The evolution of cooperation with Q-learning: the impact of information perception

Guozhong Zheng,[1] Zhenwei Ding,[2] Jiqiang Zhang,[2] Shengfeng Deng,[1] Weiran Cai,[3] and Li Chen[1,*]

[1]*School of Physics and Information Technology, Shaanxi Normal University, Xi'an 710061, P. R. China*
[2]*School of Physics, Ningxia University, Yinchuan 750021, P. R. China*
[3]*School of Computer Science, Soochow University, Suzhou 215006, P. R. China*

(Dated: July 30, 2024)

The inherent huge complexities in human beings show a remarkable diversity in response to complex surroundings, enabling us to tackle problems from different perspectives. In the realm of cooperation studies, however, existing work assumes that individuals get access to the same kind of information to make their decisions, in contrast to the facts that individuals often perceive differently. Here, within the reinforcement learning framework, we investigate the impact of information perception on the evolution of cooperation in a 2-person scenario when playing the prisoner's dilemma game. We demonstrate that distinctly different evolution processes are observed in three information perception scenarios, revealing that the structure of information significantly affects the emergence of cooperation. Notably, the asymmetric information scenario exhibits a rich dynamical process, including the cooperation emergence, breakdown, and reconstruction, akin to psychological changes in humans. Our findings indicate that the information structure is vital to the emergence of cooperation, shedding new light on establishing mutually stable cooperative relationships and understanding human behavioral complexities in general.

1. INTRODUCTION

Cooperation is the basis for the survival, development, and reproduction of human beings, and it is ubiquitous in human societies as well as for other species [1–3]. A robust cooperation between individuals effectively improves the work efficiency and benefits the collective. However, due to its complexity and subtlety, non-cooperation frequently occurs, such as global warming, overfishing, deforestation, grazing, war conflicts, and so on, which could be catastrophic. So how cooperation emerges, and under what condition it breaks down? The emergence and maintenance of cooperation have consistently been fundamental challenges for humanity [4].

Evolutionary game theory [5, 6] has served as the main framework for the study of cooperation, with prototypical models being adopted such as the prisoner's dilemma (PD) game [7]. The PD game shows that even when cooperation is beneficial for both parties as a collective, maintaining cooperation remains difficult. This difficulty arises from the fact that individuals seek to maximize their own interests and choose defection, leading to the tragedy of commons [8]. Thus, there must be some mechanisms that are able to overcome this dilemma to foster cooperation.

Several mechanisms for the emergence of cooperation have been proposed in the past decades [9, 10], including direct [11] and indirect reciprocity [12], kin and group selection [13], punishment and reward [14], network [15–17] and dynamical reciprocity [18], social diversity [19–21], reputation [22], and behavioral multimodality [23] etc. Note that these game-theoretic studies typically employ imitation learning [24], such as the Moran rule [25], Fermi updating rule [16, 26], and follow-the-best rule [27] et al. The idea behind is that individuals are more likely to imitate strategies of neighbors who are better off, which can be viewed as a simple version of social learning [28].

Reinforcement learning (RL) [29] as an alternative paradigm provides a completely different perspective to study the evolution of cooperation. There, players score different actions within different states, and by repeatedly interacting with the environment they are able to make decisions by balancing the past experience, the present reward, and the expected earnings in the future. Despite its great potential [30–32], RL as a distinct learning paradigm from imitation learning has been largely overlooked. Only recently, researchers start to apply reinforcement learning to evolutionary game theory to help understand the emergence of cooperation [33–46], trust [47], resource allocation [48, 49], and other collective behaviors for humans [45, 50].

In particular, some new insights have been obtained for the emergence of cooperation within the RL framework. For example, Zhang et al. revealed that explosive cooperation appears in the form of periodic oscillation in snowdrift games with reinforcement learning [38]. By considering the uncertainties in daily life, Wang et al. uncovered that Lévy noise promotes the evolution of cooperation through reinforcement learning [39]. A year later, they proposed a framework by combining the public goods game with an adaptive reward mechanism within the reinforcement learning framework. The simulations demonstrate that the fraction of cooperation increases significantly when the adaptive reward strategy is included [40]. He et al. extended the PD game to mobile populations, identifying three different types of spontaneous segregation. Adaptive migration enhances network reciprocity and enables the dominance of cooperation in a dense population [41]. In the two-player scenario, Ding et al. revealed that a strong memory and long-sighted expectation yield the emergence of coordinated optimal policies, where both agents act like "win-stay, lose-shift" to maintain a high level of cooperation [42]. Current game-theoretic studies also indicate the reinforcement learning is able to catalyze cooperation when it is mixed with other updating rules [43, 51, 52]. In all of above work, however, they adopt the same state setup,

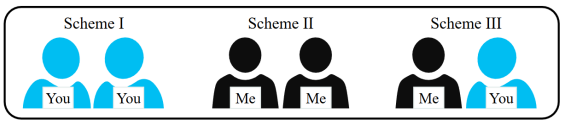

FIG. 1. **Three information schemes for playing a pairwise game.** Scheme I – "You + You" and Scheme II – "Me + Me" are both symmetric, but Scheme III – "You + Me" is asymmetric and both consider the action information of the blue player labeled with "Me".

i.e. individuals are assumed to perceive the same kind of information, either the agent's own action information [38–40] or the neighbors' [41], or both [42, 43]. These information perception is all symmetric in terms of information structure.

Yet, numerous real-world observations indicate that this symmetric perception is often not the case. In fact, factors such as age, experiences, culture, social status, wealth, personality, personal religious beliefs and sources of information directly shape an individual's unique perspective on issues [53, 54]. Likewise, indirect factors like economic, social, cultural and political environment [55], media influence, history and traditions, globalization, collectively shape societal views on problems, leading to a diverse range of perspectives. As the consequence, players may focus on the different part of the information available to them [56–59], and we have no idea about the consequence of these variations in the information intake. Therefore, it's natural to ask *how different information perceptions affect the evolution of cooperation.*

In this work, we focus on the impact of information on the evolution of cooperation within the reinforcement learning paradigm. Specifically, we employ the Q-learning algorithm [60, 61], wherein each individual possesses a Q-table as an action guide and plays the PD game [7]. By studying three simplest setups of information perception, we uncover different mechanisms of cooperation evolution in the three two-person scenarios. In two of them, the cooperation shows a first-order-like phase transition with the dilemma strength. Interestingly, in the asymmetric information scenario, the time evolution of the cooperation exhibits rich dynamical behaviors, undergoing cooperation emergence – breakdown – reestablishment. As a result, the highest cooperation preference is reached within the shortest time on average in this scenario.

This paper is organized as follows: we introduce our Q-learning model with three different information scheme in Sec. 2. In Sec. 3, we present the results. In Sec. 4, we provide a mechanistic analysis. In Sec. 5, the evolution processes for both symmetric and asymmetric information scenarios are compared. Finally, we conclude our work together with discussions in Sec. 6.

2. MODEL

We consider the two-player scenario where they play the prisoner's dilemma game (PD), each having two options: cooperation (C) or defection (D). Mutual cooperation brings each a reward R, while mutual defection leads to a punishment P for each. The mixed encounter scenario brings the cooperator the sucker's payoff S and the defector the temptation T. The payoffs need to satisfy $T>R>P>S$, and $T+S<2R$ for the collective concern. The payoff matrix is summarized as follows:

$$\mathbf{\Pi} = \begin{pmatrix} \Pi_{CC} & \Pi_{CD} \\ \Pi_{DC} & \Pi_{DD} \end{pmatrix} = \begin{pmatrix} R & S \\ T & P \end{pmatrix}, \tag{1}$$

where $R = 1.0$, $S = -b$, $T = 1 + b$, and $P = 0$ adopted in our study, corresponding to a strong version of PD [62]. $b > 0$ is the dilemma strength, a larger value of which means less likely for cooperation to survive.

In our work, players adopt the Q-learning algorithm [61], where their decision-making is guided by a two-dimensional table termed as Q-table. The Q-table in our study is as follows:

State \ Action	C (a_1)	D (a_2)
C (s_1)	Q_{s_1,a_1}	Q_{s_1,a_2}
D (s_2)	Q_{s_2,a_1}	Q_{s_2,a_2}

The state set $\mathbb{S} = \{\text{C,D}\}$ and the action set $\mathbb{A} = \{\text{C,D}\}$ are formally identical and simple. The items in the table are Q-value $Q_{s,a}$, which scores the value of the action $a \in \mathbb{A}$ within the given state $s \in \mathbb{S}$. A larger value of $Q_{s,a} > Q_{s,\hat{a}}$, the action a is more preferred rather than the other action $\hat{a}$ within the state s. While the action information available to players is definite, the set of states $\mathbb{S}$ reflects the information about the environment that individuals perceives. Different players could have different perceived information (i.e. the state set $\mathbb{S}$) which they may find useful.

Specifically, we consider three different information schemes. (I) Both players are informed of the opponent's action; (II) Both players consider one's own action information; (III) One player considers the opponent's action information, while the other considers one's own action information in the last round. Obviously, in either Scheme I or II the information used is structurally symmetric for the two players, but this is not the case in Scheme III, where they both concern the action of one player, and is thus asymmetric. The illustration of the three schemes is shown in Fig. 1.

The evolution of the two-player system follows a synchronous updating procedure. At the beginning, each player is randomly assigned an initial strategy C or D as the state, and the elements $Q_{s_l,a_m}(l,m=1,2)$ in the Q-tables are randomly assigned a value between $[0,1)$, indicating that individuals are initially unfamiliar with the environment. Given the state s at round t, (i) with a probability ϵ each player randomly chooses an action $a \in \mathbb{A}$ to conduct a trial-and-error exploration, otherwise each chooses an action a according to one's Q-table (i.e. a is selected given $Q_{s,a} > Q_{s,\hat{a}}$). (ii) Then, two players play the PD game and get a payoff π according to the matrix Eq. (1). (iii) They get their new state s' and update their Q-tables. Specifically, the element $Q_{s,a}(t)$ just referred

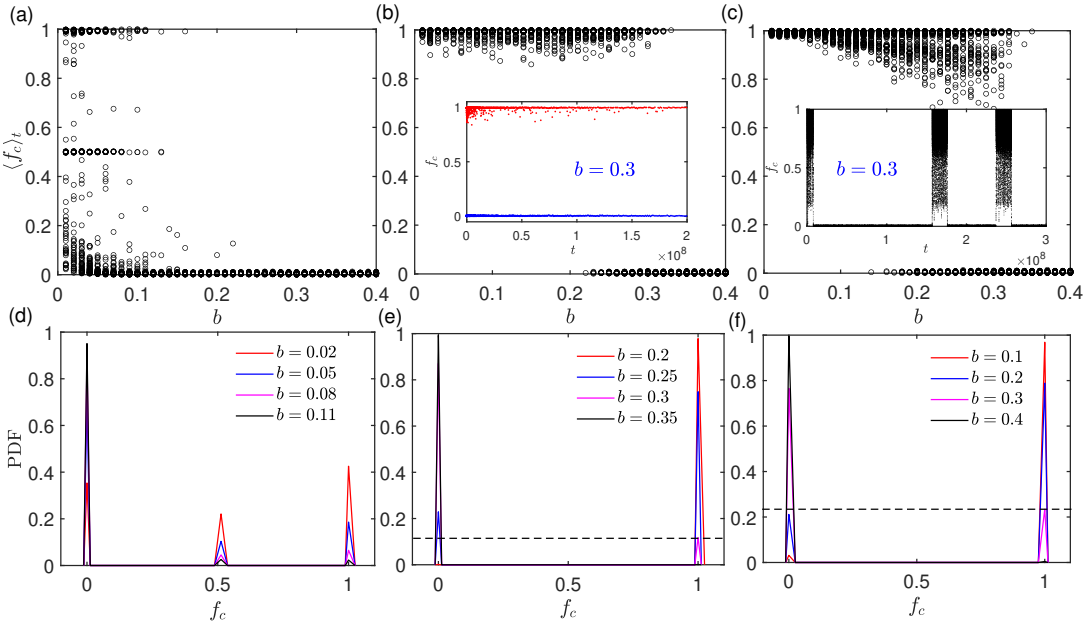

FIG. 2. **The dependence of cooperation preference on the dilemma strength within the three schemes.** (a-c) The time-averaged cooperation preference $\langle f_c \rangle_t$ versus the dilemma strength b, respectively for Scheme I, II, III. While no clear dependence is observed in Scheme I, the dependence shows a discontinuous transition of cooperation preference in Scheme II and III. The two insets shows typical time series of f_c for $b = 0.3$ in the corresponding scheme; the red and blue lines represent the results of evolution from two different initial conditions in (b). This means that once the system evolves into mutual cooperation or mutual defection, no change is expected. But persist state switches between the two solutions are always observed in (c). (d-f) The corresponding probability density function (PDF) curve of f_c, respectively, for Scheme I-III, where trimodal distribution is seen for Scheme I, and bimodal distributions are for the other two schemes. The dashed lines in (e, f) indicate the peak value of $f_c = 1$ where $b = 0.3$, where a higher value is observed in Scheme III than Scheme II. Each data is averaged 500 times after a transient of 3×10^8 rounds in (a)-(c). Other parameters: ϵ = 0.01, α = 0.1, γ = 0.9.

is updated as follows:

$$Q_{s,a}(t+1) = Q_{s,a}(t) + \alpha \left(\pi(t) + \gamma \max_{a'} Q_{s',a'}(t) - Q_{s,a}(t) \right) = (1-\alpha) Q_{s,a}(t) + \alpha \left(\pi(t) + \gamma \max_{a'} Q_{s',a'}(t) \right), \quad (2)$$

where $\alpha \in (0, 1]$ is the learning rate, capturing the contribution at the current step, a larger value of α means that the player is more forgetful as old Q-values tend to be rapidly modified. $\pi(t)$ is the payoff obtained at present round following the payoff matrix Eq. (1). $\gamma \in [0, 1)$ is the discount factor, measuring the weight of future rewards, as $\max_{a'} Q_{s',a'}(t)$ is the maximal value expected within the new state. The r.h.s. of the above equation indicates that the Q-values simultaneously contain the contribution of past experiences, reward at present and from the future.

The above process [steps (i)-(iii)] is repeated until the system reaches an equilibrium or the desired duration is completed. The three learning parameters are fixed at typical values of $\epsilon = 0.01, \alpha = 0.1, \gamma = 0.9$ throughout the study, where the players appreciate both past experiences and expected rewards in the future.

3. RESULTS

We report the evolution of cooperation for the three information schemes, where discontinuous transitions and bistability are uncovered, see Fig. 2. As shown in Fig. 2(a), when players focus on the opponent's action information (Scheme I), cooperation exhibits strong instability even at small values of temptation b. With the increase of b, the system evolves to a stable state dominated by mutual defection $f_c \approx 0$. Correspondingly, the probability density function (PDF) curves of f_c within the unstable interval in Fig. 2(d) shows a trimodal distribution. With increasing b, the peaks at 0.5 and 1 both reduce.

By contrast, when players focus on their own action information (Scheme II), Fig. 2(b) shows that the mutual cooperation ($f_c \approx 1$) is stable when $b \lesssim 0.22$. Further increasing b, however, leads to a dramatically different outcome — the system either evolves into mutual cooperation for some experiments, or the system evolves into mutual defection for some other realizations, depending on the initial conditions. Once mutual cooperation or defection is reached, the later evolution of f_c becomes quite stable, see the inset in Fig. 2(b). When $b > b_c \approx 0.32$, mutual defection is the only stable state. The observation of bistable state is strengthened by the bimodal PDF as shown in Fig. 2(e). As expected, the peak of the mu-

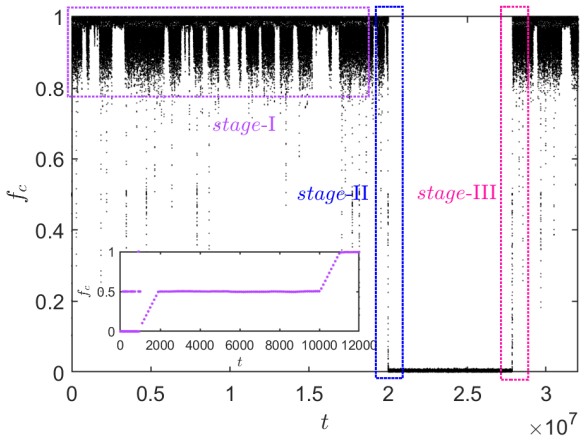

FIG. 3. **Typical time series of cooperation preference f_c in Scheme III**. A sliding window average of 500 steps is conducted. Based on the characteristics displayed in the time series, it can be divided into three stages: I) Emergence of cooperation, II) Breakdown of cooperation, and III) Rebuilding of cooperation. The inset shows the time series of f_c for the first 1.2×10^4 steps. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$, $b = 0.2$.

tual cooperation shrinks when b is increased, while the peak of mutual defection goes up. These features indicate that there is a first-order-like phase transition for the cooperation prevalence in Scheme II.

Finally, when the two players are of asymmetric information structure (Scheme III), a similar phase transition and a bimodal PDF are observed, see Fig. 2(c,f). Yet, there is an essential difference compared to Scheme II that the cooperation prevalence f_c shows a bounce between full cooperation and full defection, as shown in the inset of Fig. 2(c). In addition, detailed examination shows that when the value of b is larger, the possibility of cooperation emergence under Scheme III is higher than the value in Scheme II. For example, when $b = 0.3$, $f_c \approx 0.25$ in Scheme III while $f_c \approx 0.15$ in Scheme II.

These results suggest that the information structure has a huge impact on the evolution of cooperation, and asymmetric information leads to new complexities in the form of first-order phase transition and true bistability.

4. MECHANISM ANALYSIS

Here, we primarily analyze the mechanisms under the asymmetric scenario in Scheme III. The mechanism analyses for Schemes I and II are relatively straightforward and can be found in Appendices B and C, respectively.

To understand the mechanism in the case of information asymmetry, we now turn to the evolution of the Q-table. To be certain, we categorize the evolutionary process into three stages based on the characteristics exhibited by the typical time series of f_c shown in Fig. 3, with questions as follows:

1) Stage I: how does cooperation emerge?

2) Stage II: why does cooperation collapse?

3) Stage III: how does cooperation re-establishe afterwards?

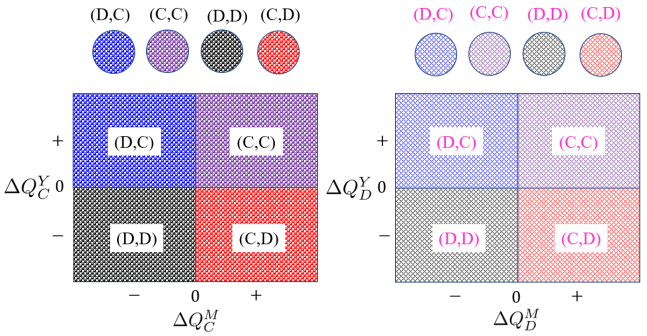

FIG. 4. **The action preference combinations of two players within two states.** The four quadrants, based on the sign of the value $\Delta Q^i_{s_l}$ ($i \in \{M, Y\}$), represent the four possible combinations of action preferences. The left and right figures correspond to the system being in state C and state D, respectively. For example, the combination (D, C) in pink indicates that in state D, individual *M* prefers action D, while individual *Y* prefers action C.

In addition to the elements $Q^i_{s_l,a_m}$ for each player i, we are particularly interested in their relative magnitude within a given row, i.e. $\Delta Q^i_{s_l} = Q^i_{s_l,a_1} - Q^i_{s_l,a_2}$. This value determines which action is preferred for player i within the given state s_l. For example, if $\Delta Q^i_{s_l}>0$, this means that for player i, the action C is preferred within the state s_l, otherwise D is supposed to be a better choice. Accordingly, we explicitly show the action preference combinations within two states (see Fig. 4) based on the sign of $\Delta Q^i_{s_l}$, where $i \in \{M, Y\}$, respectively, labels the one who considers one's own information of action ("Me") and the one who considers the opponent's action information ("You"). For example, action preference combination (D,C) represents that individual *M* chooses action D and individual *Y* chooses action C, which in different states are represented by different colors, state C (black), state D (pink).

To be certain, we start with a typical initial condition far from mutual cooperation $Q^M_{C,C} < Q^M_{C,D}$, $Q^M_{D,C} < Q^M_{D,D}$, $Q^Y_{C,C} > Q^Y_{C,D}$, $Q^Y_{D,C} < Q^Y_{D,D}$ and analyze the dynamical mechanisms.

Stage I — *Cooperation emergence*

To provide a clear and intuitive description of the evolutionary process at this stage, we divide the evolutionary mechanism of this stage into five distinct sub-stages.

Sub-stage I – *Two novices both prefer defection resulting in the combination of (D,D).*

At the beginning, both players are unfamiliar with the environment, thus they prioritize immediate payoffs and learn that D is more beneficial, leading to $\Delta Q_D^{M,Y} < 0$. Therefore, both

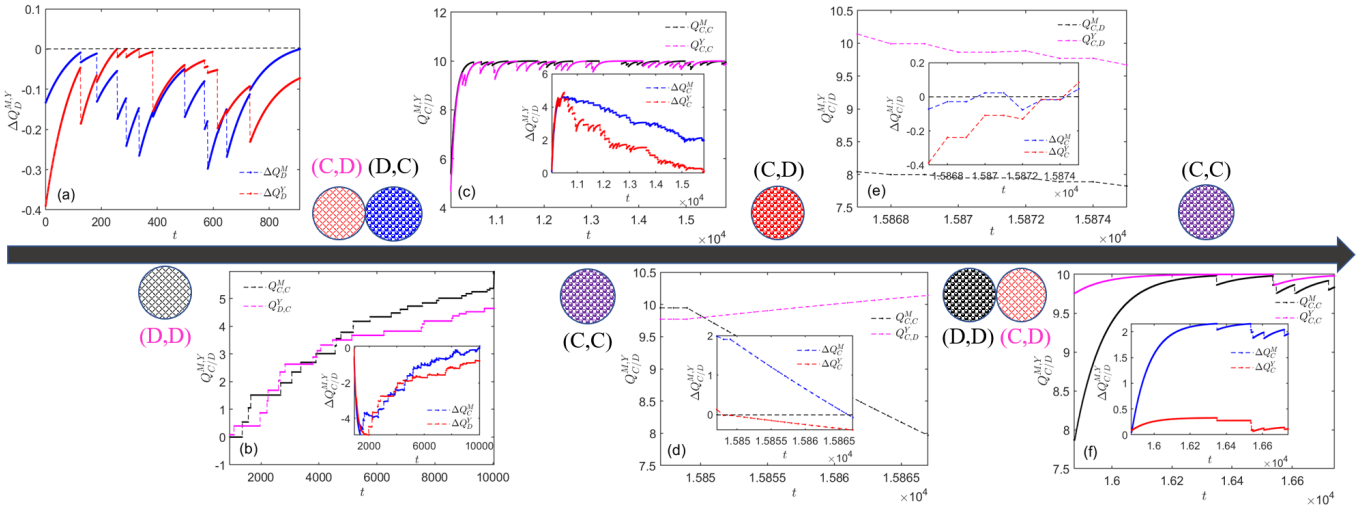

FIG. 5. **Cooperation emergence in Stage I.** It shows the evolution of action combination preferences, and the temporal evolution of $Q^{M,Y}_{s_l,a_m}$ or $\Delta Q^{M,Y}_{s_l}$ values. Here, the action preference combination (C,D)(D,C) indicates that individual *M* chooses defect in state C and cooperate in state D, causing the system to cycle between (C,D)↔(D,C). The same applies to (D,D)(C,D). The sudden declines in (a) are because of occasional cooperation by exploration, where the action of defection brings to a reward $\pi = 1 + b$. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

exhibit self-interested behavior in the form of mutual defection (D, D). However, the action preference combinations of (D, D) bring very low payoffs to both parties, which weakens the advantage of choice D and cause both values of $\Delta Q^{M,Y}_D$ back to zero [Fig. 5(a)]. The intermittent decreases are due to the action of C by exploration. When one of the values of $\Delta Q^{M,Y}_D \to 0$, meaning that their preferences in D are about to change.

Sub-stage II – *Player M's action preference shift leads to a new action preference combination (C, D) ↔ (D, C).*

When $\Delta Q^M_D > 0$ [Fig. 5(a)], the player *M*'s action preference is shifted from D to C. Correspondingly, the state of the system also undergoes the same change, and the system then enters a new action preference combination (C, D) ↔ (D, C) [Fig. 5(b)]. However, this action preference combination fails to persist in the presence of exploration.

Sub-stage III – *Exploratory behavior of both parties favors cooperation and mutual cooperation (C, C) is formed.*

Within the action preference combination (C, D) ↔ (D, C), the exploratory behavior of both parties is conducive to the growth of the utility function $Q_{s_l,C}$ [Fig. 5(b)], and the values of $Q^M_{C,C}$ and $Q^Y_{D,C}$ increase discontinuously. Correspondingly, ΔQ^M_C and ΔQ^Y_D show an increasing trend [see inset in Fig. 5(b)], indicating a gradual shift towards cooperation. Due to asymmetric information causing a faster increase in ΔQ^M_C, see Appendix A, individual *M* first transitions to cooperation in state D, leading the system to enter state C and establishing a stable positive feedback loop of mutual cooperation. As a result, the system enters a new action preference combination (C,C), the value of $Q^{M,Y}_{C,C}$ remains unchanged after continuous rise in Fig. 5(c).

Sub-stage IV – *Asymmetric information leads to exploitation of individual M by individual Y, the action preference combination (C,D) is formed.*

The action preference combination (C,C) remains unstable. The exploration behavior – defection of both players leads to an increase in $Q^{M,Y}_{C,D}$. However, due to asymmetric information, $Q^Y_{C,D}$ increases more rapidly, and ΔQ^Y_C is falling at a faster rate than ΔQ^M_C [see inset in Fig. 5(c)]. For more details, see Appendix A. Consequently, player *Y* transitions from cooperation to defection in state C first, leading the system enter the action preference combination (C, D). This combination can be viewed as a process of exploitation and tolerance. For individual *M*, positive feedback from prior mutual cooperation results in $\Delta Q^M_C > 0$, making *M* inclined to cooperate even when faced with defection, showing tolerance. Thus, individual *Y* can exploit *M* by choosing defection for a period.

Sub-stage V – *Player M implements a punishment-like policy on player Y, the corresponding action preference combination is (D,D) ↔ (C,D)*

However, tolerance within the action preference combination (C, D) is limited. Frequent exploitation by the opponent causes a continuous decline in $Q^M_{C,C}$ [Fig. 5(d)] and ΔQ^M_C to show a decreasing trend [see inset in Fig. 5(d)]. When $\Delta Q^M_C < 0$, individual *M* switches from action C to D in state C, transitioning the system to the combination preference of (D,D) ↔ (C,D). Within this combination, individual *Y*'s persistent exploitation from the previous sub-stage becomes intermittent, resulting in a reduced payoff and causing $Q^Y_{C,D}$ to start declining [Fig. 5(e)]. This can be seen as a punishment process by individual *M* towards individual *Y*. This process causes ΔQ^Y_C to rise [see inset in Fig. 5(e)]. When $\Delta Q^Y_C > 0$, individual *Y* reverts to cooperation, forming a positive feedback loop that returns the system to (C,C).

Stage I shows the evolution of cooperation emergence – exploitation and tolerance – punishment – mutual cooperation. However, the completion of this stage does not establish a stable cooperative relationship between the two play-

ers. As shown in the top panel of Fig. 6(a), the condition $\Delta Q_C^Y < 0$ occurs intermittently, indicating that individual *Y* still exploits individual *M* from time to time. As a result, the process of sub-stages III-V intermittently occurs in the subsequent evolution, causing ΔQ_C^M to fluctuate in the bottom panel of Fig. 6(a). Despite this, the system maintains a relatively high average cooperation preference (f_c>0.8) until it eventually transitions to a state of complete defection. This outcome is attributed to individual *M*'s inclination to cooperate in state D.

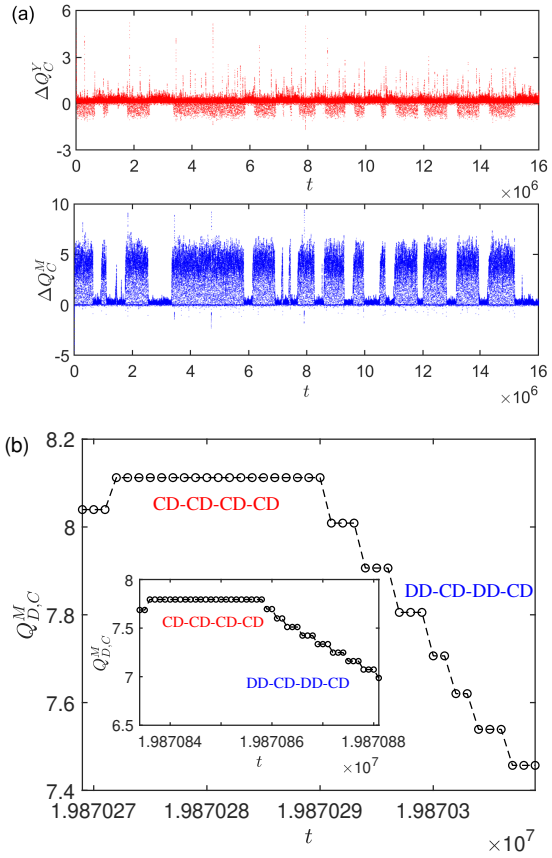

FIG. 6. **Cooperation collapse in Stage II**. (a) The time evolution of $\Delta Q_C^{Y,M}$. The upper panel shows the time evolution of ΔQ_C^Y, where intermittent occurrences of $\Delta Q_C^Y < 0$ can be observed, indicating the accumulation of exploitation of individual *M* by individual *Y*. The lower panel shows the evolution of ΔQ_C^M, with corresponding intermittent oscillations observed in the upper panel, each oscillation representing a punishment process of individual *Y* by individual *M*. (b) The time evolution of $Q_{D,C}^M$. It can be observed that the decrease in $Q_{D,C}^M$ mainly occurs during the punishment process of individual *Y* by individual *M*, with the corresponding action preference combination being (D,D)↔(C,D). The inset shows the evolution of $Q_{D,C}^M$ over different time periods. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

Stage II — *Cooperation collapse*

In Stage I, we observe that individual *M* frequently "forgives" individual *Y* and re-established mutual cooperation. However, an intriguing phenomenon emerges afterwards: individual *M* gradually loses patience and is no longer inclined to cooperate. As shown in Fig. 6(b) and the inset, the decline in $Q_{D,C}^M$ primarily occurs during sub-stage V. This indicates that each time individual *M* punishes individual *Y*, *M*'s inclination to choose cooperation in state D diminishes. Once tolerance is completely eroded, the system transitions into state D, resulting in a collapse of cooperation. Consequently, when the opponent exploits again, the system shifts to a state of mutual defection (D,D).

Stage III — *Cooperation reestablishment*

There the system transitions away from (D,D) to (C,C) again. Three distinct sub-stages can be divided in this stage.

Sub-stage I – *Simultaneous cooperative exploration breaks mutual defection, triggers (C,C) ↔ (D,D) cyclic state.*

Within the mutual defection state, the payoff π for either is zero, reducing their preference in defection. This is evidenced by the upward trend in $\Delta Q_D^{M,Y}$ shown in Fig. 7(a), the intermittent declines are due to occasional cooperative actions during exploration. Unilateral cooperation, however, only strengthens the other player's preference for defection because their preference in state C remains defection (i.e., $\Delta Q_C^{M,Y} < 0$). Simultaneous cooperation by both players can alter this situation. When both choose to cooperate, they each receive a payoff $\pi = R$, which triggers an increase in $Q_{D,C}^{M,Y}$ and leads to $\Delta Q_D^{M,Y} > 0$, indicating a reversal in preference as shown in Fig. 7(b). The system then enters a cyclical state of (C,C) ↔ (D,D). However, this action preference combination cannot be sustained under weak exploration.

Sub-stage II – *Alternating exploitation and punishment prepare for reestablishing cooperation.*

Within the action preference combination (C,C) ↔ (D,D), the exploration behavior – defection of both players leads to an increase in $Q_{D,D}^{M,Y}$. Due to asymmetric information, $Q_{D,D}^M$ increases more rapidly (for more details, see Appendix A.), causing ΔQ_D^M to decrease faster than ΔQ_D^Y [see inset in Fig. 7(b)]. Consequently, player *M* transitions from cooperation to defection in state D first, leading the system enter the action preference combination (D, C) – a process is similar to the exploitation and tolerance observed in sub-stage IV of stage I, with the roles reversed: *M* exploits *Y*, while *Y* tolerates *M*.

Then, player *Y* implements a similar punishment-like strategy on player *M*. Within the action preference combination (D, C), *M*'s continuous exploitation leads to a persistent decline in $Q_{D,C}^Y$. When $\Delta Q_D^Y \to 0$, *Y* gains no advantage in choosing either cooperation or defection, causing ΔQ_D^Y to fluctuate around zero [Fig. 7(c)]. The corresponding action preference combination is (D,D) ↔ (D,C), which then predominantly shifts to mutual defection (D,D) ↔ (D,D). This process results in an increase in ΔQ_D^M.

When $\Delta Q_D^M > 0$, individual *M* re-chooses cooperation in

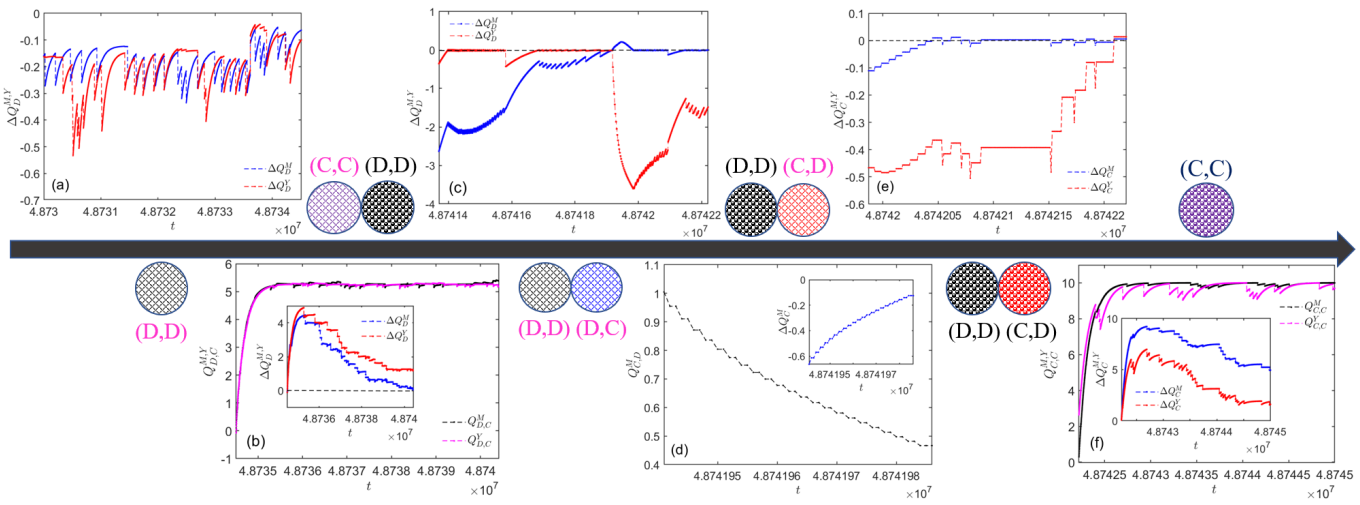

FIG. 7. **Cooperation reestablishment in Stage III**. It shows the evolution of action combination preferences, and the temporal evolution of $Q^{M,Y}_{s_l,a_m}$ or $\Delta Q^{M,Y}_{s_l}$. Here, the action preference combination (C,C)(D,D) indicates that individual *M* chooses defect in state C and cooperate in state D, causing the system to cycle between (C,C)↔(D,D). The same applies to (D,D)(C,D). The action pair (D,D)(D,C) indicates that in state D, individual *Y* alternates between cooperation and defection. Similarly, (D,D)(C,D) shows that individual *M* switches between cooperation and defection in state C. The sudden declines in (a) are because of occasional cooperation by exploration, where the action of defection brings to a reward $\pi = 1 + b$. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

state D. The roles of *M* and *Y* then reverse, repeating the previously described process. During this period, fluctuations of ΔQ^M_D around zero occasionally revert the system to state C, leading to a decrease in $Q^M_{C,D}$ [Fig. 7(d)]. When $\Delta Q^M_C > 0$, the player *M*'s action preference in state C shifts from defection to cooperation. However, due to the lack of positive returns, ΔQ^M_C fluctuates around zero [Fig. 7(e)]. This punishment-like strategy on player *Y* again results in an increase in ΔQ^Y_C. When $\Delta Q^Y_C > 0$, *Y*'s action preference in state C also shifts towards cooperation. This indicates that once the system returns to state C, the positive feedback from mutual cooperation can re-establish cooperation between both parties.

Sub-stage III – *Cooperation is reestablished when individual M chooses cooperation.*

Up to this point, the two individuals have reached a consensus to cooperate in state C. When individual *M* re-chooses cooperation, the system enters state C and mutual cooperation is successfully re-established. In Fig. 7(f), a result identical to that in Fig. 5(c) indicates that the system returns to the Stage I evolution process.

Finally, to gain an intuitive understanding of the cooperation evolution in Scheme III, we show evolutionary paths for four typical dilemma strengths b, see Fig. 8. These shown are obtained by the following procedures. Starting with all possible combinations of the two Q-tables (i.e. 4×4 cases), we monitor the evolution of these combinations, where some "attractors" are observed. For a small value of dilemma strength ($b = 0.1$), mutual cooperation is the only stable solution, while for a large value ($b = 0.4$) mutual defection is exclusively stable. For the cases in between ($b = 0.2, 0.3$), the two attractors compete with each other, the evolution of the system is up to which basin of attraction are their initial conditions located. The observations align with the overall picture discussed above.

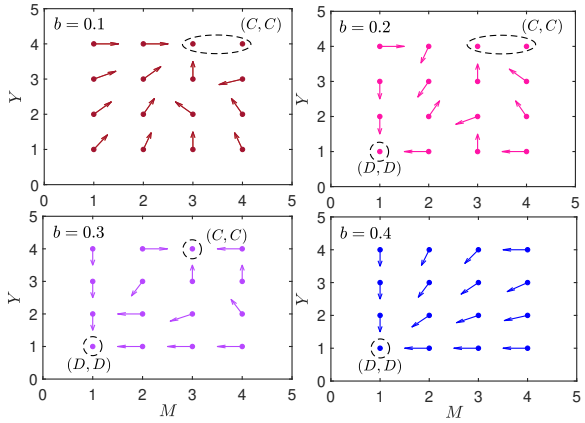

FIG. 8. **Evolutionary paths in Scheme III.** Starting with all possible settings of the initial Q-table for the two players (labeled "*Y*" and "*M*") for four typical dilemma intensities b. The axis labels $1-4$ respectively represent combinations of ($\Delta Q_C < 0$, $\Delta Q_D < 0$), ($\Delta Q_C < 0$, $\Delta Q_D > 0$), ($\Delta Q_C > 0$, $\Delta Q_D < 0$), ($\Delta Q_C > 0$, $\Delta Q_D > 0$). The arrows show the evolutionary directions of the combination type during a fixed time interval $t = 3 \times 10^5$. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

5. FURTHER COMPARISON

In this section, we first show the phase diagrams for three different schemes in Fig. 9, which reports the average cooperation preference f_c in the two key learning parameters domain

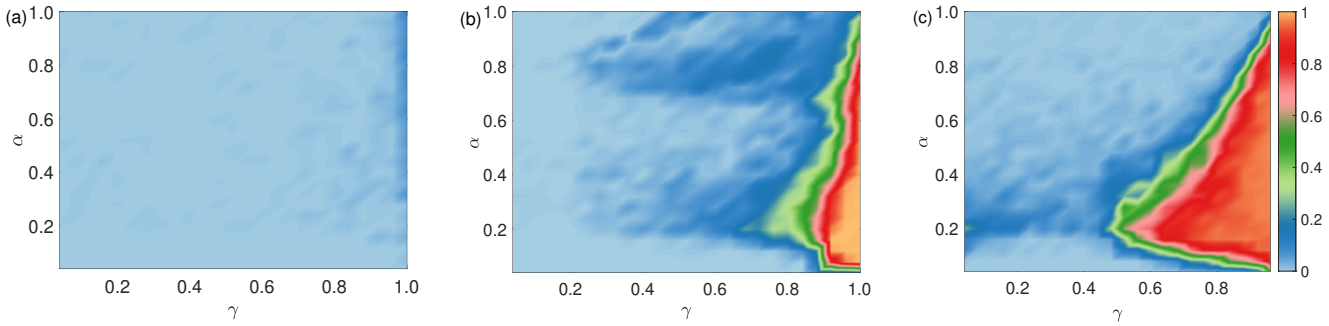

FIG. 9. (Color online) **Phase diagram.** The color-coded averaged cooperation preference f_c in the domain (γ, α), (a-c) are respectively for Scheme I - III. The red regions indicate that cooperation dominates, which often emerge for the combination of a small the learning rate α and a large the discount factor γ. Each data is averaged 100 realizations, and for each realization the data is averaged 500 rounds after a transient of 2×10^8. Other parameters : $\epsilon = 0.01$, $b = 0.2$.

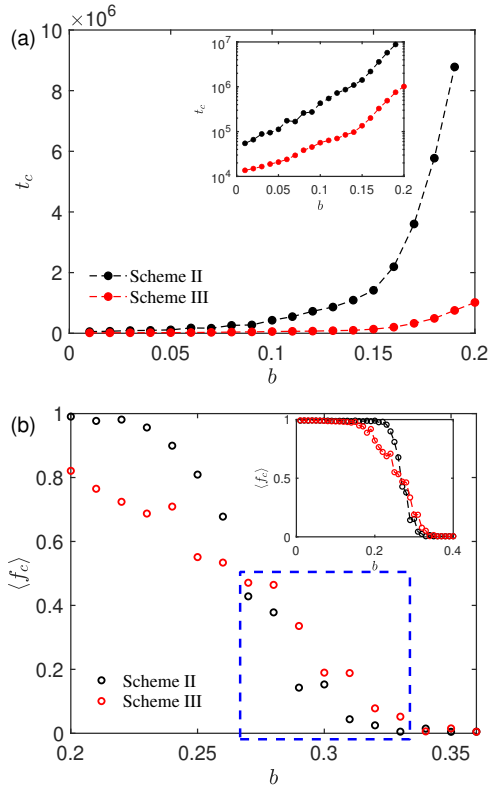

FIG. 10. **Comparison between Scheme II and III.** (a) The convergence time t_c versus the dilemma strength b, and the inset shows the same data but y-axis is taken logarithmic. (b) The averaged cooperation preference $\langle f_c \rangle$ versus b, 100 ensemble averages are conducted for each data besides the time average as we did in Fig. 2(a-c). Other parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

(γ, α) by fixing the dilemma strength $b = 0.2$. We find that in Scheme I, there is no emergence of cooperation across the region for the given b [Fig. 9(a)], instead decent levels of cooperation are observed in the other two schemes. As shown in Fig. 9(b, c), the red regions indicate that cooperation dominates ($f_c \sim 0.8$), where the learning rate α is mostly small and the discount factor γ is large. The observation can be interpreted as that a high level of cooperation emerges only when players both pay attention to their historical experiences and have a long-term vision. Besides, the region dominated by cooperation within the Scheme III is wider than that of Scheme II. Detailed examination shows that in the case of asymmetric information, a moderate degree of future expectation γ is sufficient to trigger the emergence of cooperation given a small value of learning rate α.

Apart from the average cooperation preference $\langle f_c \rangle$ at the final state, the coverage time towards the final state also matters. Fig. 10(a) shows that average converge time t_c for the system towards full cooperation are much shorter in Scheme III than the values within Scheme II. Across the whole range of b, the converge time in Scheme II is about one order larger compare to the case of Scheme III. In Fig. 10(b), we can observe that there is a crossover in the average cooperation preference as b is varied. A higher $\langle f_c \rangle$ in Scheme II is observed when $b < 0.26$, while the opposite observation is made when $b > 0.26$. The reasons behind the difference shown in Fig. 10 is closed related to the evolutionary mechanism of Scheme II, which is analyzed in the Appendix B.

6. DISCUSSION

In summary, we investigate the evolution of the iterated prisoner's dilemma game under three distinct information scenarios within the framework of reinforcement learning. Diverging from existing research on cooperation within this paradigm, our primary focus is on the impact of different information on the evolution of cooperation. Specifically, we examined three simplest information settings. Our findings reveal that the information structure significantly influences the evolution of cooperation. In the two symmetric scenarios, the direct association between the actions and states is more conducive to the emergence of cooperation. Remarkably, cooperation is more likely to emerge in the asymmetric scenario and does so more rapidly. The evolutionary dynamics exhibit characteristics of a first-order-like phase transition,

with the average cooperation preference transitions between mutual cooperation and mutual defection. Further mechanism analysis elucidates how cooperation emerges, breaks down, and reconstructs. The basin of attraction for the two stable states are identified for some typical dilemma intensities.

The majority of existing research on cooperation primarily focuses on how cooperation emerges and is sustained [24, 63]. There are very few work discussing how cooperation breaks down and the conditions necessary for cooperative reconstruction [42]. Our study provides insights into the establishment of long-term cooperation, suggesting that a moderate and tolerant understanding may prolong cooperative relationships. However, individual tolerance towards others is often limited, and relentless exploitation can lead to the collapse of cooperative ties. These observations are in line with our daily experience. Additionally, the reconstruction of cooperation is not straightforward, during this process, the player previously engaged in exploitation often finds itself in a disadvantaged position.

Certainly, our work serves as an initial effort to understand the impact of information perception on cooperation within the RL framework. We only consider three simplest information structures within two-body scenarios. In the real world, however, due to the diversity of personal and society factors [55], people may grab very different parts of information. Furthermore, people are connected within complicated social relationships which are often depicted as complex networks [64], and their information perception may also influence with each other through the networks. These complexities are far more complex beyond what we studied here and require further systematic studies in the future.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China [Grants Nos. 12075144,12165014] and Fundamental Research Funds for the Central Universities (GK202401002).

Appendix A: Asymmetric information causes imbalanced Q-value evolution between two players

Within sub-stage III of stage-I

There the system in a (D,C)↔(C,D) cyclic state. Individual *M* cooperates in state D and defects in state C, while individual *Y* cooperates in state C and defects in state D.

For individual *M*: exploratory cooperation in state C shifts the action preference from (D,C)↔(C,D) to (C,C)↔(D,C), resulting in an immediate mutual cooperation payoff of $R = 1$ and an increase in $Q^M_{C,C}$. This exploratory behavior also drives the system to state C, yielding a temptation value of $T = 1.2$ under the (D,C) preference, thus accelerating the increase in $Q^M_{C,C}$.

For individual *Y*: exploratory cooperation in state D shifts the action preference from (D,C)↔(C,D) to (C,C)↔(D,C), resulting in an immediate mutual cooperation payoff of $R = 1$ and an increase in $Q^Y_{D,C}$. However, since individual *Y* cannot directly alter the system's state, it continues along its previous trajectory into state C, where under the (D,C) action preference, it receives the payoff for sucker, $S = -0.2$, without the additional incentive seen in individual *M*.

Within sub-stage IV of stage-I

There the system in a (C,C) state, both individuals choose to cooperate in state C.

For individual *M*: exploratory defection in state C drives the system to state C, shifting the action preference combination from (C,C) to (D,C)↔(C,D), then back to (C,C) [consistent with the process in sub-stage III]. Then, an immediate temptation payoff of $T = 1.2$ is obtained and $Q^M_{C,D}$ is increased.

For individual *Y*: exploratory defection in state C shifts the action preference combination from (C,C) to (C,D), then back to (C,C). This results in an immediate temptation payoff of $T = 1.2$, leading to an increase in $Q^Y_{C,D}$. However, unlike individual *M*, *Y* cannot alter the system's state, thus bypassing the process of reverting to sub-stage III, consequently accelerating the increase in $Q^Y_{C,D}$.

Within sub-stage II of stage-III

There the system in a (C,C)↔(D,D) cyclic state. Both individuals choose to cooperate in state D and defect in state C.

For individual *M*: exploratory defection in state D shifts the action preference from (C,C)↔(D,D) to (D,C)↔(C,C), resulting in an immediate temptation payoff of $T = 1.2$ and an increase in $Q^M_{D,D}$. This exploratory behavior also drives the system to state D, yielding a mutual cooperation payoff of $R = 1$ under the (C,C) preference, thus accelerating the increase in $Q^M_{D,D}$.

For individual *Y*: exploratory defection in state D shifts the action preference from (C,C)↔(D,D) to (C,D)↔(D,D), resulting in an immediate temptation payoff of $T = 1.2$ and an increase in $Q^Y_{D,D}$. However, since individual *Y* cannot alter the system's state, it continues along its previous trajectory into state C, where under the (D,D) action preference, it receives the payoff for punishment, $P = 0$, without the additional incentive seen in individual *M*.

Appendix B: mechanism analysis in Scheme I

In Scheme I, the states of both parties directly depend on their opponent's action information. However, there is a key issue in maintaining mutual cooperation: both parties must always choose to cooperate in state C. Once one party shifts

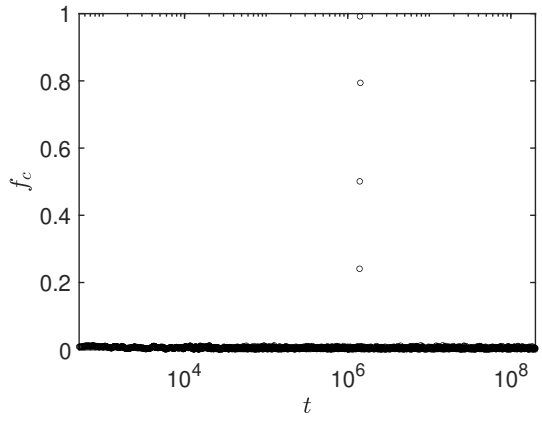

FIG. 11. Typical time series of cooperation preference f_c in Scheme I. A sliding window average of 500 steps is conducted. As can be seen from the figure, cooperation fails to emerge and be sustained. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$, $b = 0.2$.

from cooperation to betrayal, the system will enter a stage of mutual exploitation. Because actions under this scheme cannot directly change the state, it is impossible to implement punishment-like policies similar to those in Scheme II and Scheme III. Therefore, when exploratory betrayal behaviors accumulate advantages over time, the cooperation conditions will no longer be met, ultimately leading the system into a state of mutual betrayal. As can be seen from the figure 11, cooperation fails to emerge and be sustained.

For more details, we denote two individuals as $i = Y_1, Y_2$, who consider their opponent's action information. The values of Q_{s_l,a_m} and $\Delta Q^i_{s_l}$ are labeled as in the text. Even starting from an initial condition of full cooperation, i.e., $Q^{Y_1,Y_2}_{C,C} > Q^{Y_1,Y_2}_{C,D}$ and $Q^{Y_1,Y_2}_{D,C} > Q^{Y_1,Y_2}_{D,D}$, occasional exploratory choices of betrayal by both parties will lead to an increase in $Q^{Y_1,Y_2}_{C,D}$.

After the advantage of betrayal accumulates over time, satisfying $Q^{Y_1/Y_2}_{C,D} > Q^{Y_1/Y_2}_{C,C}$, one individual switches from cooperation to betrayal in state C. This initiates a continuous exploitation process (C,D)−(C,D)−(C,D). As continuous exploitation causes the other individual¡¯s tendency to cooperate in state D to decline, they eventually switch to betrayal in state D, leading to another continuous exploitation process with roles reversed (D,C)−(D,C)−(D,C). This process ultimately results in both individuals having no inclination to choose cooperation in either state, leading to the tragedy of total betrayal.

Appendix C: mechanism analysis in Scheme II

In Scheme II, the states of both parties directly depend on their respective action information. Therefore, cooperation can be rapidly established only if the random initial conditions fall within the mutual cooperative basin of attraction. If the initial conditions are closer to mutual betrayal, the prerequisite for triggering cooperation is that both parties simultaneously engage in exploratory cooperative behavior. This contrasts with Scheme III, where information can be transmitted through shared states, thereby expediting coordination. This also explains why the convergence time (t_c) for high cooperation preference in Scheme III, as depicted in Fig. 10(a), is significantly shortened.

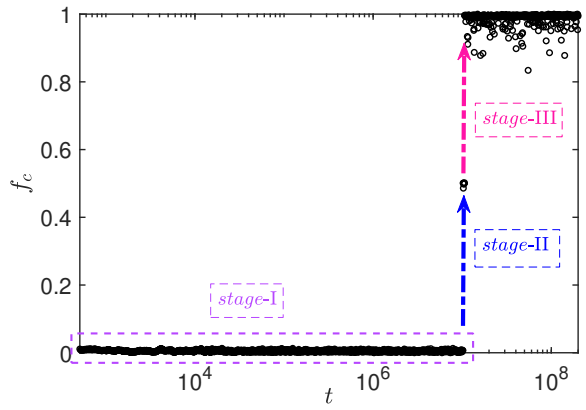

FIG. 12. Typical time series of cooperation preference f_c in Scheme II. A sliding window average of 500 steps is conducted. Based on the characteristics displayed in the time series, it can be divided into three stages: I) Mutual betrayal, II) Breaking away from mutual betrayal, and III) Establishing of cooperation. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$, $b = 0.2$.

To understand the mechanism in Scheme II, we categorize the evolutionary process into three stages based on the characteristics exhibited by the typical time series of f_c shown in Fig. 12.

1) Stage-I: Mutual betrayal.

2) Stage-II: Breaking away from mutual betrayal.

3) Stage-III: Establishing and maintaining mutual cooperation.

Here, $i = \{M_1, M_2\}$ respectively labels the two players, who consider their own action information, the values of Q_{s_l,a_m} and $\Delta Q^i_{s_l}$ are labeled in the same way as they are in the text. We initiate the study from initial conditions far from cooperation, i.e., $Q^{M_1,M_2}_{CC} < Q^{M_1,M_2}_{CD}$ and $Q^{M_1,M_2}_{DC} < Q^{M_1,M_2}_{DD}$, and analyze the mechanism in stages:

Stage I — *Mutual betrayal*

During this stage, mutual defection (D, D) does not yield any payoffs for either party, leading to an increase in $\Delta Q^{M_1,M_2}_D$. Intermittent decreases occur due to exploratory cooperation. When $\Delta Q^{M_1/M_2}_D > 0$, his/her preference shifts from defection (D) to cooperation (C). However, unilateral cooperation merely strengthens the other player's preference for

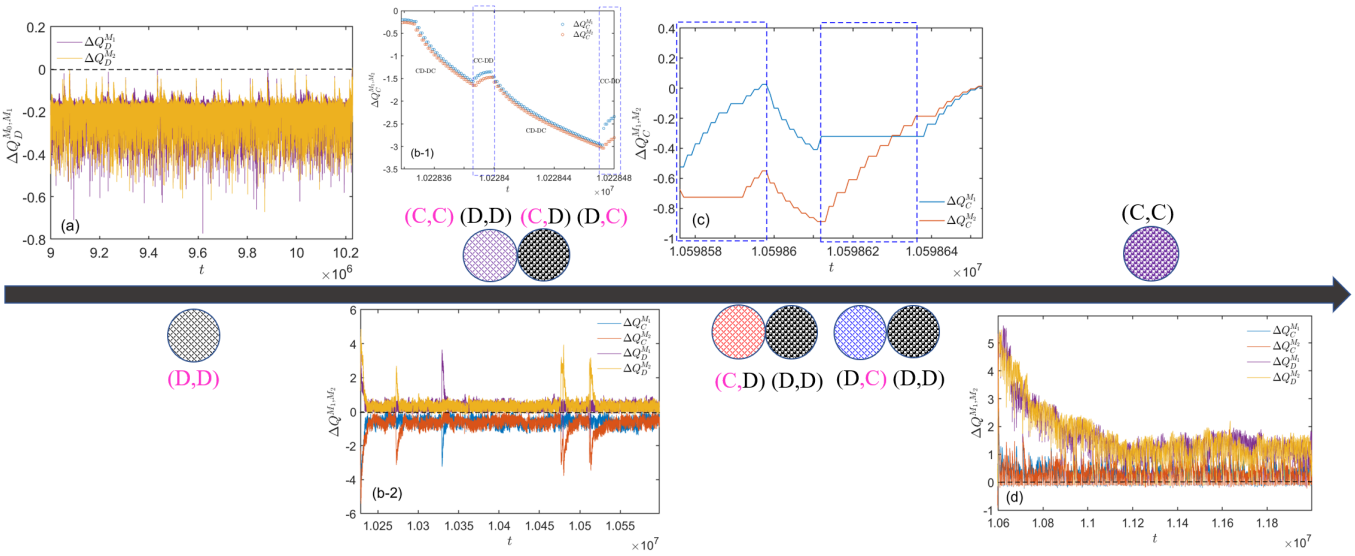

FIG. 13. Dynamical evolution process in Scheme II. The figures show the evolution of action combination preferences, and the temporal evolution of $\Delta Q_{s_l}^{M_1,M_2}$. Here, the action combination (C,C)(D,D) indicates that both individuals choose to defect in state C and to cooperate in state D. An exploratory action by one party can disrupt this synchronization, leading to (C,D)(D,C), while still maintaining the same action preferences. Thus, (C,C)(D,D) (C,D)(D,C) represent the alternation between synchronized and unsynchronized states under the influence of exploratory actions. The action combinations (C,D)(D,D) and (D,C)(D,D) indicate that one individual chooses to defect in state D, while the other defects in state C and cooperates in state D. Therefore, (C,D)(D,D) (D,C)(D,D) represent this process occurring sequentially and swapping the positions of the two individuals. This can be viewed as an alternating exploitation and punishment process. Parameters: $\epsilon = 0.01$, $\alpha = 0.1$, $\gamma = 0.9$.

defection, as the only perceivable change is an increased payoff for maintaining the original action. Therefore, breaking the (D, D) preference through a unilateral shift is challenging. As shown in Fig. 13(a), $\Delta Q_D^{M_0,M_1}$ fluctuates but remains consistently below 0.

Stage II — *Breaking away from mutual betrayal*

Sub-stage I – *Simultaneous cooperative exploration breaks mutual defection, triggers (C,C) ↔ (D,D) cyclic state.*

When both individuals simultaneously engage in exploratory cooperative behavior, they achieve positive payoffs R, which leads to a continuous increase in $Q_{D,C}^{M_1,M_2}$. This results in $\Delta Q_D^{M_1,M_2} > 0$ and a reversal in preference [Fig. 13(b-2)]. The system then cycles between (C,C) ↔ (D,D). Subsequent exploratory behavior disrupts this synchronization, forming the combinations (C,D) ↔ (D,C). Consequently, synchronization and asynchronization alternate [Fig. 13(b-1)], with both parties choose to cooperate in state D and defect in state C. However, this action preference combination fails to persist with weak exploration.

Sub-stage II – *Alternating exploitation and punishment prepare for establishing cooperation.*

Within the above action preference combination, both individuals' exploratory defection in state D leads to intermittent increases in $Q_{D,D}^{M_1,M_2}$. When $Q_{D,D}^{M_1/M_2} > Q_{D,C}^{M_1/M_2}$, the system forms the action preference combination (C,D) ↔ (D,D), with (D,D) occurring more frequently. This can be viewed as a process where one party punishes the other, resulting in an increasing trend in $\Delta Q_C^{M_1/M_2}$[Fig. 13(c)]. When $\Delta Q_C^{M_1/M_2} > 0$, the action preference in state C shifts towards cooperation. As indicated by the rectangular dotted boxes, when this process occurs sequentially for both individuals, their preference for defection in state C transitions to cooperation, establishing a (C,C) positive feedback loop.

Stage III — *Establishing and maintaining mutual cooperation*

In contrast to Scheme III, where individual Y continuously exploits individual M through asymmetric information, leading to the collapse of cooperation, the case of symmetric information presents a different dynamic. Here, the choice of betrayal by either party directly transitions their respective states to state D. As depicted in Fig. 13(d), guided by the Q-table, both parties tend to opt for cooperation in state D, returning to state C and simultaneously enhancing $Q_{D,C}^{M}$. Consequently, the stability of the cooperative relationship emerges from both parties' propensity to choose cooperation in state D.

* Email address: chenl@snnu.edu.cn

[1] R. Axelrod and W. D. Hamilton, Science **211**, 1390 (1981).
[2] J. Maynard Smith and E. Szathmáry, *The Major Transitions in Evolution* (Oxford University Press, 1995).

[3] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc, Physics Reports **948**, 1 (2022).
[4] E. Pennisi, Science **309**, 93 (2005).
[5] M. A. Nowak and K. Sigmund, Science **303**, 793 (2004).
[6] C. P. Roca, J. A. Cuesta, and A. Sánchez, Physics of Life Reviews **6**, 208 (2009).
[7] M. Doebeli and C. Hauert, Ecology Letters **8**, 748 (2005).
[8] A. Rapoport and A. M. Chammah, *Prisoner's dilemma: A study in conflict and cooperation*, Vol. 165 (University of Michigan press, 1965).
[9] M. A. Nowak, Science **314**, 1560 (2006).
[10] M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Physics Reports **687**, 1 (2017).
[11] R. L. Trivers, The Quarterly Review of Biology **46**, 35 (1971).
[12] M. A. Nowak and K. Sigmund, Nature **393**, 573 (1998).
[13] D. C. Queller, Nature **201**, 1145 (1964).
[14] K. Sigmund, C. Hauert, and M. A. Nowak, Proc. Natl. Acad. Sci. U.S.A. **98**, 10757 (2001).
[15] M. A. Nowak and R. M. May, Nature **359**, 826 (1992).
[16] G. Szabó and C. Tőke, Phys. Rev. E **58**, 69 (1998).
[17] Z. Wang, A. Szolnoki, and M. Perc, Scientific Reports **3**, 1183 (2013).
[18] R. Liang, Q. Wang, J. Zhang, G. Zheng, L. Ma, and L. Chen, Physical Review E **105**, 054302 (2022).
[19] M. Perc and A. Szolnoki, Physical Review E **77**, 011904 (2008).
[20] F. C. Santos, M. D. Santos, and J. M. Pacheco, Nature **454**, 213 (2008).
[21] R. Liang, J. Zhang, G. Zheng, and L. Chen, Physica A: Statistical Mechanics and its Applications **567**, 125726 (2021).
[22] C. Xia, J. Wang, M. Perc, and Z. Wang, Physics of Life Reviews **46**, 8 (2023).
[23] L. Ma, J. Zhang, G. Zheng, R. Liang, and L. Chen, Chaos, Solitons & Fractals **171**, 113452 (2023).
[24] C. P. Roca, J. Cuesta, and A. Sánchez, Physics of Life Reviews **6**, 208 (2009).
[25] V. Knight, M. Harper, N. E. Glynatsi, and O. Campbell, PLOS ONE **13**, 1 (2018).
[26] G. Szabó, J. Vukov, and A. Szolnoki, Phys. Rev. E **72**, 047107 (2005).
[27] M. Nowak and K. Sigmund, Nature **364**, 56 (1993).
[28] A. Bandura and R. H. Walters, *Social Learning Theory*, Vol. 1 (Englewood cliffs Prentice Hall, 1977).
[29] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT press, 2018).
[30] D. Lee, Nature Neuroscience **11**, 404 (2008).
[31] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Nature **529**, 484 (2016).
[32] A. Subramanian, S. Chitlangia, and V. Baths, Neural Networks **145**, 271 (2022).
[33] S. Tanabe and N. Masuda, Journal of Theoretical Biology **293**, 151 (2012).
[34] T. Ezaki, Y. Horita, M. Takezawa, and N. Masuda, PLoS Computational Biology **12**, e1005034 (2016).
[35] Y. Horita, M. Takezawa, K. Inukai, T. Kita, and N. Masuda, Scientific Reports **7**, 39275 (2017).
[36] H. Ding, G. Zhang, S. Wang, J. Li, and Z. Wang, Physica A: statistical mechanics and its applications **536**, 122551 (2019).
[37] L. Fan, Z. Song, L. Wang, Y. Liu, and Z. Wang, Chaos: An Interdisciplinary Journal of Nonlinear Science **32**, 123140 (2022).
[38] S. Zhang, J. Zhang, L. Chen, and X. Liu, Nonlinear Dynamics **99**, 3301 (2020).
[39] L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, and Z. Wang, Nonlinear Dynamics **108**, 1837 (2022).
[40] L. Wang, L. Fan, L. Zhang, R. Zou, and Z. Wang, New Journal of Physics **25**, 073008 (2023).
[41] Z. He, Y. Geng, C. Du, L. Shi, and Z. Wang, New Journal of Physics **24**, 123038 (2022).
[42] Z. Ding, G. Zheng, C. Cai, W. Cai, L. Chen, J. Zhang, and X. Wang, Chaos, Solitons & Fractals **175**, 114032 (2023).
[43] Y. Geng, Y. Liu, Y. Lu, C. Shen, and L. Shi, Applied Mathematics and Computation **427**, 127182 (2022).
[44] Z. Yang, L. Zheng, M. Perc, and Y. Li, Applied Mathematics and Computation **463**, 128364 (2024).
[45] Y. Shi and Z. Rong, IEEE Transactions on Circuits and Systems II: Express Briefs **69**, 2463 (2022).
[46] J. Zhang, Z. Rong, G. Zheng, J. Zhang, and L. Chen, Journal of Physics: Complexity **5**, 025006 (2024).
[47] G. Zheng, J. Zhang, J. Zhang, W. Cai, and L. Chen, New Journal of Physics **26**, 053041 (2024).
[48] M. Andrecut and M. Ali, Physical Review E **64**, 067103 (2001).
[49] S. Zhang, J. Dong, L. Liu, Z. Huang, L. Huang, and Y. Lai, Physical Review E **99**, 032302 (2019).
[50] M. S. Tomov, E. Schulz, and S. J. Gershman, Nature Human Behaviour **5**, 764 (2021).
[51] X. Han, X. Zhao, and H. Xia, Chaos, Solitons & Fractals **164**, 112684 (2022).
[52] A. Sheng, J. Zhang, G. Zheng, J. Zhang, W. Cai, and L. Chen, arXiv preprint arXiv:2406.11121 (2024), https://doi.org/10.48550/arXiv.2406.11121.
[53] R. Dawkins, *The Selfish Gene* (Oxford University Press, 1989).
[54] J. Molinas, World Development **26**, 413 (1998).
[55] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (Oxford University Press, 2004).
[56] J. H. Kagel, C. Kim, and D. Moser, Games and Economic Behavior **13**, 100 (1996).
[57] P. D. Allison, American Sociological Review **43**, 865 (1978).
[58] N. Feltovich, Econometrica **68**, 605 (2000).
[59] A. McAvoy and C. Hauert, PLOS Computational Biology **11**, 1 (2015).
[60] C. J. C. H. Watkins, *Learning from delayed rewards (Ph.D. thesis)*, Ph.D. thesis (1989).
[61] P. Watkins, Christopher J. C. H.and Dayan, Machine Learning **8**, 279 (1992).
[62] R. Axelrod and W. D. Hamilton, Science **211**, 1390 (1981).
[63] K. Sigmund, *The Calculus of Selfishness* (Princeton University Press, 2010).
[64] M. Newman, *Networks* (Oxford university press, 2018).