FRONTIERS OF PHYSICS

RESEARCH ARTICLE
Volume 19 / Issue 4 / 40201 / 2024

Self organizing optimization and phase transition in reinforcement learning minority game system

Si-Ping Zhang[1], Jia-Qi Dong[2], Hui-Yu Zhang[1], Yi-Xuan Lü[1], Jue Wang[1], Zi-Gang Huang[1,†]

1 Key Laboratory of Biomedical Information Engineering of Ministry of Education, Key Laboratory of Neuro-informatics & Rehabilitation Engineering of Ministry of Civil Affairs, and Institute of Health and Rehabilitation Science, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

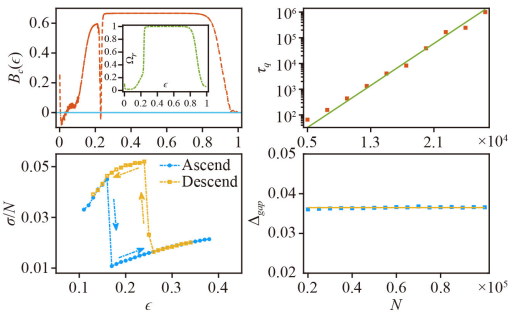
2 Lanzhou Center for Theoretical Physics and Key Laboratory of Theoretical Physics of Gansu Province, Lanzhou University, Lanzhou 730000, China

Corresponding author. E-mail: †huangzg@xjtu.edu.cn

Received September 19, 2023; accepted November 1, 2023

ABSTRACT

Whether the complex game system composed of a large number of artificial intelligence (AI) agents empowered with reinforcement learning can produce extremely favorable collective behaviors just through the way of agent self-exploration is a matter of practical importance. In this paper, we address this question by combining the typical theoretical model of resource allocation system, the minority game model, with reinforcement learning. Each individual participating in the game is set to have a certain degree of intelligence based on reinforcement learning algorithm. In particular, we demonstrate that as AI agents gradually becomes familiar with the unknown environment and tries to provide optimal actions to maximize payoff, the whole system continues to approach the optimal state under certain parameter combinations, herding is effectively suppressed by an oscillating collective behavior which is a self-organizing pattern without any external interference. An interesting phenomenon is that a first-order phase transition is revealed based on some numerical results in our multi-agents system with reinforcement learning. In order to further understand the dynamic behavior of agent learning, we define and analyze the conversion path of belief mode, and find that the self-organizing condensation of belief modes appeared for the given trial and error rates in the AI system. Finally, we provide a detection method for period-two oscillation collective pattern emergence based on the Kullback–Leibler divergence and give the parameter position where the period-two appears.



Keywords oscillatory evolution, collective behaviors, phase transition, reinforcement learning, minority game

1 Introduction

The emergence of collective behaviors [1–10] is an intriguing enigma in self-organizing systems, which is ubiquitous in living systems (e.g., starling flock, ant colony, fish swarm and microbial swarms), human social systems (e.g., the formation of social norms, political behavior and collective decision-making), and computer simulation systems (e.g., Game of Life). As a typical self-organizing systems, resource allocation systems that a

Higher Education Press

https://doi.org/10.1007/s11467-023-1378-z

basis of the modern economy and society emerge diverse collective behaviors, in which the mechanism of minority wins [11] often dominates individual behavior, and patterns of collective behavior of individuals are critical to the efficiency and functional realization of systems.

To uncover the dynamical mechanism of collective behavior evolution in resources allocation system, a large number of theoretical models involving minority games have been proposed [12] and developed [13–23] in the last 20 years, aiming to reveal various collective behavior patterns and their corresponding dynamic mechanisms. The idea of minority game has been also widely used in the theoretical modeling and analysis of various real resource allocation systems, such as resource allocation in cloud manufacturing [24], collective decision making in fish schools [25], and arbitrageurs decision-making behavior in financial markets [26]. Another focused core problem is to control the dynamical evolution of the system by means of an external intervention to suppress herd behavior, e.g., pinning control [15], because the suppression of herd effect can not only effectively improve the efficiency of system resource utilization, but also maintain long-term stable sustainable resource supply.

In recent years, machine learning is advancing rapidly, especially as one of the learning paradigm reinforcement learning plays an important role in the field of artificial intelligence research. Reinforcement learning is an algorithm that makes optimal decisions through continuous interactive learning with the environment by trial and error. So far, it has been widely and successfully used in finance field [27–29], disease prediction [30, 31], game decision system [22, 32–35] and other research fields. It is exciting to note that reinforcement learning can also be applied to complex systems with multiple agents. For example, cooperative behavior can be promoted by Lévy noise and some interesting collective modes of oscillatory cooperation have emerged in evolutionary game systems with reinforcement learning [36–38]. This will provide a new perspective for the challenges of revealing the emergence mechanism of collective behavior. As shown in our previous work [39–41], the resource allocation systems combined Reinforcement learning with delayed rewards and minority game can emerge with self-restoring collective behavior to resist outside interference, and find that the system can evolve to the optimal resource allocation state without external intervention.

However, the reward mechanism or the duration of the reward can make a big difference in the effect of social group behavior based on a series of social surveys [42], and the authors also emphasized that weaker performers need to be rewarded immediately, rather than in the distant future to address their poverty. Inspired by this thought, we further investigate the emergence of collective behavior in resource allocation systems with the reinforcement learning manner under the condition of immediate rewards, and the decision-making. Without loss of generality, our multi-agent AI systems compete for limited resources in order to maximize their own payoff through Q-learning algorithms, and it is just that the reward mechanism is different from our previous work [39]. Some typical questions are whether the resource allocation can still be optimized under the immediate reward to agent, whether herding can be effectively inhibited, and what kind of paradigm of collective behavior can emerge. In this paper, we address these questions of how a multi-agent AI system can regulate the operation of the entire game system based on the suppression of adverse dynamic behavior. In addition, we also analyze the effects of various learning parameters and system size on the oscillatory evolution of collective behavior in the AI systems. Addressing these issues is of paramount importance because the establishment of such a framework can further understand the collective behavior emergence and its mechanism of large-scale multi-agent complex game systems. It is also the primary step to design human-machine system, which is the inevitable trend in the future.

This paper is organized as follows. In Section 2, we introduce reinforcement learning algorithms with the immediate reward mechanism into minority game systems (RLMG systems). To show the evolution of collective behavior of the AI systems, some major numerical simulation results are presented on multiple combinations of learning parameters in Section 3. We also revealed an interesting and important phenomenon that first-order phase transition appears with exploration rate in the multi-agent AI systems and two phases can also be observed (period-two oscillatory mode and non-period oscillatory mode) in Section 4.1. In order to further understand the learning process of the agent, we analyze the conversion path between belief modes in Section 4.2 and give the self-organizing condensation phenomenon of belief mode in the appendix. Moreover, a finding is that ergodic breaking of the system occurs and resource allocation performs better than random systems without external interference for a small exploration rate in Section 4.3. Then, in Section 4.4, a detection method for period-two oscillation collective pattern emergence based on the Kullback–Leibler (KL) divergence is introduced and give the parameter position where the period-two appears. Finally, a discussion and conclusion is provided in Section 5.

2 RLMG model

Without loss of generality, our system contains N agents competing for m limited resources denoted by r, and $r = 1, 2, ..., m$. The capacity of each resource is set C_r, which is the maximum number of agents that each resource can hold. For simplicity, the resource capacity

C_r is consistent, which is $C_r = N/m$. The vector $\boldsymbol{\rho}(t) = [\rho_1(t), \rho_2(t), \cdots, \rho_m(t)]^T$ to reflect the agent's preference for resources in time t. The component $\rho_r(t)$ represents the proportion of agents with resource r. If $N\rho_r(t) \leq C_r$, the individuals who chooses the resource r are winner because the number of agent dose not exceed the maximum load of the resource in this round. On the contrary, if $N\rho_r(t) > C_r$, the agents are loser due to resource overload in this round.

Q-learning is a typical reinforcement learning algorithm [43], it is used to combine with the minority game in our model. The goal of the agent empowered by Q-learning is to learn an optimal strategy through continuous interaction with the environment. That is, starting from the current state, it maximizes the expected value of the total reward in any and all subsequent steps. The relationship between the state of each agent, the actions available and the rewards obtained can be parameterized by the Q-function. Strategies that obtain higher rewards is strengthened by updating the Q-function in each round game. In order to seek for higher rewards, the agents always try to improve their wisdom via the Q-function and trial and error during the course of the game among agents. The update of the agent status and Q-function are just equivalent to the synchronization updating of Monte Carlo simulations [44, 45]. Each agent is empowered by reinforcement learning algorithm: Q-learning. The available state set is $S = \{s_1, \cdots, s_m\}$ and its elements indicate the resource label that each agent can occupy, and the action set is $\mathcal{A} = \{a_1, \cdots, a_m\}$ and its elements represent the actions that each agent can choose in any state s for Q learning algorithm. In most typical Q-learning tasks, the elements in $\mathcal{A}$ and S are distinguished by their environment properties. They are just same in our minority game model for both the action set and the state set.

The Q function is a time-dependent memory matrix combination $S \times \mathcal{A}$ as

$$\boldsymbol{Q}(t) = \begin{pmatrix} Q_{s_1 a_1}(t) & \cdots & Q_{s_1 a_m}(t) \\ \vdots & \ddots & \vdots \\ Q_{s_m a_1}(t) & \cdots & Q_{s_m a_m}(t) \end{pmatrix}.$$

The rows of this matrix represent the possible states of the agent, and the columns represent actions. If the agent i is in the state s and takes action a at the time step t, according to the Bellman equation, the element Q_{sa} of this matrix is updated as follows:

$$\begin{aligned} Q_{sa}(t+1) &= g(Q(t), R(t)) \\ &= Q_{sa}(t)(1-\alpha) + \alpha\left[R(t) + \gamma Q^{\max}_{s'a'}(t)\right], \quad (1) \end{aligned}$$

where the subscripts s and a denote the current state of agent and the action that the agent may take, respectively, $\alpha \in (0, 1]$ is the learning rate, and $R(t)$ is the reward obtained by the agent for taking action a in time step t. The parameter $\gamma \in [0, 1)$ is the discount factor

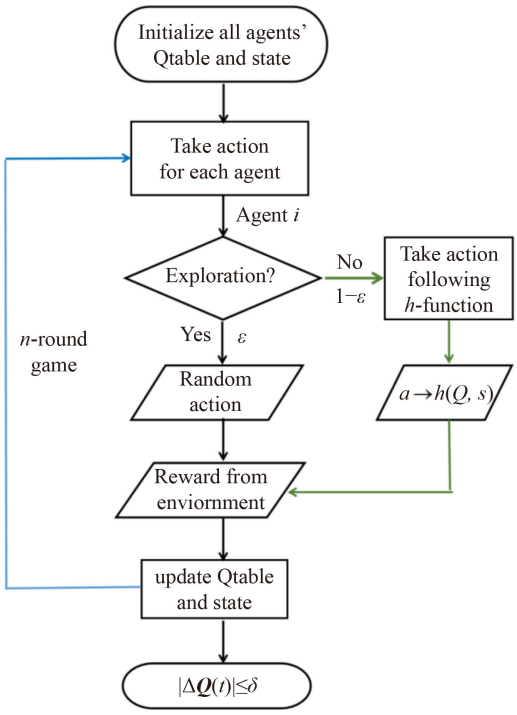


Fig. 1 The flow chart of protocol for Q-learning minority games. Green arrows indicates that agent i takes action following h-function in the logic diagram. δ is an arbitrarily given small constant and $\delta > 0$.

that measures the impact of future rewards. Agents with $\gamma = 0$ are short-sighted that they only care about current interests, while those with the larger value of γ are considered for a longer term. $Q^{\max}_{s'a'} = \max_{a'}(Q_{s'a'})$ is an optimal estimate of future values at s', the state s' is derived from the output of action a based on s at the current time step. According to Eq. (1), Q-function is the accumulation of historical experience through the iterative update of memory matrix elements, it reflects the quantitative relationship between state, action, and reward. As the time to explore the environment increases, the Q-function performs better and better based on the feedback of the action reward.

In our model, each agent in the system has its own memory matrix Q. As it starts to interact with the environment, the adaptability of the Q-function will increase rapidly. Just like the basic steps of Q-learning algorithm, while agents make decisions mostly based on their own memory matrix, it is also necessary to select the action with a certain randomness for trial and error. In summary, for a given set of parameters α and γ, the Q-learning algorithm is shown in Fig. 1:

1) Randomly initialize all agents' memory matrix Q and their state s.

2) Each agent takes the action following h-function with probability $1 - \epsilon$, that is

$$a(t) = h(Q(t), s)) = \arg \max_{a_1,\cdots,a_m \in \mathcal{A}} \{Q_{sa_1}(t), \cdots, Q_{sa_m}(t)\}, \quad (2)$$

or randomly selects an action with probability ϵ (also known as exploration rate) for each round. Meanwhile, if the action leading to the state is the current winning (minority) state, the agent gets a reward $R(t) = 1$ from the environment. Instead, the state is the failed (majority) state, the reward $R(t) = 0$.

3) The agent updates the element value Q_{sa} of memory matrix following Eq. (1), and its state is also updated as $s(t+1) = a(t)$.

4) Repeat the process 2)–3) until the system becomes stable or reaches a preset termination condition.

As stated in Ref. [42], differences in reward mechanisms can lead to very different dynamics behavior. If the feedback reward was delayed by a time step in our previous work [39], an explosion oscillation collective behavior that tends to the optimal state emerges in that system. However, in our current minority game system, the agent empowered by reinforcement learning receives an immediate reward for an action during the process of understanding environment. A remarkable feature is that the collective behavior presents the gentle oscillation around the optimal state without the explosion phenomenon, and as the exploration rate ϵ increases, the system undergoes a phase transition. In addition, our minority game system combined with RL is also different from other studied game systems [15], because our system takes into account the complex coupled memory taken action by agent in the past and adjusted by parameters α, γ and the reward $R(t)$ to optimize the allocation of resources.

3 Simulation results

In this paper, we focus on the case with two resources ($m = 2$) that the simplest minority game model with Q-learning agents. Thus, each agent could be in one of two resources (states) $S = \{s_1, s_2\}$ or $S = \{s_r | r = 1, 2\}$, and it could choose one resource as its action $\mathcal{A} = \{a_1, a_2\}$ or $\mathcal{A} = \{a_r | r = 1, 2\}$. The Q-function can be expressed as

$$\boldsymbol{Q}(t) = \begin{pmatrix} Q_{s_1a_1}(t) & Q_{s_1a_2}(t) \\ Q_{s_2a_1}(t) & Q_{s_2a_2}(t) \end{pmatrix}.$$

Actually, Q-function is the Cartesian product form about the state set S and the action set $\mathcal{A}$. The resource capacity is limited to $C_1 = C_2 = N/2$. On the evolution beginning, all the elements in agents' Q-function are randomly initialized in the interval of $(0, 1)$, and the agents' states are randomly specified from $\{s_1, s_2\}$.

Here, the occupation ratio of the resource r denoted by $\rho_r(t)$ is defined as

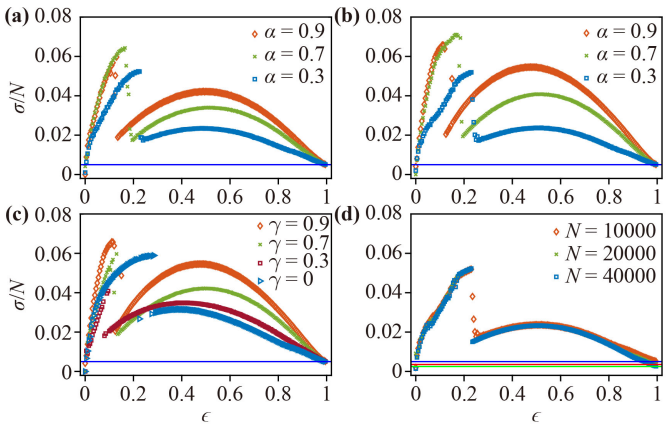


Fig. 2 Curves of σ as the function of the exploration rate ϵ with different learning rates and discount factors. The parameters α, γ and the system size N in each subplot are **(a)** $\gamma = 0.7$ and $N = 10000$; **(b)** $\gamma = 0.9$ and $N = 10000$; **(c)** $\alpha = 0.9$ and $N = 10000$; **(d)** $\alpha = 0.3$ and $\gamma = 0.9$. The straight line in each panel represents the standard deviation of a binomial random system, i.e., $\epsilon = 1$. (d) Three lines of different colors represent different system sizes. The Q function of each agent is initialized randomly before simulations.

$$\rho_r(t) = \frac{A_r(t)}{N}, \quad (3)$$

where $A_r(t) = \sum_{i=1}^{N} \delta(s^i(t) - s_r)$ is the number of agent selecting resource r, $s^i(t)$ is the state of agent i and δ is Dirac delta function, and $\delta = 1$ if the state of agent i is s_r at time t, otherwise $\delta = 0$. Obviously, the optimal resource allocation scheme is $\rho_r = C_r/N = 1/2$ in the RLMG system. In order to measure the performance of the RLMG system, a commonly used metric is

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T} (A_r(t) - C_r)^2, \quad (4)$$

which reflects the fluctuation of $A_r(t)$ deviating to the optimal allocation of resources over a long period [15]. In order to eliminate the influence of limited system size, the form σ/N is used.

Figure 2 shows σ of the RLMG systems as the function of the exploration rate ϵ. A striking result is that σ jumps discontinuously at a specific exploration rate denoted by ϵ_c, which is widely observed in different parameter combinations as shown in Figs. 2(a)–(c). This discontinuous jump indicates the dynamical behavior of the system undergoes a first-order phase transition and the transition behavior is further discussed in Section 4.1. In addition, for fixed discount factors $\gamma = 0.7$ and $\gamma = 0.9$ as shown in Figs. 2(a) and (b), the transition point ϵ_c decreases monotonically as α increases. Figure 2(c) shows that ϵ_c also relies on discount factors γ for a fixed learning rate α but it has no monotonic relation. For all combinations of learning parameters α and γ, σ is independent of system size except for the region ϵ close to 1,

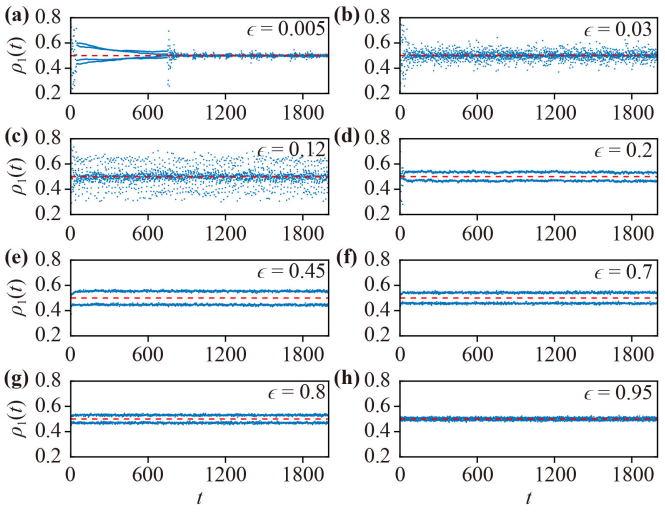


Fig. 3 The evolutionary time series $\rho_1(t)$ for the different ϵ. The exploration rates ϵs are smaller than ϵ_c in **(a, b)**. The time series close to the transition point ϵ_c is exhibited in **(c)**. Those for $\epsilon \in [\epsilon_c, 1]$ are shown in **(d–h)**. The red dashed line indicates the average occupation ratio C_r/N. The learning parameters and the number of agents are identical, i.e. $\alpha = 0.9$, $\gamma = 0.9$, and $N = 10000$.

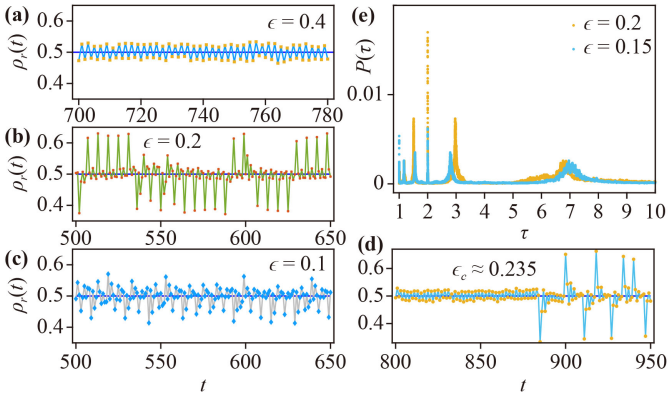


Fig. 4 Examples of several oscillating modes of resource load ρ_r are in the RLMG system, respectively $\epsilon = 0.4$, $\epsilon = 0.2$, $\epsilon = 0.1$, $\epsilon_c \approx 0.235$. The learning parameters are $\alpha = 0.3$ and $\gamma = 0.9$, and the system size is $N = 10000$. **(a)** A two period oscillation mode. **(b, c)** A non-period oscillation with different amplitude A. **(d)** The switch between period-two and other oscillation modes near ϵ_c. **(e)** The power spectrum of ρ_r with the period τ.

as the occupation behavior of agents is almost random and breaks the collective mode as shown in Fig. 2(d). From simulations of different learning parameters, it is find that the qualitative behaviors of σ with respect to ϵ are identical except for the position of the transition point ϵ_c.

The discontinuous phase transition breaks the regular oscillation of the time series $\rho_r(t)$. Taking learning parameters $\alpha = 0.9$ and $\gamma = 0.9$ [the orange diamond symbol in Fig. 2(b)] as a concrete example, the evolutionary time series of $\rho_r(t)$ near and far from the transition point ϵ_c are presented in Fig. 3. A stable period-two oscillation appears in the range $\epsilon > \epsilon_c$, while it is broken and turns into an irregular oscillation for $\epsilon < \epsilon_c$. In fact, stable period-two oscillation is a common collective behavior in minority game models [14, 15, 46], and it is a kind of herd behaviors which generates inefficient resource allocation. To suppress herd behaviors an outfield control, such as pin control [46], often needs to be applied. However, the RLMG system composed of AI agents can effectively suppress herding at very low exploration rate and organizes itself into an optimal resource allocation state by breaking regular oscillation as shown in Fig. 3(a). With the increase of ϵ, the amplitude of the irregular oscillation is enlarged rapidly, eventually leading to a stronger herding behavior. Figures 3(d)–(h) show temporal evolutions of $\rho_r(t)$ at the right of the transition point ϵ_c. As ϵ increases, the amplitude of the regular oscillation first increases and then decreases, which corresponds to the non-monotonic behavior of σ in Fig. 2 and the peak position of σ is denoted by ϵ_m. Further research shows that the amplitude of the regular oscillation increases with α and γ, respectively, and ϵ_m decreases with α, but there is no obvious pattern for discounting factors γ. For any given learning parameters combination α and γ, the amplitude and ϵ_m are independent of the system size N. As $\epsilon \to 1$, the amplitude of the periodic oscillation gradually decreases until it is dominated by noise, and $\rho_r(t)$ completely degenerates into a random oscillation, since all AI agents randomly access any resource r.

4 Analysis

4.1 The discontinuous transition

In this subsection, we investigate the jump of resource load standard deviation σ. A major question is whether the jump of σ means that the intelligent system undergoes a first-order phase transition. To answer this question, the collective oscillation modes on both sides of the transition points, shown in Figs. 4(a) and (b), have to be distinguished. At the right side of the transition point, $\rho_r(t)$ exhibits a relatively stable period-two oscillation mode (PTOM), for example $\epsilon = 0.4$ in Fig. 4(a), and this mode only maintains itself until the noise intensity overwhelms it at $\epsilon \to 1$. This is a robust dynamic phase in RLMG systems because this oscillation pattern can not be destroyed with finite noise, which is discussed in Section 4.4. It is worthy to mention that the amplitude of this oscillation mode shows a non-monotonic varying with the increase of ϵ, but the period-two oscillation is still maintained. However, the nature of this non-monotonic form of σ (a single oscillation mode determined only by

the oscillation amplitude) is the result of competition between two driving, namely, the AI agent's exploration (reflected by ϵ) and the pursuit of profit maximization (reflected by memory matrix Q-function). The position of σ reaching its maximum is denoted by ϵ_m. In the interval $(\epsilon_c, \epsilon_m]$, the revenue of agents dominates the evolution of the system, and the degree of herding behaviors are enhanced with ϵ. In the interval $(\epsilon_m, 1]$, the exploration of agents plays a leading role, which makes the elements in the Q-function to be randomly selected and updated. Therefore, herd effect is suppressed to some extent, but the optimization performance is not better than random systems.

On the left side of the transition point ϵ_c, there are abundant oscillatory evolution modes mainly manifested by the difference of oscillation frequency and amplitude in RLMG system. These oscillations are called non-period oscillation mode (non-POM). It should be emphasized that, for a given exploration rate $\epsilon \in (0, \epsilon_c)$, the evolution of $\rho_r(t)$ switches flexibly between different oscillatory modes instead of a fixed oscillatory evolution pattern as shown in Figs. 4(b) and (c). To depict the oscillatory behavior, we investigate the power spectral $P(\tau)$ of $\rho_r(t)$ which is defined as following:

$$P(\tau) = \lim_{T\to\infty} \frac{|F_T(2\pi/\tau)|^2}{2\pi T}, \tag{5}$$

where $F_T(2\pi/\tau)$ is the Fourier transform of $\rho_r(t)$ with length T. Figure 4(e) shows two examples of how the power spectrum $P(\tau)$ is distributed over the period τ on the left of ϵ_c. This is a special dynamic phase of MG systems with reinforcement learning agents. As the exploration rate approaches the transition point, an obvious switch occurs between the two phases as shown more intuitively in Fig. 4(d), i.e., PTOM phase and non-POM phase. One can see from Fig. 4 that the self-organized collective behavior emerges in the system for $\epsilon < 1$. In an interval of greater exploration rate $\epsilon \in (\epsilon_c, 1]$, the period-two oscillation mode keeps while the amplitude changes. The key factor is that in this exploration rate interval, the Q-function can self-organize into a very robust structure. In contrast, for a small exploration rate $\epsilon \in (0, \epsilon_c]$, the matrix structure is fragile, so that a variety of structures emerge and are owned by small systems of different sizes.

To analyze the transition between PTOM and non-POM, the order parameter Ω_r is defined as following:

$$\Omega_r = \frac{1}{T}\sum_{t=1}^{T} \text{sign}[\rho_r(2t) - C_r/N], \quad t = 1, 2, ..., T, \tag{6}$$

where sign(x) is the symbolic function that sign$(x) = 1$ for $x > 0$ and sign$(x) = -1$ for negative x. The length of the system time series that we take is $2T$. The order parameter Ω_r is within $[0, 1]$. If $\Omega_r = 1$, the system is oscillated with period two (PTOM), i.e., in ordered phase. Otherwise, if

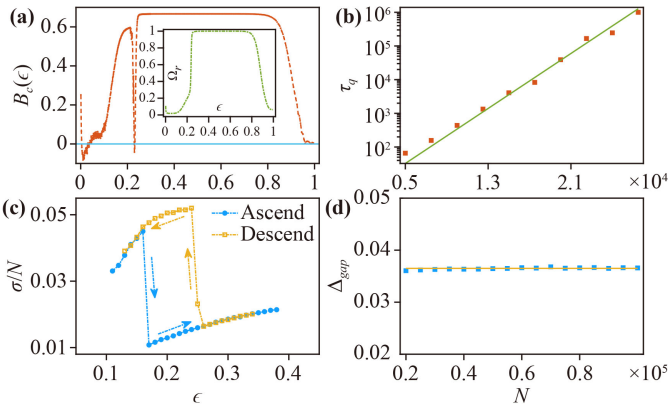


Fig. 5 Some phenomena near the phase transition point ϵ_c include critical slowing down, hysteresis like loop, Binder cumulant moment, etc. The learning parameters are $\alpha = 0.3$ and $\gamma = 0.9$, and the system size is $N = 10000$. **(a)** Binder cumulant moment $B_c(\epsilon)$ as a function of explore rate ϵ, the insert is Ω_r versus ϵ. **(b)** The critical state lifetime is taken as an exponential function of system size N, namely $\tau_q \sim \exp(\nu N)$. The solid green line is the fitting data. **(c)** σ/N with the hysteresis loop shown by ϵ, the blue solid circle is increasing for ϵ and the yellow square is decreasing for ϵ. **(d)** The gap Δ_{gap} of σ/N with system size N between bistable states near ϵ_c. The solid yellow line is the fitting data.

$\Omega_r = 0$, the period-two oscillation is broken, which means it turning into non-POM. Taking learning parameters $\alpha = 0.3$ and $\gamma = 0.9$ as an example, in which $\epsilon_c \approx 0.235$, Ω_r changes suddenly as shown in the illustration of Fig. 5(a), which is consistent with the jump point of σ in Fig. 2(b). As ϵ increases, the order parameter decreases gradually, which reflects the destruction effect of noise on the periodic oscillation, and the exploration rate is denoted by ϵ_k that the period-two oscillation is exhaustively broken for randomness of exploration.

According to Binder's research about phase transition [47, 48], the transition should be a first-order phase transition, if Binder cumulant moment of the order parameter changes to a negative value at the transition point ϵ_c. For the transition breaking PTOM, the Binder cumulant moment $B_c(\epsilon)$ of Ω_m is defined as

$$B_c(\epsilon) = 1 - \frac{\langle \Omega_r^4 \rangle}{3\langle \Omega_r^2 \rangle^2}. \tag{7}$$

$B_c(\epsilon)$ is shown as a function of ϵ in the main panel of Fig. 5(a). Obviously, when ϵ reaches $\epsilon_c = 0.235$ where the discontinuous jump of σ occurs, B_c changes suddenly to a negative value. This is a significant evidence that the system undergoes a first-order phase transition at ϵ_c. However, it is also noticed that B_c also decays to a negative value at $\epsilon \approx 0.01$. One possible reason is that there is additional complex oscillatory mode switching at very low noise intensity. Unfortunately, this switch is difficult

to detect in more detail.

This transition also has other general characteristics of the first-order phase transition. Constructing the evolution of system by slowly increasing (or decreasing) exploration rate ϵ, the hysteresis loop phenomenon of σ near ϵ_c is observed as shown in Fig. 5(c), which is in accordance with other discontinuous phase transitions [49–51]. Ref. [52] has emphasized that a more severe slowing-down occurs at the first-order phase transition point, since the free energy barriers separate the ordered and disordered phases. The dynamics of the system present an exponential slowing-down with the system size N near transition point. To detect the slowing-down form, the lifetime τ_q of the PTOM for a fixed system size N is defined as follows:

$$\tau_q = \frac{1}{n_q}\sum_{T'=0}^{n_q}\Gamma(T'), \tag{8}$$

where $\Gamma(T')$ is the total number of time steps of a complete PTOM between two adjacent switching about modes of oscillation, and n_q is the total number of PTOM in a long enough time series of $\rho_r(t)$. In our system, the lifetime τ_q is also found to have the exponential slowing-down at the transition point. That is, the duration of PTOM or non-POM as a exponential function of system size N, i.e., $\tau_q \sim \exp(\nu N)$ [see Fig. 5(b)]. This indicates that the RLMG system has structure of bistable state around the transition point, which is an typical feature of first-order transition. Figure 5(d) demonstrates that the gap of σ between PTOM and non-POM at transition point ϵ_c does not vanish with system size N.

4.2 The formation and transformation of belief mode

Because the AI agent has only two states at any time, namely s_1 or s_2, and the actions it can take for each state are a_1 or a_2. Each AI agent must take an action to participate in the minority game for maximizing its payoff based on feedback from the environment through reinforcement learning algorithm, and then update its Q-function and state. As we all know, AI agent constantly try and error with a certain probability ϵ to explore the optimal strategy in typical Q-learning algorithms, or take action following h-function with the probability $1-\epsilon$, see Fig. 1. If an AI agent makes decisions based on h-function, its Q-function structure is further solidified. Conversely, if the AI agent takes a random action to explore the environment, the Q-function structure is slightly broken down and reconstructed. Therefore, agent learning can be broken down into two processes, the former is called memory reinforcement updating process (mR-process) and the latter is trial and error or exploration updating process (tE-process). Specifically, if tE-process occurs with a small probability $(m-1)\epsilon/m \ll 1$, it can be regarded as a disturbance to mR-process.

In order to further understand the dynamic behavior of RLMG system, we tracked the learning process of a randomly selected AI agent adapting to the environment. The belief mode of AI agent based on Q-function and state s as $s|a$ is defined, which is the action a corresponding to the largest element Q_{sa} in the row of s. Therefor, for a fixed set of states S and fixed set of actions $\mathcal{A}$, the belief mode set is represented as $\mathcal{B} = \{s|a : s \in S, a \in \mathcal{A}\}$ and is time-dependent. The belief mode that an agent belongs to indicating a preference for a certain resource. That is, the AI agent's belief is the action a when the state is s at the current moment t for the belief mode $s|a$. To quantify the strength of AI agent preference for resources, we defined the belief mode function as follows:

$$|\mathcal{B}_{s|a_m}(t)| = \frac{1}{n-1}\sum_{a'\in\mathcal{A}\setminus a_m}^{n-1} Q(s,a_m)(t) - Q(s,a')(t), \tag{9}$$

$$\begin{aligned} a_m &= \arg\max_{a\in\mathcal{A}} Q(s,a) \\ &= \{a|\forall Q : Q(s,a_m) > Q(s,a)\}, \end{aligned} \tag{10}$$

where $|\mathcal{B}_{sa_m}|$ is the average gap between the largest element and other elements at in the row of state s. The larger the gap, the stronger the robustness of $s|a$ belief mode, which implies a more persistent preference of agent for action a_m. Based on Eq. (9), we find that the number of belief modes depends on the dimension of the Q-function. A larger Q-function dimension corresponds to a larger number of belief modes. For example, there are nine belief modes for a 3×3 matrix and eight belief modes for a 4×2 matrix.

In our system, only two resources are set, called A and B, then $s_1 = A$, $s_2 = B$ and $a_1 = A$, $a_2 = B$. Therefor, there are four belief modes: $A|A$, $A|B$, $B|A$ and $B|B$ for each AI agent in this two-resource RLMG system. For mR-process and tE-process, the transformation of belief mode $\mathcal{B}$ can be fully represented by a double-coupled directed network, in which belief mode is the node of the network and directed edge is the transformation direction between modes. The coupling between the two layers reflects the interaction of the two processes as shown in Fig. 6. Certain transitions occur between modes in the process of learning and adapting to the environment. The belief mode of the AI agent needs to go through the fixed evolution path, which includes two processes: mR-process (Solid arrow, the top layer in Fig. 6) and tE-process (Dash line, the down layer in Fig. 6). Each transformation between belief modes, including the self-transformation (namely, $s|a \to s|a$), is accompanied by the update of the corresponding Q matrix elements (the end of the arrow). If only according to the mR-process, AI agent's belief mode transformation path is as follows: $s_1|a_1 \to s_1|a_2 \to \cdots \to s_2|a_1 \to s_1|a_1$, which is an ordered

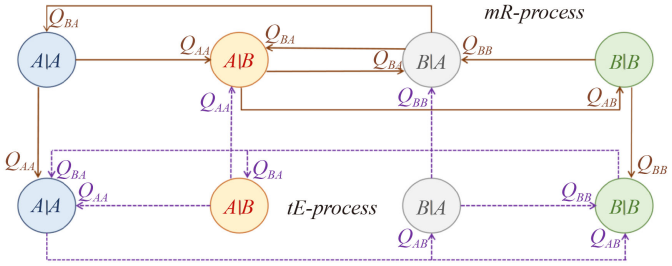


Fig. 6 Evolution scheme of belief mode in two-resource RLMG system. There are two types evolution paths of belief mode: *mR*-process and *tE*-process. In *mR*-process, the evolution paths of belief mode are represented by solid lines (the top layer). In *tE*-process, the evolution paths are represented by dash lines (the down layer). The arrows between the layers indicate that the AI agent's belief mode itself is maintained through either *mR*-process or *tE*-process. At the end of the arrow, the element Q_{sa} of Q-function is updated when the *mR*-process or *tE*-process occurred.

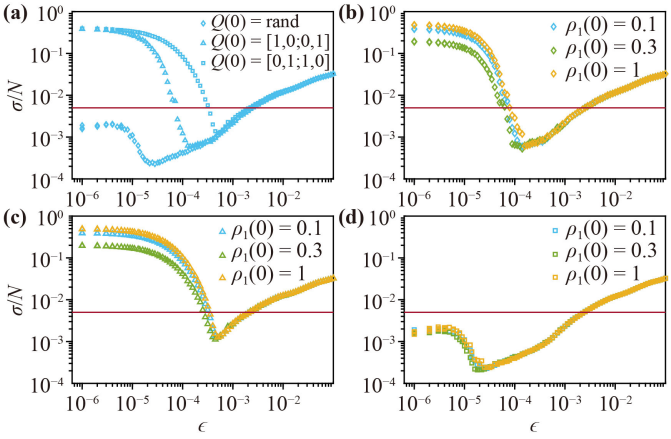


Fig. 7 The initialization sensitivity and the self-organization of optimal resources allocation. The learning parameters are $\alpha = 0.3$ and $\gamma = 0.9$, and the system size is $N = 10000$. **(a)** The initial occupation ratio of resource 1 is 0.1, i.e., $\rho_1(0) = 0.1$, and the Q-function has different initializations, which are $[1, 0; 0, 1]$, $[0, 1; 1, 0]$, and random in intervals $(0, 1)$. **(b–d)** The system has a given initialization Q-function and the initial occupation ratios $\rho_1(0)$ are different, respectively, $\rho_1(0) = 0.1$, $\rho_1(0) = 0.3$ and $\rho_1(0) = 1$. (b) The Q-function is randomly initialized in interval $(0, 1)$. (c) $\boldsymbol{Q}(0) = [1, 0, 0, 1]$, (d) $\boldsymbol{Q}(0) = [0, 1, 1, 0]$.

closed loop. However, the occurrence of *tE*-process destroys the orderliness of this closed loop and leads to a jump transformation between modes by adopting the opposite strategy of the h-function.

The belief mode of each AI agent in our formalization is reflected by its own Q matrix and current status. For a given game dynamics, the essence of agent learning is the mutual conversion of belief modes to maximize returns. As in our RLMG system, the $A|B$ and $B|A$ modes flow more toward the $B|B$ and $A|A$ modes until the system reaches a dynamic equilibrium for a fixed exploration rate ϵ shown in Fig. A1 of the Appendix, that is, the self-organizing condensation of belief modes appeared in the system. There exists strong nonlinear correlation effect between *mR*-process and *tE*-process in the process of agent adaptive reinforcement learning. Different from the noise disturbance of classical nonlinear dynamics system, *tE*-process not only perturbs the agent's current decision, but also perturbs or unlocks its belief mode which may be responsible for the oscillatory convergence phenomenon of the system. It will then indirectly affect the subsequent decision-making of the agent through the *mR*-process with greater probability. The evolutionary behavior of belief modes $\mathcal{B}_{s|a_m}(t)$ is discussed more detail by means of visualization in the appendix.

4.3 Initial sensitivity for low exploration rate

In this subsection, we focus on the dynamic behavior when the exploration rate ϵ approaches 0. An interesting phenomenon is that RLMG system shows more efficient resource allocation than a random system when ϵ is close to 0 for almost all possible combination of parameters (see Fig. 7). In other words, optimal allocation of resources gradually emerges under the condition of small noise in RLMG system. The solid red lines in Fig. 7 represent the resource load fluctuation of a complete random system, which is a constant $\sigma = 1/(2\sqrt{N})$. A large number of numerical results show that the RLMG system has an optimal exploration rate ϵ_o, i.e., the exploration rate corresponding to the minimum σ, where $\sigma(\epsilon_o) < 1/(2\sqrt{N})$. This is a very significant finding, which means the deviation of resource load σ is independent of system size N. For the constraint of Q-function in reinforcement learning algorithm with the small noise, the system can emerge spontaneously with a collective mode that effectively suppress herding behavior.

In addition to the self-organizing optimization, the collective behavior of the system also shows sensitivity to initial conditions with low exploration rates. The initial occupation ratio of any resource and initial element distribution of agents' Q-functions can both affect σ of future system. As shown in Fig. 7(a), for the Q-function initialization of three different structures, which are $\boldsymbol{Q}(0) = [1, 0; 0, 1]$, $\boldsymbol{Q}(0) = [0, 1; 1, 0]$, and random initializing all elements of $\boldsymbol{Q}(0)$ in range of $(0, 1)$, respectively. The σ presents different optimal position ϵ_o, while they all have the same initial occupation ratio $\rho_r(0) = 0.1$. In Figs. 7(b) and (c), the initialization structure of Q-function of the AI agents in the system is the same, but the $\rho_r(0)$s are different. The system is observed to have the same optimal position ϵ_o for a given $\boldsymbol{Q}(0)$ initialization, but the curves of $\sigma(\epsilon)$ are distinct. This phenomena are similar to the glass transition observed in some theoretical models of glass, that is, the system undergoes a

glass transition when the temperature, which corresponds to the exploration rate in our system, drops rapidly. The system undergoes a process from ergodic to ergodic breaking state. An interesting finding is that the optimal position ϵ_o of the system resources configuration is uniquely determined by the initialization of the $\boldsymbol{Q}$ table, independent of the initial proportion of the resource load $\rho_1(0)$. The relative value of the initial Q-function elements plays a major role in the position of ϵ_o.

4.4 The period-two oscillation mode breaking at $\epsilon = 1$

As mentioned in the previous Section 4.1, the period-two oscillation mode gradually loses stability and cannot be identified when exploration rate ϵ is close to 1 in the multi AI-agent system. An important question is whether this regular oscillation pattern (i.e., PTOM) can be destructed for finite exploration rate $\epsilon_k < 1$ or it is an finite size effect. To investigate this problem, the KL divergence D is introduced to depict the destruction behavior of the collective oscillation:

$$D(P(A_r)|\mathcal{N}_N(A_r)) = \int P(A_r) \log \frac{P(A_r)}{\mathcal{N}_N(A_r)} \, \mathrm{d}A_r. \quad (11)$$

Specifically, KL divergence is used to measure the difference between two distributions. In this case, the two distributions correspond to $P(A_r)$ that the numbers of agents A_r occupying resource r and $\mathcal{N}_N(A_r)$ that the distribution of A_r with $\epsilon = 1$, respectively. For period-two oscillation, $P(A_r)$ is a typical bimodal distribution. As ϵ approaches 1, $P(A_r)$ is going to degenerate to Gaussian distribution, denoted by $\mathcal{N}_N(A_r)$.

Figure 8(a) demonstrates the bimodal probability density function (PDF) $P(A_r)$ with $(1-\epsilon)\sqrt{N} = 14$ (green) and 24 (red). The emergence of bimodal structures distribution indicates that the multi-agent system is manipulated by the reinforcement learning algorithms, and thus deviated from the random system. Moverover, the bimodal structure further implies that a period-two oscillation mode emerges in the system through dynamical bifurcation. The gap between two peaks increases with the exploration rate decreasing, that is, the more distinct PTOM behavior can be observed. Different shapes of empty symbols represent different system sizes. Interestingly, it is found that the PDF of various system sizes can be re-scaled to a single curve by a simple relation $(A_r - C_r)/\sqrt{N}$. This is consistent with that σ is independent with system size N in period-two oscillation region, shown in Fig. 2(d).

Figure 8(b) shows the variation curve of KL divergence with ϵ in logarithmic coordinates. The KL deviation with the exploration rate ϵ re-scale relationship is given as following:

$$D(P(A_r)|\mathcal{N}_N(A_r)) \sim ((1-\epsilon)\sqrt{N})^{\mu} \quad (12)$$

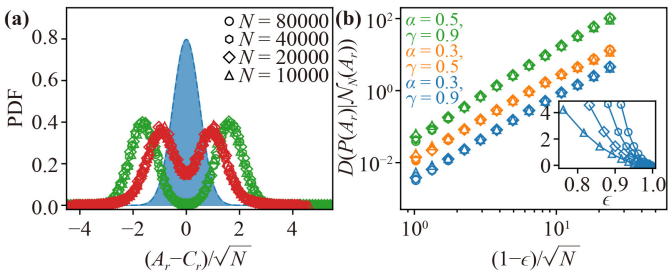


Fig. 8 As ϵ moves away from 1, the period-two oscillatory collective behavior of the system can be observed gradually. **(a)** It is statistical distribution of the resource load ρ_r for different system sizes, based on the ensemble average of a longer time series when the system reaches steady state, different color curves correspond to different exploration rates ϵ, the red empty symbol is $\epsilon = 0.76$, the green is $\epsilon = 0.85$ and the blue shade $\epsilon = 1$ (A completely random system). The system shares learning parameters $\alpha = 0.3$ and $\gamma = 0.9$. **(b)** It is KL Divergence of resource load ρ_r distributed between the given exploration rate $\epsilon \leq 1$ and a random system for different combinations of learning parameters α and γ. The insert shows the trend of KL deviation in linear coordinates without re-scaling. The system size $N = 10000$.

for different system sizes. This power law behavior can show that the oscillatory collective behavior of period-two does not suddenly appear under certain ϵ_k, but is always present in the interval $\epsilon \in (\epsilon_c, 1]$. This is analogous to the phase transition at the critical value $\epsilon = 1$, where only one phase exists in $\epsilon \in (\epsilon_c, 1]$, namely, PTOM. Therefore, the order Ω_m of the system begins to decrease at ϵ_k, simply because the strong noise drowns out the oscillating pattern of period-two. In addition, we found that the KL divergence is not independent of the system size when $\epsilon \to 1$, given by the different shape symbols of certain color, but also does not depend on learning parameter combinations, given by the different color symbols. For clarity, the gap between curves of different colors, including green symbol $\alpha = 0.5$, $\gamma = 0.9$, yellow symbol $\alpha = 0.3$, $\gamma = 0.5$ and blue symbol $\alpha = 0.3$, $\gamma = 0.9$, shown in Fig. 8(b) is caused by artificial translation.

5 Discussion and conclusion

The emergence of collective behavior from the various complex systems which consist of a large number of interacting components is universal in ecosystem, social and economic system. Collective behavior can be positive, such as efficient teamwork, bird foraging and animal migration, or negative, such as investor herding in the stock market, stampedes and traffic jams. The herding behavior is one of the typical collective behaviors of complex resource allocation systems and can spread quickly through the system causing it to collapse. The best performance of such a system is usually measured

by the sustainable and maximum use of resources by all individuals in the system. One of the core objectives of management resource allocation system is to avoid the occurrence of collective behavior with herd property through reasonable mechanism regulation. Machine learning paradigm provides a new perspective for the study of collective behavior emergence in complex systems. In addition, artificial intelligence will increasingly penetrate into all aspects of human society, such as self-driving car clusters and drone clusters, authorized by various machine learning paradigms in the future. Therefore, it is of great practical significance to explore the emergent of collective behavior and its mechanism through the marriage of AI and complex systems, especially the adaptive regulation of collective behavior evolution in complex systems consisting of a large number of AI agents. More generally, the main goal of our research is how machine learning in AI agents system with immediate reward affects collective dynamics in complex systems.

In this paper, we introduced the minority game, a typical model of resource allocation systems, incorporating reinforcement learning (RL): Q-Learning algorithm, by building individual simplified AI models. In our AI agents complex system, the individuals are intelligent to some extent, and they are capable of reinforcement learning. Intelligent groups can self-organize and evolve to reach a predetermined goal based on the feedback reward of the environment and the simplified Q-learning algorithm. With the exploration rate ϵ increasing (similar to the temperature of thermodynamic systems), an interesting transition phenomenon is also found in AI complex systems, that is first-order phase transition. Further research shows that two phases can also be observed for the AI system: period-two oscillatory evolution mode and non-period oscillatory evolution mode. This means that the system is regulated by the exploration rate ϵ of the learning parameter in the reinforcement learning algorithm, and the multi-agent system can emerge with two kinds of collective behavior (regular stable oscillation and irregular oscillation). As with traditional thermodynamic systems, a exponential slowing-down has also been discovered near the transition point ϵ_c. In short, different from the existing complex multi-agent game system with reinforcement learning, in our system, firstly, the agent has a simple memory matrix (Q-function) with low computational complexity. Secondly, each agent decision is independent, and the interaction between agents only comes from the feedback reward of the environment. Third, based on this simple model, intelligent game groups can realize self-organizing optimization for a given combination of parameters and switching behavior patterns.

In addition, an important phenomenon is that when the trial-and-error rate ϵ of the agent is very low, the system self-organize into a state of optimal resource allocation without any external intervention at ϵ_o, which is much more optimized than the random system. This indicates that large-scale complex systems composed of AI individuals can realize self-organization and collaboration based on individual interaction and reinforcement learning, and then emerge positive collective behaviors under certain learning parameters. The definition of belief mode and its transformation path give the emergent paradigm and evolutionary mechanism of the collective behavior of multi-agent systems from a more microscopic perspective. For example, for a specific parameter combination, with the advance of the learning process, mode $A|A$ and mode $B|B$ gradually dominate, while mode $A|B$ and $B|A$ gradually erode, but they will not disappear. By defining the belief modes approach, a clearer physical picture is presented to understand the evolution of the RLMG system dynamics behavior. We give the identity transformation path of an AI individual when it interacts with the environment, and the transformation path is universal regardless of the learning parameters for a given game dynamics. Then the emergence of period-two oscillation modes is demonstrated under high exploration rate, and concluding that the collective behavioral emergence of AI complex systems is qualitatively strongly dependent on the exploration rate ϵ and the learning parameter α, γ which can determine the point of the phase transition ϵ_c.

Finally, in our RLMG system there are four key exploration rates summarized below: ϵ_o, ϵ_c, ϵ_m, ϵ_k. The smallest ϵ_o implies an optimal resource allocation. ϵ_c represents the transition point of the first order phase transition, ϵ_m means the system enters an trade-off between the two driving, namely the Q-function and the noise effect, and the maximum ϵ_k indicates that the period-two oscillation mode is gradually destroyed by noise and cannot be clearly observed. As our theoretical results suggest that slightly different reward mechanisms have a great impact on the evolution of collective behavior in real social systems [42], for example, a delayed reward was set up in our previous work [39] which emerged the collective behavior with periodic explosions and can self-organize recovery against external disturbances. For immediate reward settings, however, the system exhibited a more varied and mildly oscillating behavior that optimizes resource allocation.

Our work can provide a basic theoretical framework for the integration of machine learning and complex systems. The systematic study of complex game systems from the perspective of machine learning can further help us to understand the emergence and evolution of diversified oscillating collective behavior patterns and their mechanisms. Because agents with a certain level of intelligence are closer to the activity pattern of biological groups. Based on our findings, there are still a lot of concrete unanswered questions that are very interesting. Different from the traditional game system, the agent with reinforcement learning ability can form a preference

belief mode for alternative actions by its Q-function. How the flow between belief mode drives the emergence of collective behavior in complex system is a significant and interesting question. In addition, from control point of view, by adjusting the learning parameters of the machine learning algorithm, system size or game parameters, it is of practical importance to make the system emerge positive collective behavior, or restrain unfavorable collective behavior.

Declarations The authors declare that they have no competing interests and there are no conflicts.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 12105213), China Postdoctoral Science Foundation (No. 2020M673363), and the Natural Science Basic Research Program of Shaanxi (No. 2021JQ-007).

Appendix: Self-organizing condensation of belief mode in RLMG system

In this section, we focus on the evolutionary behavior of belief modes as the agent becomes familiar with the environment. For simplicity, the resource configuration system still sets two resources as A and B. The set of belief mode contains only four elements: $\mathcal{B} = \{A|A, A|B, B|A, B|B\}$. In Section 4.2, the set of belief modes and the belief strength are defined for AI agents in RLMG game system [Eq. (9)]. One of our concerns is the evolution behavior of belief modes as the system moves from a random initial state to a steady state for the AI system. Here, a concrete example is given to shown the belief evolution in the AI system that the learning parameters are set to $\alpha = 0.3$, $\gamma = 0.9$ and $\epsilon = 0.01$.

We take snapshots of the system during training and observe the evolution of four belief modes as shown in Fig. A1. The occupancy of the agent for the four belief modes is nearly equal to about 1/4 as shown in Fig. A1(a), due to the random initialization of Q matrix and the state of the agent at the initial time $t = 1$. Belief modes begin to flow or transform between them with the evolution of AI system or agents learning because the elements in the Q-function are updated to adapt the dynamic environment. Therefor, the ratio of belief modes changes rapidly adaptation in the snapshot Fig. A1(b) at time $t = 50$. That is, the occupancy of $B|A$ mode is close to 0.5, and the occupancy of $A|B$ mode is close to 0. This indicates that the inflow in mode $B|A$ is greater than the outflow, and the net flow is positive. As the system is trained, for example, until time step $t = 257$ or a longer time in Figs. A1(c) and (d), the net flow of the mode $B|A$ or $A|B$ becomes negative. In fact, the property of belief modes $B|A$ and $A|B$, which is a speculative belief, is to cause the agent to change decisions

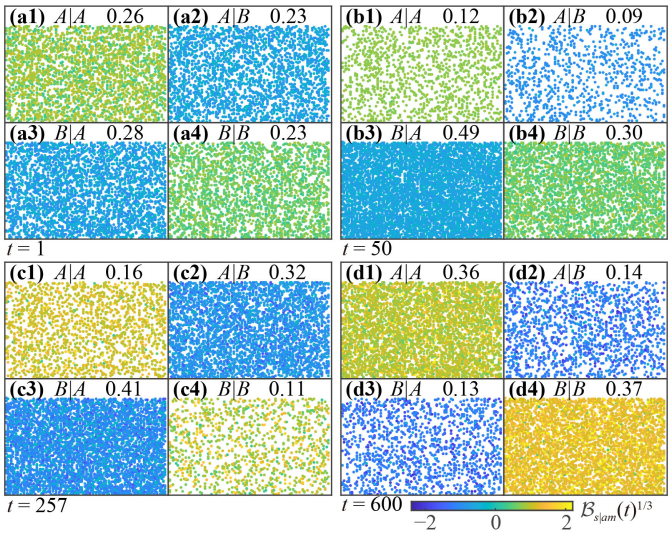


Fig. A1 Evolution snapshot that the agent's occupancy ratio of four belief modes for the RLMG systems in time step $t = 1$, $t = 50$, $t = 257$, $t = 600$, respectively. The learning parameters is $\alpha = 0.3$, $\gamma = 0.9$ and the system size $N = 10000$. All agents are fixed to nodes of the regular grid in the RLMG system. If the agent's belief mode is $s|a$, its correspond ing node is highlighted in grid with belief mode $s|a$. The color indicates belief strength $\mathcal{B}$ [Eq. (9)]. **(a1–d1)** The grids with $A|A$ belief mode. **(a2–d2)** The grids with $A|B$ belief mode. **(a3–d3)** The grids with $B|A$ belief mode. **(a4–d4)** The grids with $B|B$ belief mode. The number at the top of each subplot indicates the occupancy density of the belief at the current moment.

at adjacent moments, and trigger the emergence of a oscillations collective behavior. For certain reward mechanisms, the oscillating belief mode is absorbed by other belief mode to maximize overall RLMG system benefits. As the evolution of the RLMG system is close to stable, the occupancy of oscillatory modes of $A|B$ and $B|A$ decreases gradually, large-scale condensation of agents occurs on $A|A$ and $B|B$ modes. Therefore, modes $B|A$ and $A|B$ are more like a medium that can guide the system into a high ordered collective behavior for any learning parameters. But the resource load temporal $\rho_r(t)$ is disordered or non-period oscillation in the left of the transition point ϵ_c.

We summarize the evolutionary condensation of belief modes into the following steps with four properties: (i) The belief modes of all agents in the system are occupied uniformly and randomly at the initial time. (ii) With the self-organizing learning of the AI system, the belief modes $A|B$ and $B|A$ show large asymmetric fluctuations. The system is in a large oscillation, because the main occupation states of the agents at this stage are $A|B$ and $B|A$. (iii) In the third stage of system evolution, the proportion of belief mode states starts to change relatively smoothly. The oscillation modes $A|B$ and $B|A$ gradually flow to $A|A$ and $B|B$. Therefore, the resource preference

ρ_r is in a stable oscillation state, and the oscillation amplitude is gradually reduced. (iv) The system converges to a relatively stable state and forms a dynamic equilibrium. A large number of agents condense on $A|A$ and $B|B$ belief modes.

In a word, as the RLMG system is constantly trained, a large number of agents gradually self-organize on the $A|A$ and $B|B$ belief modes, and then reduce the fluctuation amplitude of the system. What is interesting is that this condensation of the $A|A$ and $B|B$ modes occurs for any given exploration rate ϵ, although the resource load $\rho_r(t)$ is a non-period oscillation to the left of the phase transition point.

References

1. D. J. Sumpter, Collective Animal Behavior, Princeton University Press, 2010
2. A. Procaccini, A. Orlandi, A. Cavagna, I. Giardina, F. Zoratto, D. Santucci, F. Chiarotti, C. K. Hemelrijk, E. Alleva, G. Parisi, and C. Carere, Propagating waves in starling, Sturnus vulgaris, flocks under predation, *Anim. Behav.* 82(4), 759 (2011)
3. H. King, S. Ocko, and L. Mahadevan, Termite mounds harness diurnal temperature oscillations for ventilation, *Proc. Natl. Acad. Sci. USA* 112(37), 11589 (2015)
4. C. R. Reid and T. Latty, Collective behaviour and swarm intelligence in slime moulds, *FEMS Microbiol. Rev.* 40(6), 798 (2016)
5. Y. T. Lin, X. P. Han, B. K. Chen, J. Zhou, and B. H. Wang, Evolution of innovative behaviors on scale-free networks, *Front. Phys.* 13(4), 130308 (2018)
6. L. M. Ying, J. Zhou, M. Tang, S. G. Guan, and Y. Zou, Mean-field approximations of fixation time distributions of evolutionary game dynamics on graphs, *Front. Phys.* 13(1), 130201 (2018)
7. N. T. Ouellette, A physics perspective on collective animal behavior, *Phys. Biol.* 19(2), 021004 (2022)
8. H. Murakami, M. S. Abe, and Y. Nishiyama, Toward comparative collective behavior to discover fundamental mechanisms underlying behavior in human crowds and nonhuman animal groups, *J. Robot. Mechatron.* 35(4), 922 (2023)
9. I. B. Muratore and S. Garnier, Ontogeny of collective behaviour, *Philos. Trans. R. Soc. Lond. B* 378(1874), 20220065 (2023)
10. Y. Liang and J. P. Huang, Robustness of critical points in a complex adaptive system: Effects of hedge behavior, *Front. Phys.* 8(4), 461 (2013)
11. W. B. Arthur, Inductive reasoning and bounded rationality, *Am. Econ. Rev.* 84(2), 406 (1994), 106th Annual Meeting of the American-Economic-Association, BOSTON, MA, JAN 03-05, 1994
12. D. Challet and Y. Zhang, Emergence of cooperation and organization in an evolutionary game, *Physica A* 246(3–4), 407 (1997)
13. T. Zhou, B. H. Wang, P. L. Zhou, C. X. Yang, and J. Liu, Self-organized Boolean game on networks, *Phys. Rev. E* 72(4), 046139 (2005)
14. Z. G. Huang, J. Q. Zhang, J. Q. Dong, L. Huang, and Y. C. Lai, Emergence of grouping in multi-resource minority game dynamics, *Sci. Rep.* 2(1), 703 (2012)
15. J. Q. Zhang, Z. G. Huang, J. Q. Dong, L. Huang, and Y. C. Lai, Controlling collective dynamics in complex minority-game resource-allocation systems, *Phys. Rev. E* 87(5), 052808 (2013)
16. J. Q. Dong, Z. G. Huang, L. Huang, and Y. C. Lai, Triple grouping and period-three oscillations in minority-game dynamics, *Phys. Rev. E* 90(6), 062917 (2014)
17. A. Cuesta, O. Abreu, and D. Alvear, Methods for measuring collective behaviour in evacuees, *Saf. Sci.* 88, 54 (2016)
18. X. H. Li, G. Yang, and J. P. Huang, Chaotic–periodic transition in a two-sided minority game, *Front. Phys.* 11(4), 118901 (2016)
19. L. Chen, Complex network minority game model for the financial market modeling and simulation, *Complexity* 2020, 8877886 (2020)
20. S. Biswas and A. K. Mandal, Parallel Minority Game and its application in movement optimization during an epidemic, *Physica A* 561, 125271 (2021)
21. T. Ritmeester and H. Meyer-Ortmanns, Minority games played by arbitrageurs on the energy market, *Physica A* 573, 125927 (2021)
22. B. Majumder and T. G. Venkatesh, Mobile data offloading based on minority game theoretic framework, *Wirel. Netw.* 28(7), 2967 (2022)
23. J. Linde, D. Gietl, J. Sonnemans, and J. Tuinstra, The effect of quantity and quality of information in strategy tournaments, *J. Econ. Behav. Organ.* 211, 305 (2023)
24. D. Carlucci, P. Renna, S. Materi, and G. Schiuma, Intelligent decision-making model based on minority game for resource allocation in cloud manufacturing, *Manage. Decis.* 58(11), 2305 (2020)
25. A. Swain and W. E. Fagan, Group size and decision making: experimental evidence for minority games in fish behaviour, *Anim. Behav.* 155, 9 (2019)
26. T. Ritmeester and H. Meyer-Ortmanns, The cavity method for minority games between arbitrageurs on financial markets, *J. Stat. Mech.* 2022(4), 043403 (2022)
27. Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, Deep direct reinforcement learning for financial signal representation and trading, *IEEE Trans. Neural Netw. Learn. Syst.* 28(3), 653 (2017)
28. Z. Jiang, D. Xu, and J. Liang, A deep reinforcement learning framework for the financial Portfolio management problem, arXiv: 1706.10059 (2017)
29. H. Yang, X. Y. Liu, S. Zhong, and A. Walid, in: *Proceedings of the First ACM International Conference on AI in Finance*, ICAIF'20, Association for Computing Machinery, New York, NY, USA, 2021
30. J. A. Cruz and D. S. Wishart, Applications of machine learning in cancer prediction and prognosis, *Cancer Inform.* 2, 59 (2007)
31. J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, *Proc. 27th Int. Conf. Neural Inf. Process. Syst.* 1, 1799 (2014)
32. 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, Mastering the game of Go with deep neural networks and tree search, *Nature* 529(7587), 484 (2016)

33. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, *Nature* 518(7540), 529 (2015)
34. H. Huang, Y. Cai, H. Xu, and H. Yu, A multiagent minority-game-based demand-response management of smart buildings toward peak load reduction, *IEEE Trans. Comput. Aided Des. Integrated Circ. Syst.* 36(4), 573 (2017)
35. M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 (2018)
36. 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.* 99(4), 3301 (2020)
37. L. Wang, D. Jia, L. Zhang, P. Zhu, M. Perc, L. Shi, and Z. Wang, Lévy noise promotes cooperation in the prisoner's dilemma game with reinforcement learning, *Nonlinear Dyn.* 108(2), 1837 (2022)
38. J. Xu, L. Wang, Y. Liu, and H. Xue, Event-triggered optimal containment control for multi-agent systems subject to state constraints via reinforcement learning, *Nonlinear Dyn.* 109(3), 1651 (2022)
39. S. P. Zhang, J. Q. Dong, L. Liu, Z. G. Huang, L. Huang, and Y. C. Lai, Reinforcement learning meets minority game: Toward optimal resource allocation, *Phys. Rev. E* 99(3), 032302 (2019)
40. S. P. Zhang, J. Q. Zhang, Z. G. Huang, B. H. Guo, Z. X. Wu, and J. Wang, Collective behavior of artificial intelligence population: Transition from optimization to game, *Nonlinear Dyn.* 95(2), 1627 (2019)
41. 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.* 99(4), 3301 (2020)
42. A. V. Banerjee and E. Duflo, Poor economics: A radical rethinking of the way to fight global poverty, Public Affairs, 2012
43. C. J. Watkins and P. Dayan, *Q*-learning, *Mach. Learn.* 8, 279 (1992)
44. M. Cao, A. S. Morse, and B. D. Anderson, Coordination of an asynchronous multi-agent system via averaging, *IFAC Proceedings Volumes* 38(1), 17 (2005)
45. H. L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi, Maximum likelihood reconstruction for Ising models with asynchronous updates, *Phys. Rev. Lett.* 110(21), 210601 (2013)
46. J. Q. Zhang, Z. G. Huang, Z. X. Wu, R. Su, and Y. C. Lai, Controlling herding in minority game systems, *Sci. Rep.* 6(1), 20925 (2016)
47. K. Binder, Theory of first-order phase transitions, *Rep. Prog. Phys.* 50(7), 783 (1987)
48. K. Binder, Applications of Monte Carlo methods to statistical physics, *Rep. Prog. Phys.* 60(5), 487 (1997)
49. G. Grégoire and H. Chaté, Onset of collective and cohesive motion, *Phys. Rev. Lett.* 92(2), 025702 (2004)
50. M. Nagy, I. Daruka, and T. Vicsek, New aspects of the continuous phase transition in the scalar noise model (SNM) of collective motion, *Physica A* 373, 445 (2007)
51. J. M. Encinas and C. E. Fiore, Influence of distinct kinds of temporal disorder in discontinuous phase transitions, *Phys. Rev. E* 103(3), 032124 (2021)
52. A. D. Sokal, Course 16 - Simulation of Statistical Mechanics Models, Elsevier, 2006