mathcal{U}(\mathcal{A}), & \theta \leq \epsilon \end{cases} \tag{5}$$

where $\epsilon \in (0, 1]$ indicates the exploration rate, θ signifies a random number ranging from 0 to 1, and $\arg\max_{a' \in \mathcal{A}} Q_{s,a'}(t)$ refers to the action with the maximum Q value in the row corresponding to state s. In addition, $\mathcal{U}(\mathcal{A})$ denotes a uniform distribution over $\mathcal{A}$, implying that every action has the same probability of being selected. In other words, the agent leverages its prior knowledge (Q-table) to determine the current optimal policy with probability $1-\epsilon$ and explores the environment randomly with probability ϵ.

IV. Model

This study develops a hybrid strategy updating mechanism based on higher-order networks to examine the evolutionary dynamics of collective cooperation, as shown in Fig. 1. This model is composed of three parts. 1) *Higher-order Network Construction:* We construct SRHs to characterize the topological structure of the population, where nodes denote agents and hyperedges represent interacting groups. 2) *Initialization:* We initialize the types (autonomous or imitation learners) and strategies (cooperation or defection) of agents. 3) *Evolutionary Process:* Autonomous and imitation learners adopt the Q-learning algorithm and the Fermi rule, respectively, for the iterative evolution. Nomenclature summarizes a list of symbols used in our model.

A. Higher-Order Network Construction

A hypergraph $\mathcal{H}(\mathcal{X}, \mathcal{E})$ consists of a set $\mathcal{X} = \{x_1 = 1, \ldots, x_N = N\}$ with N nodes and a set $\mathcal{E} = \{e_1 = 1, \ldots, e_M = M\}$ with M hyperedges. Each hyperedge can link an arbitrary number of nodes, representing a group of interacting agents. The order of a hyperedge, denoted by k, refers to the number of nodes it contains. The hyperdegree $d_H(x)$ of node x is defined as the number of hyperedges to which it belongs. To accurately characterize higher-order interactions among agents, we use SRHs to represent the underlying topology of the population. In SRHs, the node hyperdegree follows a power-law distribution, and each hyperedge contains the same number of nodes. The method for constructing SRHs [54] is described below.

1) *Initialization:* Generate k nodes $v_1, v_2, \ldots, v_k$, and use them to construct a hyperedge $E_1 = \{v_1, v_2, \ldots, v_k\}$.
2) *Hyperedge Growth:*
   a) Select $m \in [1, k)$ existing nodes $v_1, v_2, \ldots, v_m$ from the hypergraph according to a preferential attachment mechanism. The probability $\Pi d_H(x)$ of selecting node x is defined as
   $$\Pi d_H(x) = \frac{d_H(x)}{\sum_{y \in N} d_H(y)}. \tag{6}$$
   b) Add $(k-m)$ new nodes $v_{m+1}, \ldots, v_k$.
   c) Connect the selected m existing nodes with the $(k-m)$ newly added nodes to form a new hyperedge $E_i = \{v_1, \ldots, v_m, v_{m+1}, \ldots, v_k\}$.
3) Repeat step 2) until the hypergraph is constructed.

In particular, when $k = 2$, the hypergraph simplifies to a conventional graph, obtaining the Barabási–Albert (BA) model [55]. In Fig. 2, we illustrate the initial three steps of the evolutionary process for an SRH with $k = 3$ and $m = 1$, where the red circles represent newly added nodes and the orange-shaded areas indicate newly formed hyperedges. Following the hypergraph generation algorithm, we construct SRHs with $N = 10\,000$, $m = 2$, and $k = 3, 5, 7$, respectively. As shown in Fig. 3, the hyperdegree distributions in these hypergraphs follow a power-law distribution, demonstrating that these hypergraphs exhibit scale-free properties.

B. Initialization

To capture the diversity of human decision-making patterns, we categorize agents into two groups: autonomous learners and imitation learners. Each agent has a probability α of being an autonomous learner and a probability $1-\alpha$ of being an imitation learner. The distinction between these two groups is reflected in their decision-making mechanisms: autonomous learners rely on the Q-learning algorithm, while imitation learners follow the Fermi rule for their decisions.

For autonomous learners, each agent has a state set $\mathcal{S}$ and an action (strategy) set $\mathcal{A}$, where $\mathcal{S} = \mathcal{A} = \{C, D\}$, with C representing cooperation and D denoting defection. Note that although the elements in $\mathcal{S}$ and $\mathcal{A}$ are the same, the connotations of actions and states differ. In other words, the current state of an agent is determined by its action in the previous step, and the action taken in the current state becomes

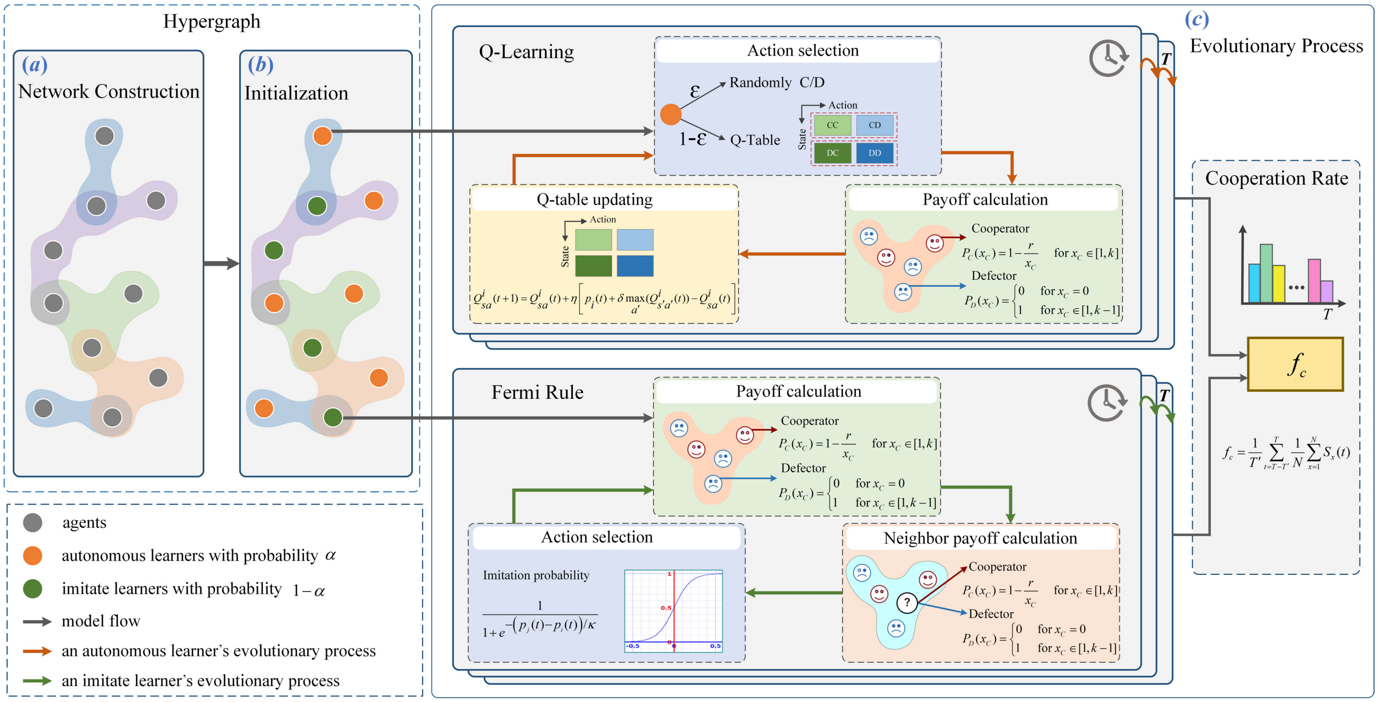


Fig. 1. Hybrid strategy updating framework based on higher-order networks. It encompasses three key elements. (a) Higher-order network construction. (b) Initialization. (c) Evolutionary process.

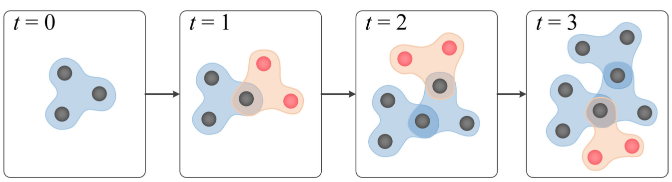


Fig. 2. Schematic of the evolving hypergraph with $k = 3$ and $m = 1$, illustrating the initial three steps of the evolution process.

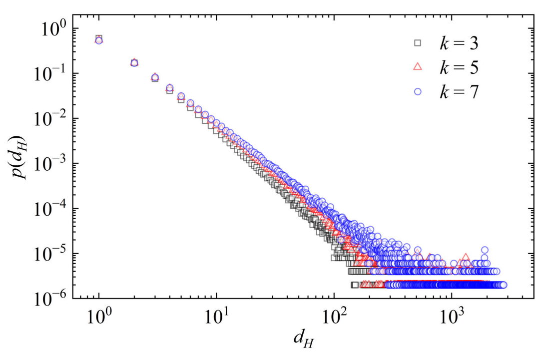


Fig. 3. Hyperdegree distributions of SRHs with $N = 10\,000$, $m = 2$, and $k = 3$, 5, and 7, respectively, on a logarithmic scale. We average the results over 50 independent realizations to ensure their reliability.

its state in the next step. In addition, since autonomous learners use the Q-learning algorithm to determine their actions during the evolutionary process, each autonomous learner is equipped with an independent Q-table. The agent's Q-Table, together with its state, constitutes its long-term memory for decision-making. The Q-table is a 2×2 matrix composed of state (row)–action (column) combinations, as defined in the following equation:

$$\mathbf{Q}^i(t) = \begin{bmatrix} Q^i_{s_1,a_1}(t) & Q^i_{s_1,a_2}(t) \\ Q^i_{s_2,a_1}(t) & Q^i_{s_2,a_2}(t) \end{bmatrix} = \begin{bmatrix} Q^i_{C,C}(t) & Q^i_{C,D}(t) \\ Q^i_{D,C}(t) & Q^i_{D,D}(t) \end{bmatrix} \tag{7}$$

where each element in the Q-table represents the value for agent i performing a particular action in a certain state during round t. Initially, every agent is randomly designated a state with an equal chance. Given that agents lack prior knowledge of their environment, their Q-tables are initialized to zero [56].

For imitation learners, each agent has an action (strategy) set $\mathcal{A} = \{C, D\}$. Initially, each agent randomly selects an action from $\mathcal{A}$ with the same probability.

C. Evolutionary Process

1) Autonomous Learners: Autonomous learners perform iterative evolution within the reinforcement learning framework. The evolutionary process consists of three components: action selection, payoff calculation, and Q-table updating.

a) Action selection: At each round, autonomous learners first identify their current state based on the action taken in the previous round. They then apply the ϵ-greedy method for action selection to achieve a balance between exploration and exploitation. Specifically, agents have a probability of 1-ϵ to select the action with the maximum Q value in the current state from their Q-tables. For instance, if $Q^i_{C,C}(t) > Q^i_{C,D}(t)$ in (7), then at round t, agent i will decide to cooperate when in the cooperation state. In addition, agents choose an action at random with probability ϵ. In line with prior works [51], [52], we set the parameter ϵ to 0.02.

b) Payoff calculation: Unlike the multiplayer snowdrift game on conventional pairwise networks, the version on hypergraphs takes hyperedges—connecting multiple agents—as the basic interaction units. Each agent participates in the game within the hyperedges it belongs to. Thus, the total payoff of agent i can be represented by the following equation:

$$\pi_i(t) = \sum_{h=1}^{d_H(i)} \pi_i^h \tag{8}$$

$$\pi_i^h = \begin{cases} 1 - a_i r / x_C^h, & \text{for } x_C^h > 0 \\ 0, & \text{for } x_C^h = 0 \end{cases} \tag{9}$$

where $d_H(i)$ is agent i's hyperdegree, π_i^h denotes the payoff agent i receives in its hth hyperedge, a_i signifies the action (strategy) taken by agent i with $a_i = 1$ for cooperation and $a_i = 0$ for defection, and x_C^h indicates the count of cooperators in the h-th hyperedge of agent i. Clearly, the payoff of agent i is directly affected by the dilemma intensity r.

In particular, since agents are involved in different numbers of games, for the convenience of comparison, this study adopts the normalized payoff, as described in the following equation:

$$p_i(t) = \pi_i(t) / d_H(i). \tag{10}$$

c) Q-table updating: After obtaining payoffs, agents update their Q-tables. Specifically, if the state and action of agent i in round t are s and a, respectively, then the element $Q_{s,a}^i(t)$ is updated according to the following equation:

$$Q_{s,a}^i(t+1) = (1-\eta) Q_{s,a}^i(t) + \eta \left[p_i(t) + \delta \max_{a' \in \mathcal{A}} Q_{s',a'}^i(t) \right] \tag{11}$$

where η represents the learning rate, δ denotes the discount factor, $p_i(t)$ refers to agent i's payoff in round t, and $\max_{a' \in \mathcal{A}}(Q_{s',a'}^i(t))$ corresponds to the maximum Q value in the row of the next state s'. Consistent with previous studies [11], [51], we choose $\eta = 0.8$ and $\delta = 0.8$.

2) Imitation Learners: Imitation learners conduct iterative evolution under the imitation dynamics. The evolutionary process also encompasses three components: payoff calculation, neighbor payoff calculation, and action selection.

a) Payoff calculation: For imitation learner i, the payoff is calculated using the same method as autonomous learners, according to (8)–(10).

b) Neighbor payoff calculation: Agent i randomly selects a neighboring agent j as its model from those connected to i through hyperedges. The payoff for agent j is also computed using (8)–(10).

c) Action selection: After obtaining the payoffs of agents i and j, agent i adopts the action of agent j with probability F given by the Fermi rule. The probability F is related to the payoff difference between the two, as expressed in the following equation:

$$F\left(a_j(t) \to a_i(t+1)\right) = \frac{1}{1 + \exp\left[\left(p_i(t) - p_j(t)\right)/\kappa\right]} \tag{12}$$

where κ denotes the noise intensity, which characterizes the level of uncertainty or irrationality during the strategy updating process. When $\kappa \to 0$, the updating process becomes deterministic. Conversely, when $\kappa \to \infty$, agents are in a noisy environment where they can only update their strategies randomly. In accordance with previous studies [3], [18], we set the parameter $\kappa = 0.1$.

The evolutionary processes of autonomous and imitation learners are iterated through Monte Carlo simulations (MCSs), which consist of T steps. A complete Monte Carlo step comprises one iteration of N agents, that is, each agent, on average, has one opportunity to update its action (strategy) during a Monte Carlo step. When the system arrives at a stationary state ($T = 2 \times 10^5$), we observe the average rate of strategists on hypergraphs during the last 3000 steps. The architecture of this study is presented in Algorithm 1.

By analyzing Algorithm 1, it can be observed that in each time step, all N agents are, on average, selected once for processing. The time complexity for each selected agent depends on its type. For autonomous learners, assuming that the average hyperdegree of agents is $\langle d_H \rangle$, the time complexities of action selection, payoff calculation, and Q-table updating are $O(1)$, $O(\langle d_H \rangle)$, and $O(1)$, respectively. For imitation learners, the time complexities of payoff calculation, neighbor payoff calculation, and action selection are $O(\langle d_H \rangle)$, $O(\langle d_H \rangle)$, and $O(1)$, respectively. Therefore, the time complexity for each time step is $O(N \langle d_H \rangle)$. Since $\langle d_H \rangle \ll N$ can be regarded as a constant in this study and the algorithm contains T time steps, the total time complexity of the algorithm is $O(\mathrm{TN})$.

V. Results

A. Evolution of Agent Behaviors

We first examine how the steady-state cooperation rate f_c in higher-order networks depends on the cost-to-benefit ratio (dilemma intensity) r and the proportion of autonomous learners α, as shown in Fig. 4. It can be observed that autonomous learners, who adopt the reinforcement learning algorithm, have a dual effect on the evolution of cooperation. Specifically, at low dilemma intensity ($r < r_c$), f_c decreases as α rises; whereas at high dilemma intensity ($r > r_c$), f_c increases with α. Moreover, the presence of autonomous learners does not alter the overall trend of the network cooperation rate in relation to r. That is, regardless of the proportion of autonomous learners ($\alpha \in (0, 1)$), the cooperation rate always declines as the dilemma intensity increases. In the traditional case where imitation learners occupy the entire population ($\alpha = 0$), cooperators can survive only when r is relatively small (approximately $r < 0.7$). In contrast, the inclusion of autonomous learners ($0 < \alpha \leq 1$) allows cooperation to persist even at high values of r. And even under high dilemma intensity, when α is sufficiently large ($\alpha > 0.8$ approximately), cooperators always perform better than defectors ($f_c > 0.5$).

To better understand these differences, we examine the time series of the cooperation rate f_c in higher-order networks under varying dilemma intensities, as displayed in Fig. 5. Here, Fig. 5(a) depicts how the network cooperation rate evolves over time under different values of α at a low dilemma

Algorithm 1 Hybrid Strategy Updating Framework on Higher-Order Networks

Input: α, η, δ, ϵ, $\theta \sim \mathcal{U}(0, 1)$, game parameters, and the maximum Monte Carlo step T;

Generate a hypergraph $\mathcal{H}(\mathcal{X}, \mathcal{E})$ with N agents;
for *each agent* $i \in N$ **do**
  if $\theta < \alpha$ **then**
    Agent i is an autonomous learner (AL);
    Initialize i's state s_i to C or D at random;
    Initialize all elements of $\mathbf{Q}^i$ to zero;
  else
    Agent i is an imitation learner (IL);
    Initialize i's action a_i to C or D at random;
  end
end
for *time step* $t \in [1, T]$ **do**
  for $l < N$ **do**
    Select an agent i at random;
    if *Agent* i *is an AL* **then**
      if $\theta < \epsilon$ **then**
        Choose an action ξ randomly from $\mathcal{A}$;
        Set $a_i(t) = \xi$;
      else
        if $Q^i_{s_i,C}(t) > Q^i_{s_i,D}(t)$ **then**
          Set $a_i(t) = C$;
        else
          Set $a_i(t) = D$;
        end
      end
      Calculate the payoff $p_i(t)$ using 8–10;
      Set $s_i(t+1) = a_i(t)$;
      Update the Q-table using (11);
    else
      Calculate the payoff $p_i(t)$ using 8–10;
      Choose one neighbor j of agent i;
      Calculate $F(a_j(t) \to a_i(t+1))$;
      if $\theta < F(a_j(t) \to a_i(t+1))$ **then**
        Set $a_i(t+1) = a_j(t)$;
      end
    end
  end
end

intensity ($r = 0.1$). The results show that when α is small, the cooperation rate f_c exhibits a monotonic increasing trend; whereas when α is large, the cooperation rate f_c first rises, then decreases, and finally stabilizes. The reason for this phenomenon is that at the low dilemma intensity, the cooperation strategy is inherently advantageous. In this case, (i) when α is small, the majority of agents are imitation learners. Through imitation, defectors learn the cooperation strategy, thus forming a collective effect of cooperation. As a result, cooperative behaviors spread rapidly, and the cooperation rate increases monotonically. (ii) As α increases, the number of autonomous learners grows correspondingly, meaning that more agents update their strategies through the Q-learning algorithm rather than simple imitation. In the early stages of evolution, cooperation yields relatively high payoffs, resulting in an upward trend in the cooperation rate. However, due to the exploration ability of autonomous learners, over time they may realize that defection can bring higher short-term payoffs in certain situations (e.g., in a fully cooperative environment). This prompts some agents to switch to defection, causing a decline in the cooperation rate. In other words, the exploration feature of Q-learning may destroy the already formed cooperative behaviors in local environments, eventually leading to a decrease in the overall cooperation rate.

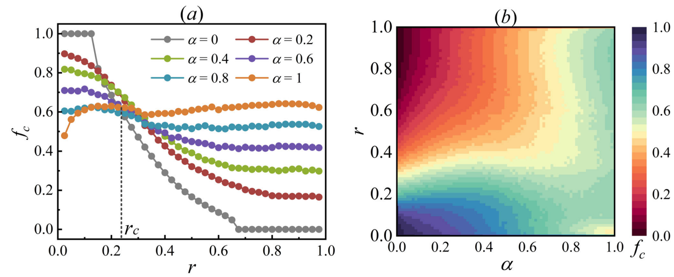


Fig. 4. (a) Line chart and (b) heat map illustrating the steady-state cooperation rate f_c with respect to the dilemma intensity r and the fraction of autonomous learners α. The results are averaged over ten realizations. It is obvious that autonomous learners exert a dual effect on the evolution of cooperation: at low dilemma intensity ($r < r_c$), the cooperation rate f_c decreases as α increases, whereas at high dilemma intensity ($r > r_c$), f_c increases with α. Parameters: $N = 2000$, $k = 3$, $m = 2$, $\eta = 0.8$, $\delta = 0.8$, $\epsilon = 0.02$, $\kappa = 0.1$, and $T = 2 \times 10^5$. Notably, the results presented here are also valid for different values of k.

Fig. 5(b) and (c) shows the temporal evolution of the network cooperation rate under different values of α at high dilemma intensities ($r = 0.5$ and $r = 0.7$). We observe that when α is small, the cooperation rate f_c shows a monotonically decreasing trend; whereas when α is large, the cooperation rate f_c first declines, then rises, and eventually stabilizes. This is because at high dilemma intensities, the temptation to defect is stronger. In this case, (i) when α is small, imitation learners dominate the population. The imitation effect prompts more agents to adopt the defection strategy, bringing about a drop in the cooperation rate. As α increases, the number of autonomous learners also grows, i.e., more agents use the Q-learning method to update their strategies. In the early stages of evolution, autonomous learners might try to defect. However, over time, through iterative learning and strategy adjustments, they gradually realize that while defection yields higher immediate payoffs, cooperation can maximize long-term payoffs in repeated interactions. In other words, the agents' capacity for autonomous learning helps reverse the proliferation of defective behaviors, resulting in an improved cooperation rate over time. In conclusion, the time series of evolution demonstrates that the exploration and autonomous learning abilities of autonomous learners using the Q-learning algorithm suppress cooperation at low dilemma intensities, but effectively promote cooperation at high dilemma intensities.

Furthermore, Fig. 6 depicts the equilibrium distribution of the number of cooperators, x_C, within each hyperedge (interaction group) for varying values of α at low ($r = 0.1$)

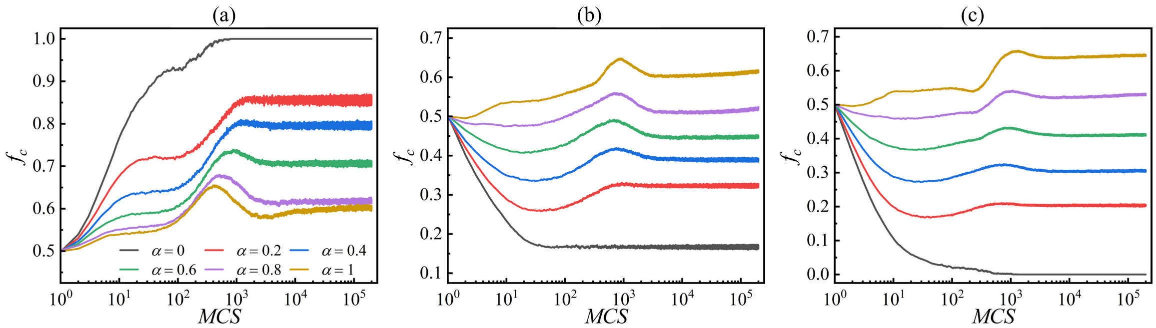


Fig. 5. Average cooperation rate as a function of time for various values of α at varying dilemma intensities. For panels (a)–(c), the dilemma intensity r corresponds to 0.1, 0.5, and 0.7, respectively. The evolutionary time series reveal that the exploration and autonomous learning capabilities inherent in the Q-learning algorithm suppress cooperation at low dilemma intensities but effectively promote cooperation at high dilemma intensities. Other parameters are the same as those in Fig. 4.

and high ($r = 0.7$) dilemma intensities. Since we adopt SRHs with $k = 3$, x_C takes values in the set $\{0, 1, 2, 3\}$. 1) When the dilemma intensity is low ($r = 0.1$) and α is small, hyperedges with $x_C = 3$ (i.e., fully cooperative hyperedges) account for a relatively high proportion. This is because, in this case, the cost of cooperation is low and most agents in the network are imitation learners, making it easier for cooperators to form tight clusters to resist the invasion of defectors. As α increases, the proportion of hyperedges with $x_C = 3$ declines, while the proportions of mixed hyperedges with $x_C = 1$ and $x_C = 2$ increase, resulting in a scenario where cooperators and defectors coexist. This is because a higher α introduces more autonomous learners into the network, who rely on the Q-learning algorithm to update their strategies rather than simply imitating their neighbors' cooperative behavior. Autonomous learners possess certain exploration ability, which disrupts previously stable cooperative clusters, leading to an increased existence of mixed hyperedges with $x_C = 1$ and $x_C = 2$. This phenomenon also suggests that the presence of Q-learning, to some extent, interferes with the already formed stable cooperative behaviors. 2) When the dilemma intensity is high ($r = 0.7$) and α is small, hyperedges with $x_C = 0$ (i.e., fully defective hyperedges) dominate the network. This is because the high cost of cooperation makes defection more profitable, and most agents in the network are imitation learners. After observing that defectors obtain higher payoffs, imitation learners mimic defective behavior, resulting in the rapid spread of defection. As α increases, the proportion of hyperedges with $x_C = 0$ decreases, while those with $x_C = 1$ rises; simultaneously, some hyperedges with $x_C = 2$ and $x_C = 3$ emerge. This is because the number of autonomous learners in the network who use Q-learning to make decisions grows with the increase in α. Although defection provides higher immediate returns under high dilemma intensity, autonomous learners gradually recognize the long-term advantages of cooperation through repeated interactions by iteratively updating their Q-tables. As a result, they adopt the cooperative strategy to avoid falling into the vicious cycle of mutual defection, leading to an increase in hyperedges with $x_C = 1$. This further demonstrates the adaptability of Q-learning in complex decision-making environments, enabling autonomous learners to learn cooperative strategy under specific conditions (e.g., when surrounded by defectors), to maximize long-term utility.

We further investigate how the endogenous parameters of the Q-learning algorithm—the discount factor δ and the learning rate η—affect the evolution of cooperation, as shown in Fig. 7. The results indicate that the cooperation rate f_c declines as the discount factor δ increases across various dilemma intensities. This trend arises from the role of δ in shaping the decision-making of autonomous learners: a higher δ places greater emphasis on long-term rewards, while a lower δ prioritizes immediate gains. During the evolutionary process, defection typically yields higher short-term payoffs, whereas cooperation requires repeated interactions to manifest its advantages (e.g., reciprocity). Thus, a higher δ encourages autonomous learners to adopt cooperative strategies that maximize long-term returns, thereby increasing the presence of cooperators within hyperedges. However, in the multiplayer snowdrift game, defectors benefit more from the presence of cooperators due to the "free-riding" effect. As imitation learners update their strategies using the Fermi rule, they are more likely to imitate neighbors with higher payoffs—often defectors. Thus, although a higher δ promotes cooperation among autonomous learners, it also indirectly boosts the relative success of defectors, leading imitation learners to favor defection. This negative feedback mechanism ultimately results in a decline in the overall cooperation rate as δ rises.

The impact of the learning rate η on the cooperation rate f_c depends on the dilemma intensity. Specifically, at low dilemma intensity, a lower learning rate tends to promote higher cooperation levels, whereas at high dilemma intensity, a higher learning rate is more conducive to the emergence of cooperation. This phenomenon can be attributed to the role of the learning rate η in controlling the speed of Q-table updates: a higher η allows agents to rapidly adjust their strategies based on new experiences, while a lower η makes them rely more heavily on past experiences. When the dilemma intensity is low, the temptation to defect is minimal, and agents with lower learning rates—who adapt more slowly—are more likely to retain cooperative strategies learned from past interactions, thereby sustaining higher cooperation levels. Conversely, under high dilemma intensity, higher learning rates enable agents to quickly adjust their strategies and form

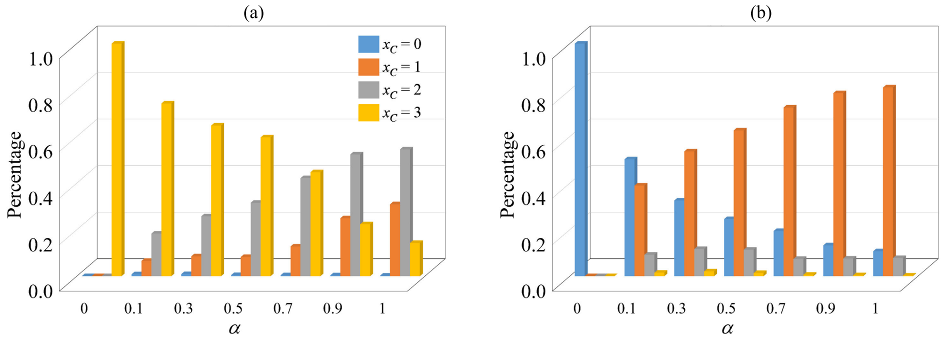


Fig. 6. Distribution of the number of cooperators, x_C, within each hyperedge (interaction group) at steady state for different values of α when (a) $r = 0.1$ and (b) $r = 0.7$. Since SRHs with $k = 3$ are used, $x_C \in \{0, 1, 2, 3\}$. Other parameters are the same as those in Fig. 4.

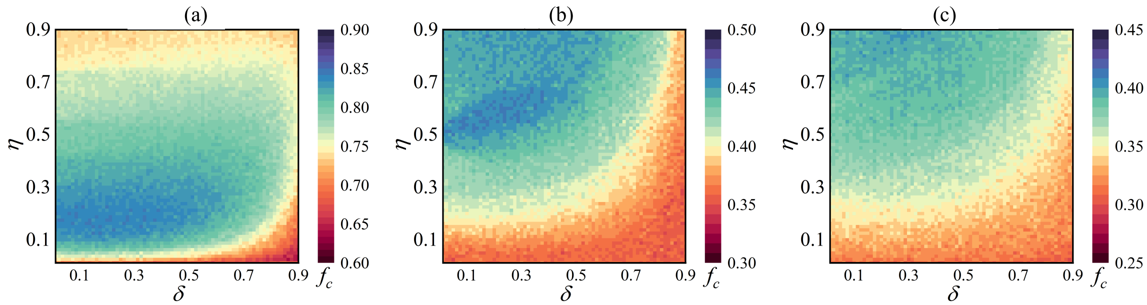


Fig. 7. Steady-state cooperation rate f_c as a function of the learning rate η and the discount factor δ for different dilemma intensities. As δ increases, f_c declines under various dilemma intensities, whereas the effect of η on f_c varies with the dilemma intensity. Other parameters are the same as those in Fig. 4. (a) $r = 0.1$. (b) $r = 0.5$. (c) $r = 0.7$.

cooperative clusters that can effectively resist defection—particularly in hypergraph structures, where cooperators tend to group together.

B. Impact of Group Size

Next, we investigate the influence of group size on cooperative evolution using SRHs, where each hyperedge consists of k nodes, representing a group size of k. Fig. 8 shows how the steady-state cooperation rate varies with group size on SRHs, as k ranges from 3 to 9. The results indicate that under different dilemma intensities, the network cooperation rate declines as the group size increases. This is because of the fact that in smaller groups, each agent's contribution has a greater influence on the collective performance of the group, thus motivating agents to cooperate. Conversely, in larger groups, the impact of an agent's contribution weakens, rendering some agents to choose noncooperative strategy, namely, "free-riding." This trend aligns with observations in the real world. For example, in smaller groups, communication and information exchange among members are often more efficient, promoting the rapid development of trust and reciprocity [57]. However, as group size expands, communication costs also rise. It is difficult for individuals to fully understand the behavior of other members, thereby reducing their expectations and confidence in cooperation, and "free-riding" behavior is more likely to occur.

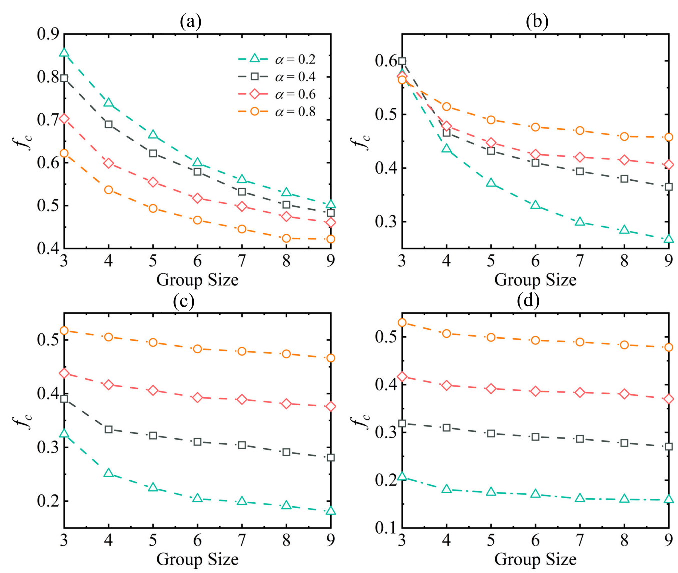


Fig. 8. Equilibrium cooperation rate f_c in relation to group size for varying dilemma intensities on SRHs. In SRHs, every hyperedge consists of k nodes, corresponding to a group size of k. From panel (a) to (d), the values of dilemma intensity r are 0.1, 0.3, 0.5, and 0.7, respectively. Results show that smaller group sizes are more favorable for cooperation. Other parameters are the same as those in Fig. 4.

To better understand why cooperation is more prevalent in smaller groups, we analyze the average payoffs of cooperators and defectors as a function of group size, as shown in Fig. 9. The results reveal that under low dilemma intensity ($r = 0.1, 0.3, 0.5$), cooperators achieve higher average payoffs than defectors in smaller groups. However, this advantage gradually diminishes and eventually reverses as group size

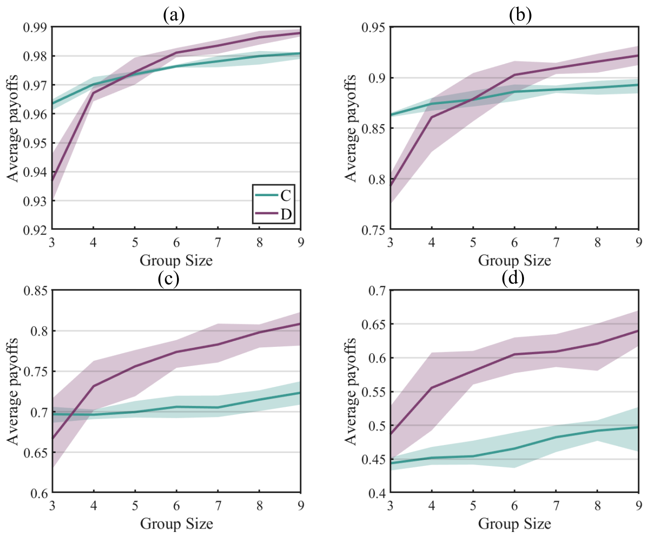


Fig. 9. Steady-state average payoffs of cooperators and defectors with respect to group size. The solid lines denote the mean values, while the shaded areas represent the min–max range over 50 independent runs. (a) $r = 0.1$. (b) $r = 0.3$. (c) $r = 0.5$. (d) $r = 0.7$.

increases, with the payoff gap continuing to widen. This is because, in smaller groups, cooperators have a greater marginal contribution—their behaviors have a stronger impact on the group—making them more likely to receive positive payoff feedback. As a result, imitation learners, observing the success of cooperators, are more inclined to adopt cooperative strategies, and autonomous learners are also more likely to reinforce cooperative behaviors through Q-learning. In contrast, in larger groups, individual impact on the group diminishes, thereby weakening the incentives for cooperation and ultimately leading to a lower cooperation rate.

C. Theoretical Analysis

To validate the above numerical simulation results, we conduct an approximation analysis using the strategy transition probability and the mean-field method. Given that agents in the system are classified into two types, autonomous learners (proportion α) and imitation learners (proportion $1-\alpha$), we define the corresponding strategy transition probabilities for each type. For autonomous learners, $W^{(A)}_{C\to D}$ and $W^{(A)}_{D\to C}$ represent the probabilities of transitioning from cooperation to defection and from defection to cooperation, respectively, during the evolutionary process. Similarly, for imitation learners, $W^{(I)}_{C\to D}$ and $W^{(I)}_{D\to C}$ denote the corresponding transition probabilities. These probabilities are calculated as follows.

1) Autonomous Learners: Autonomous learners update their strategies according to the Q-learning algorithm, with transition probabilities determined by the ϵ-greedy policy. Take $W^{(A)}_{C\to D}$ as an example

$$W^{(A)}_{C\to D} = \frac{1}{2}\epsilon + (1-\epsilon)\, I\,(Q_D > Q_C) \tag{13}$$

where ϵ is the exploration rate, representing the probability that the agent selects an action—cooperation or defection—at random (i.e., random decision-making). $1/2\epsilon$ reflects the probability of switching from cooperation to defection under random decision-making. Conversely, $1-\epsilon$ denotes the probability that the agent chooses the action with the highest Q value (i.e., rational decision-making). The indicator function $I(\cdot)$ returns 1 if $Q_D > Q_C$ and 0 otherwise. Thus, $(1-\epsilon)I(Q_D > Q_C)$ corresponds to the probability of transitioning from cooperation to defection under rational decision-making. Similarly, $W^{(A)}_{D\to C}$ is defined as

$$W^{(A)}_{D\to C} = \frac{1}{2}\epsilon + (1-\epsilon)\, I\,(Q_C > Q_D)\,. \tag{14}$$

2) Imitation Learners: Imitation learners update their strategies based on the Fermi rule, where the transition probabilities are determined by the payoff difference between the focal agent and its neighbors. The probabilities $W^{(I)}_{C\to D}$ and $W^{(I)}_{D\to C}$ are given by

$$W^{(I)}_{C\to D} = \frac{1}{1+\exp\left(\frac{\pi_C-\pi_D}{\kappa}\right)} \tag{15}$$

$$W^{(I)}_{D\to C} = \frac{1}{1+\exp\left(\frac{\pi_D-\pi_C}{\kappa}\right)} \tag{16}$$

where π_C and π_D represent the payoffs received by cooperators and defectors, respectively, and κ is the noise intensity.

Based on the above equations, we can calculate the aggregate probability $W_{C\to D}$ ($W_{D\to C}$) that agents collectively transition from cooperation (defection) to defection (cooperation) during the evolutionary process

$$W_{C\to D} = \alpha W^{(A)}_{C\to D} + (1-\alpha)\, W^{(I)}_{C\to D} \tag{17}$$

$$W_{D\to C} = \alpha W^{(A)}_{D\to C} + (1-\alpha)\, W^{(I)}_{D\to C}. \tag{18}$$

According to the mean-field method, the rate of change in the system's cooperation level f_c can be approximately expressed by the following equation:

$$\dot{f}_c = (1-f_c)\, W_{D\to C} - f_c W_{C\to D}. \tag{19}$$

As the system reaches equilibrium, $\dot{f}c = 0$, namely, $(1-f_c)W_{D\to C} - f_c W_{C\to D} = 0$. Accordingly, the cooperation level f_c of the system in the steady state can be presented as the following equation:

$$f_c = \frac{W_{D\to C}}{W_{D\to C} + W_{C\to D}} = 1\Big/\left(1+\frac{W_{C\to D}}{W_{D\to C}}\right). \tag{20}$$

Based on the above analysis, we present the theoretical and numerical results for the cooperation rate, as demonstrated in Fig. 10. The strong agreement between the two demonstrates that the theoretical model effectively captures the distinct strategy updating dynamics of autonomous and imitation learners.

D. Robustness of Results

To evaluate the robustness of our findings under varying exploration rates, we perform sensitivity experiments on the parameter ϵ. Specifically, following prior studies [58], [59], we set $\epsilon \in \{0.02, 0.1, 0.2, 0.4\}$, and the corresponding results are presented in Fig. 11. Although changes in the exploration rate do have some impacts on the cooperation level, the overall

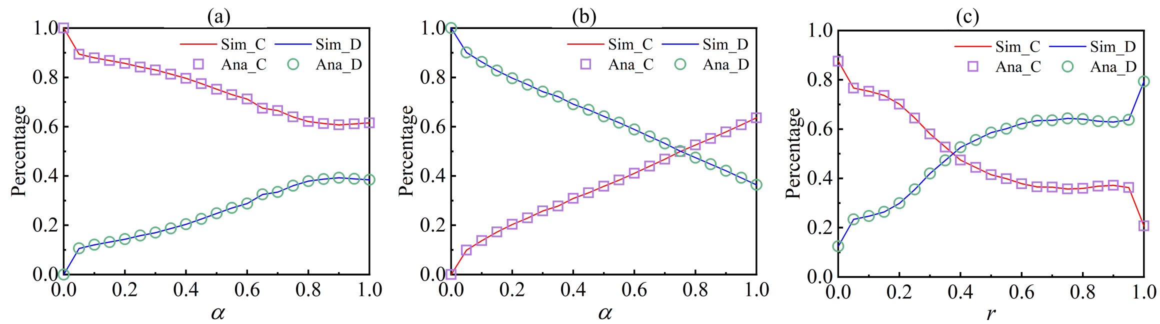


Fig. 10. Percentage of cooperators and defectors depending on (a) α under low dilemma intensity ($r = 0.1$), (b) α under high dilemma intensity ($r = 0.7$), and (c) r when $\alpha = 0.5$. Hollow squares and circles denote numerical results, while solid lines indicate the theoretical predictions. Other parameters are the same as those in Fig. 4.

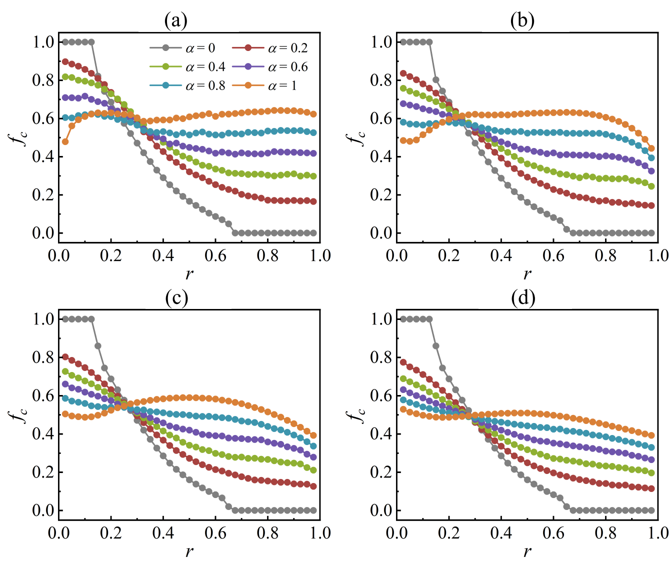


Fig. 11. Steady-state cooperation rate f_c as a function of the dilemma intensity r and the proportion of autonomous learners α for different exploration rates. (a) $\epsilon = 0.02$, (b) $\epsilon = 0.1$, (c) $\epsilon = 0.2$, and (d) $\epsilon = 0.4$. These results demonstrate that our main conclusions remain robust across varying exploration rates. Other parameters are the same as those in Fig. 4.

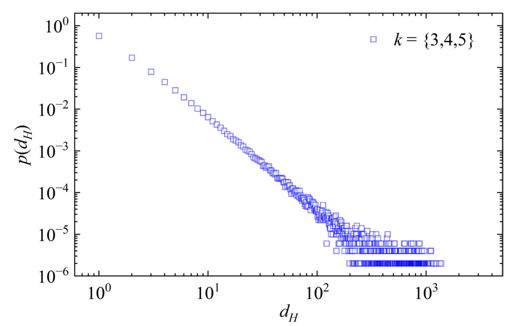


Fig. 12. Hyperdegree distributions of MSRHs with $N = 10\,000$ and $k = \{3, 4, 5\}$ on a logarithmic scale. Results are averaged over 50 independent realizations.

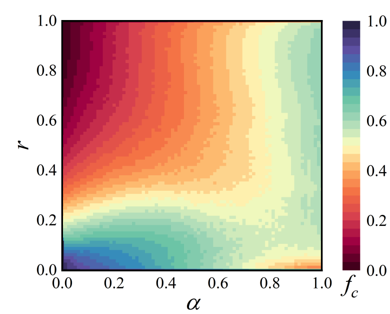


Fig. 13. Steady-state cooperation rate f_c with respect to the dilemma intensity r and the proportion of autonomous learners α on MSRHs. It can be seen that the results obtained on MSRHs are consistent with those on SRHs. Other parameters are the same as those in Fig. 4.

trend remains consistent: reinforcement learning promotes cooperation under high dilemma intensity but suppresses cooperation under low dilemma intensity. This demonstrates that our main conclusions are robust across different exploration abilities. In addition, we also validate our findings using alternative reinforcement learning algorithms—SARSA [60] and BM [61]—to model autonomous learners' decision-making. The results from these alternative algorithms are consistent with those obtained using the Q-learning method, further supporting the reliability of our conclusions.

The underlying network topology also plays a crucial role in the evolution of group cooperation. To test the robustness of our findings across different network topologies, we therefore introduce member heterogeneous SRHs (MSRHs) and uniform random hypergraphs (URHs). Specifically, 1) Similar to SRHs, MSRHs exhibit a power-law hyperdegree distribution; however, the orders of hyperedges are heterogeneous and follow a uniform distribution $p(k) \sim (1/n)$, where n is the total number of k can take. Following the method described in Ref. [62], we generate MSRHs with $k = \{3, 4, 5\}$ and $n = 3$. The resulting hyperdegree distribution is plotted in Fig. 12, and the experimental results on MSRHs are reported in Fig. 13. 2) For URHs, each hyperedge contains the same number of nodes, but the hyperdegree spectrum follows a normal distribution. According to the rules outlined in [10], we construct URHs with $k = 3$, 5, and 7, respectively, whose hyperdegree distribution is displayed in Fig. 14, and the experimental results on URHs are shown in Fig. 15. It can be observed that the results obtained on both MSRHs and URHs are consistent with those on SRHs. These findings demonstrate the robustness of our conclusions across different types of hypergraph topologies.

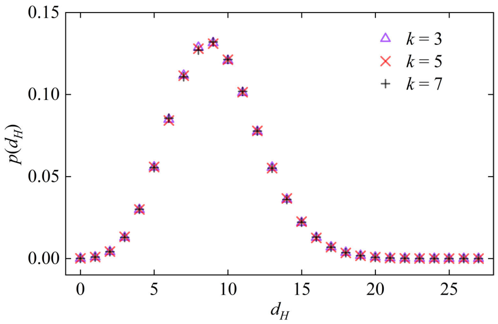


Fig. 14. Hyperdegree distributions of URHs with $N = 10\,000$ and $k = 3$, 5, and 7, respectively. Results are averaged over 50 independent realizations.

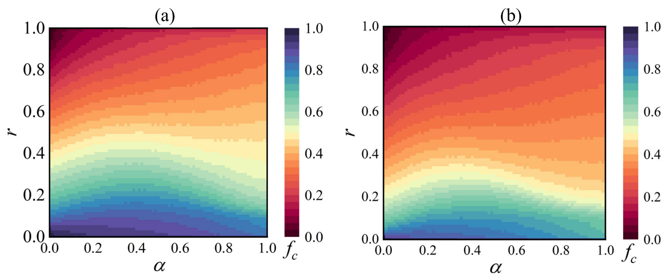


Fig. 15. Steady-state cooperation rate f_c as a function of the dilemma intensity r and the proportion of autonomous learners α on URHs. It can be observed that the results obtained on URHs are consistent with those on SRHs. Other parameters are the same as those in Fig. 4. (a) $k = 3$. (b) $k = 4$.

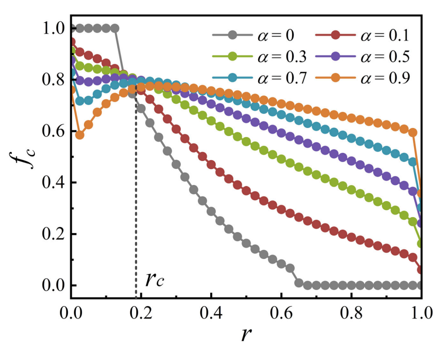


Fig. 16. Average cooperation rate f_c varying with the dilemma intensity r for different values of α. In the above studies, the types of agents in the network remain unchanged throughout the evolutionary process. Here, we break this limitation: in each MCS, each agent becomes an autonomous learner with probability α, and an imitation learner with probability $1 - \alpha$. Interestingly, we find that whether imitation and autonomous learners can convert does not affect the overall experimental results. Other parameters are the same as those in Fig. 4.

Notably, in the above studies, the types of agents in the network—imitation learners and autonomous learners—are fixed and remain unchanged throughout the evolutionary process. Here, we break this limitation: In each MCS, each agent has a probability α of acting as an autonomous learner and a probability $1 - \alpha$ of acting as an imitation learner. With this setup, we further analyze the impact of dynamic transitions between imitation and autonomous learners on cooperative behaviors, as illustrated in Fig. 16. It is interesting to observe that whether imitation and autonomous learners can convert does not affect the overall experimental results.

VI. Conclusion and Discussion

Collective cooperation is a widespread phenomenon; while seemingly contradicting Darwinian theory, it is considered the cornerstone of the development of human society [63]. Higher-order networks, represented by hypergraphs, can naturally describe structured group interactions, becoming a powerful tool for exploring the emergence of collective cooperation [13], [64]. In real-world scenarios, individuals exhibit diverse behavioral patterns, reflecting significant differences in their decision-making processes. This heterogeneity not only affects the interaction mode among individuals but also alters the propagation features of cooperative behaviors within the population. Therefore, investigating the hybrid strategy updating mechanism based on higher-order networks is essential for comprehending the emergence and maintenance of cooperative behaviors.

In this study, we categorize agents into imitation and autonomous learners based on social learning and behaviorism theories, and then develop a hybrid strategy updating mechanism based on hypergraphs to examine the influence of different types of agents on the evolutionary dynamics of group cooperation. This framework uses the Q-learning algorithm to model autonomous learners, highlighting their independent decision-making characteristics, and uses the Fermi rule to represent imitation learners, characterizing their probability-based imitation behaviors. Through a series of numerical simulations, we observe that autonomous learners, compared with imitation learners, exert a dual effect on group cooperation: they inhibit cooperation under low dilemma intensity but promote cooperation under high dilemma intensity. In addition, we examine the role of group size in cooperative evolution within higher-order networks under the hybrid updating mechanism, revealing that smaller group sizes are more conducive to the formation of cooperation. Finally, we theoretically substantiate the validity of our numerical simulation results using the strategy transition probability and the mean-field method. These findings provide valuable insights into understanding how the hybrid strategy updating mechanism affects cooperative evolution on higher-order networks within social dilemmas.

This study primarily investigates the mechanisms by which the hybrid updating framework influences group cooperation. Although we validate the robustness of different updating rules under probability changes during the evolutionary process, individuals in real-world settings continuously adjust their behavioral patterns based on personal experience, knowledge, and environmental factors [65], [66]. In other words, individual decision-making approaches (i.e., strategy updating rules) evolve dynamically, rather than being simply determined by probabilistic conditions. Therefore, it is important to model the evolution of individual decision-making patterns and subsequently examine their impact on group cooperation. In addition, future work can also attempt to explore how the hybrid strategy updating mechanism influences the evolution of collective cooperation in dynamic, higher-order environments [67], [68]. Overall, our research provides a new framework for exploring the impact of the hybrid strategy updating rule on the evolution and emergence of cooperation in higher-order networks, laying the foundation for studies that align more closely with real-world scenarios.

References

[1] R. Rider, "The evolution of cooperation," *Science*, vol. 5, nos. 3–4, pp. 406–409, 1984.

[2] M. Nowak and R. Highfield, *Supercooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. New York, NY, USA: Simon and Schuster, 2011.

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

[4] S. Boccaletti et al., "The structure and dynamics of networks with higher order interactions," *Phys. Rep.*, vol. 1018, pp. 1–64, May 2023.

[5] F. Battiston et al., "The physics of higher-order interactions in complex systems," *Nature Phys.*, vol. 17, no. 10, pp. 1093–1098, 2021.

[6] F. Battiston et al., "Networks beyond pairwise interactions: Structure and dynamics," *Phys. Rep.*, vol. 874, pp. 1–92, Aug. 2020.

[7] S. Majhi, M. Perc, and D. Ghosh, "Dynamics on higher-order networks: A review," *J. Roy. Soc. Interface*, vol. 19, no. 188, Mar. 2022, Art. no. 20220043.

[8] J. Grilli, G. Barabás, M. J. Michalska-Smith, and S. Allesina, "Higher-order interactions stabilize dynamics in competitive network models," *Nature*, vol. 548, no. 7666, pp. 210–213, Aug. 2017.

[9] I. Iacopini, G. Petri, A. Barrat, and V. Latora, "Simplicial models of social contagion," *Nature Commun.*, vol. 10, no. 1, p. 2485, Jun. 2019.

[10] U. Alvarez-Rodriguez, F. Battiston, G. F. de Arruda, Y. Moreno, M. Perc, and V. Latora, "Evolutionary dynamics of higher-order interactions in social networks," *Nature Hum. Behaviour*, vol. 5, no. 5, pp. 586–595, Jan. 2021.

[11] Y. Xu et al., "Reinforcement learning and collective cooperation on higher-order networks," *Knowl.-Based Syst.*, vol. 301, Oct. 2024, Art. no. 112326.

[12] F. Malizia, A. Corso, L. V. Gambuzza, G. Russo, V. Latora, and M. Frasca, "Reconstructing higher-order interactions in coupled dynamical systems," *Nature Commun.*, vol. 15, no. 1, p. 5184, Jun. 2024.

[13] A. Civilini, O. Sadekar, F. Battiston, J. Gómez-Gardeñes, and V. Latora, "Explosive cooperation in social dilemmas on higher-order networks," *Phys. Rev. Lett.*, vol. 132, no. 16, Apr. 2024, Art. no. 167401.

[14] A. Govaert, P. Ramazi, and M. Cao, "Rationality, imitation, and rational imitation in spatial public goods games," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 3, pp. 1324–1335, Sep. 2021.

[15] G. Como, F. Fagnani, and L. Zino, "Imitation dynamics in population games on community networks," *IEEE Trans. Control Netw. Syst.*, vol. 8, no. 1, pp. 65–76, Mar. 2021.

[16] A. Sheng, Q. Su, L. Wang, and J. B. Plotkin, "Strategy evolution on higher-order networks," *Nature Comput. Sci.*, vol. 4, no. 4, pp. 274–284, Apr. 2024.

[17] A. Civilini, N. Anbarci, and V. Latora, "Evolutionary game model of group choice dilemmas on hypergraphs," *Phys. Rev. Lett.*, vol. 127, no. 26, Dec. 2021, Art. no. 268301.

[18] A. Kumar, S. Chowdhary, V. Capraro, and M. Perc, "Evolution of honesty in higher-order social networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 104, no. 5, Nov. 2021, Art. no. 054308.

[19] Y.-J. Ma, Z.-Q. Jiang, F.-S. Fang, M. Perc, and S. Boccaletti, "Social norms and cooperation in higher-order networks," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 480, no. 2293, Jul. 2024, Art. no. 20240066.

[20] M. S. Anwar, G. K. Sar, M. Perc, and D. Ghosh, "Collective dynamics of swarmalators with higher-order interactions," *Commun. Phys.*, vol. 7, no. 1, p. 59, Feb. 2024.

[21] W. Wang et al., "Epidemic spreading on higher-order networks," *Phys. Rep.*, vol. 1056, pp. 1–70, Mar. 2024.

[22] L. Neuhäuser, A. Mellor, and R. Lambiotte, "Multibody interactions and nonlinear consensus dynamics on networked systems," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 101, no. 3, Mar. 2020, Art. no. 032310.

[23] P. K. Pal, M. S. Anwar, M. Perc, and D. Ghosh, "Global synchronization in generalized multilayer higher-order networks," *Phys. Rev. Res.*, vol. 6, no. 3, Jul. 2024, Art. no. 033003.

[24] R. Ghorbanchian, J. G. Restrepo, J. J. Torres, and G. Bianconi, "Higher-order simplicial synchronization of coupled topological signals," *Commun. Phys.*, vol. 4, no. 1, p. 120, Jun. 2021.

[25] S. Dutta, P. Kundu, P. Khanra, C. Hens, and P. Pal, "Perfect synchronization in complex networks with higher-order interactions," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 108, no. 2, Aug. 2023, Art. no. 024304.

[26] H. Ohtsuki, C. Hauert, E. Lieberman, and M. A. Nowak, "A simple rule for the evolution of cooperation on graphs and social networks," *Nature*, vol. 441, no. 7092, pp. 502–505, May 2006.

[27] G. Szabó and G. Fáth, "Evolutionary games on graphs," *Phys. Rep.*, vol. 446, nos. 4–6, pp. 97–216, Jul. 2007.

[28] P. Ramazi, J. Riehl, and M. Cao, "The lower convergence tendency of imitators compared to best responders," *Automatica*, vol. 139, May 2022, Art. no. 110185.

[29] H. Ohtsuki and M. A. Nowak, "The replicator equation on graphs," *J. Theor. Biol.*, vol. 243, no. 1, pp. 86–97, Nov. 2006.

[30] F. Zhang, J. Wang, H. Gao, X. Li, and C. Xia, "Role of strategy update rules in the spatial memory-based mixed strategy games," *Eur. Phys. J. B*, vol. 94, no. 1, p. 22, Jan. 2021.

[31] J. Shi, C. Liu, and J. Liu, "Hypergraph-based model for modeling multi-agent Q-learning dynamics in public goods games," *IEEE Trans. Netw. Sci. Eng.*, vol. 11, no. 6, pp. 6169–6179, Nov. 2024.

[32] B.-Y. Li, Z.-N. Zhang, G.-Z. Zheng, C.-R. Cai, J.-Q. Zhang, and L. Chen, "Cooperation in public goods games: Leveraging other-regarding reinforcement learning on hypergraphs," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 111, no. 1, Jan. 2025, Art. no. 014304.

[33] A. Bandura, *Social Learning Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1977.

[34] B. Skinner, "Is it behaviorism?" *Behav. Brain Sci.*, vol. 9, no. 4, p. 716, 1986.

[35] M. Feng, B. Pi, L.-J. Deng, and J. Kurths, "An evolutionary game with the game transitions based on the Markov process," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 54, no. 1, pp. 609–621, Jan. 2024.

[36] H. Guo, Z. Song, M. Perc, X. Li, and Z. Wang, "Third-party intervention of cooperation in multilayer networks," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 53, no. 11, pp. 6646–6657, Nov. 2023.

[37] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, US: MIT Press, 2018.

[38] Y. Tao, K. Hu, P. Wang, X. Zhao, and L. Shi, "A double-edged sword: Diverse interactions in hypergraphs," *New J. Phys.*, vol. 26, no. 9, Sep. 2024, Art. no. 093021.

[39] J. Wang, J. Nie, S. Guo, M. Özer, C. Xia, and M. Perc, "Mixing prisoner's dilemma games on higher-order networks," *Neurocomputing*, vol. 607, Nov. 2024, Art. no. 128439.

[40] Y. Lu et al., "Understanding the role of neutral species by means of high-order interaction in the rock-paper-scissors dynamics," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 109, no. 1, Jan. 2024, Art. no. 014313.

[41] D. Schlager, K. Clauß, and C. Kuehn, "Stability analysis of multiplayer games on adaptive simplicial complexes," *Chaos, Interdiscipl. J. Nonlinear Sci.*, vol. 32, no. 5, May 2022, Art. no. 053128.

[42] J. Pan, L. Zhang, W. Han, and C. Huang, "Heterogeneous investment promotes cooperation in spatial public goods game on hypergraphs," *Phys. A, Stat. Mech. Appl.*, vol. 609, Jan. 2023, Art. no. 128400.

[43] W.-Y. Chen, J.-C. Pan, W.-C. Han, and C.-W. Huang, "Evolutionary public goods games on hypergraphs with heterogeneous multiplication factors," *Acta Phys. Sinica*, vol. 71, no. 11, 2022, Art. no. 110201.

[44] C. Ma, A. Li, Y. Du, H. Dong, and Y. Yang, "Efficient and scalable reinforcement learning for large-scale network control," *Nature Mach. Intell.*, vol. 6, no. 9, pp. 1006–1020, Sep. 2024.

[45] B. Pi, L.-J. Deng, M. Feng, M. Perc, and J. Kurths, "Dynamic evolution of complex networks: A reinforcement learning approach applying evolutionary games to community structure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 10, pp. 8563–8582, Oct. 2025.

[46] X. Shi, Y. Li, C. Du, C. Chen, G. Zong, and W. Gui, "Reinforcement learning-based optimal control for Markov jump systems with completely unknown dynamics," *Automatica*, vol. 171, Jan. 2025, Art. no. 111886.

[47] S. Lu and Y. Wang, "Past-performance-driven strategy updating promote cooperation in the spatial prisoner's dilemma game," *Appl. Math. Comput.*, vol. 491, Apr. 2025, Art. no. 129220.

[48] Z. Song, H. Guo, D. Jia, M. Perc, X. Li, and Z. Wang, "Reinforcement learning facilitates an optimal interaction intensity for cooperation," *Neurocomputing*, vol. 513, pp. 104–113, Nov. 2022.

[49] S.-P. Zhang, J.-Q. Zhang, L. Chen, and X.-D. Liu, "Oscillatory evolution of collective behavior in evolutionary games played with reinforcement learning," *Nonlinear Dyn.*, vol. 99, no. 4, pp. 3301–3312, Mar. 2020.

[50] Y. Usui and M. Ueda, "Symmetric equilibrium of multi-agent reinforcement learning in repeated prisoner's dilemma," *Appl. Math. Comput.*, vol. 409, Nov. 2021, Art. no. 126370.

[51] L. Wang et al., "Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning," *Nonlinear Dyn.*, vol. 108, no. 2, pp. 1837–1845, Apr. 2022.

[52] H. Guo, Z. Wang, Z. Song, Y. Yuan, X. Deng, and X. Li, "Effect of state transition triggered by reinforcement learning in evolutionary prisoner's dilemma game," *Neurocomputing*, vol. 511, pp. 187–197, Oct. 2022.

[53] R. Chiong and M. Kirley, "Effects of iterated interactions in multiplayer spatial evolutionary games," *IEEE Trans. Evol. Comput.*, vol. 16, no. 4, pp. 537–555, Aug. 2012.

[54] T. Gradowski and A. Krawiecki, "Majority-vote model on scale-free hypergraphs," *Acta Phys. Polonica A*, vol. 127, no. 3a, pp. A-55–A-58, Mar. 2015.

[55] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[56] A. K. Sadhu and A. Konar, "An efficient computing of correlated equilibrium for cooperative Q-learning-based multi-robot planning," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 50, no. 8, pp. 2779–2794, Aug. 2020.

[57] C. Xia, J. Wang, M. Perc, and Z. Wang, "Reputation and reciprocity," *Phys. Life Rev.*, vol. 46, pp. 8–45, Sep. 2023.

[58] P. Dai, W. Yu, and D. Chen, "Distributed Q-learning algorithm for dynamic resource allocation with unknown objective functions and application to microgrid," *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12340–12350, Nov. 2022.

[59] Z. Li, S. Pan, and Y. Qin, "Multiuser scheduling algorithm for 5G IoT systems based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4643–4653, Apr. 2023.

[60] A. Y. Majid, S. Saaybi, V. Francois-Lavet, R. V. Prasad, and C. Verhoeven, "Deep reinforcement learning versus evolution strategies: A comparative survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 9, pp. 11939–11957, Sep. 2024.

[61] T. Ezaki, Y. Horita, M. Takezawa, and N. Masuda, "Reinforcement learning explains conditional cooperation and its moody cousin," *PLOS Comput. Biol.*, vol. 12, no. 7, Jul. 2016, Art. no. e1005034.

[62] Z. Wang, R. Li, X. Jin, and H. Ding, "Emergence of social norms in metanorms game with high-order interaction topology," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 3, pp. 1057–1072, Jun. 2023.

[63] Z. Zeng, M. Feng, P. Liu, and J. Kurths, "Complex network modeling with power-law activating patterns and its evolutionary dynamics," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 55, no. 4, pp. 2546–2559, Apr. 2025.

[64] J. Li, X. Wu, J. Lü, and L. Lei, "Enhancing predictive accuracy in social contagion dynamics via directed hypergraph structures," *Commun. Phys.*, vol. 7, no. 1, p. 129, Apr. 2024.

[65] G. Gigerenzer and P. M. Todd, *Simple Heuristics That Make us Smart*. London, U.K.: Oxford Univ. Press, 1999.

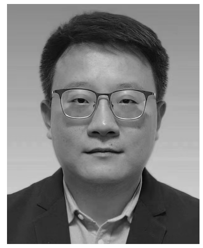
[66] W. Wood and D. T. Neal, "A new look at habits and the habit-goal interface," *Psychol. Rev.*, vol. 114, no. 4, pp. 843–863, 2007.

[67] S. Chowdhary, A. Kumar, G. Cencetti, I. Iacopini, and F. Battiston, "Simplicial contagion in temporal higher-order networks," *J. Phys., Complex.*, vol. 2, no. 3, Sep. 2021, Art. no. 035019.

[68] Y. Xu, J. Wang, C. Xia, and Z. Wang, "Higher-order temporal interactions promote the cooperation in the multiplayer snowdrift game," *Sci. China Inf. Sci.*, vol. 66, no. 12, Dec. 2023, Art. no. 222208.

Yan Xu received the Ph.D. degree in computer science from Tianjin University of Technology, Tianjin, China, in 2025.

She is currently a Lecturer with Tianjin Police Institute, Tianjin. Her current research interests include network security, complex network, and evolutionary game theory.



Dawei Zhao (Member, IEEE) received the Ph.D. degree in cryptology from Beijing University of Posts and Telecommunications, Beijing, China, in 2014.

He is currently a Professor with Shandong Computer Science Center (National Supercomputer Center in Jinan), Jinan, China. His main research interests include network security, complex network, and epidemic spreading dynamics.

Tina Perc Benko received the B.Sc. degree in biology from the University of Maribor, Maribor, Slovenia, in 2005, where she is currently pursuing the Ph.D. degree in theoretical biology.

She is currently a Research Assistant with the University of Maribor. Her current research interests are evolutionary game theory and complex systems.

Chengyi Xia (Member, IEEE) received the Ph.D. degree in control theory and control engineering from Nankai University, Tianjin, China, in 2008.

At present, he is a Full Professor with the School of Artificial Intelligence, Tiangong University, Tianjin. He has published more than 70 articles in the peer-reviewed journals. His current research interests include complex systems modeling and analysis, network science, cyber security, epidemic propagation, and evolutionary game theory.

Matjaž Perc is a member of Academia Europaea and European Academy of Sciences and Arts, and among top 1% most cited physicists according to 2020, 2021, 2022, and 2023 Clarivate Analytics data. He is also the 2015 recipient of the Young Scientist Award for Socio and Econophysics from the German Physical Society, and the 2017 USERN Laureate. In 2018, he received the Zois Award, which is the highest national research award in Slovenia. In 2019, he became the fellow of the American Physical Society. Since 2021, he has been a Vice Dean of Natural Sciences with European Academy of Sciences and Arts.