INTERFACE

royalsocietypublishing.org/journal/rsif

Research

Cite this article: Huang F, Cao M, Wang L. 2020 Learning enables adaptation in cooperation for multi-player stochastic games. *J. R. Soc. Interface* **17**: 20200639.
http://dx.doi.org/10.1098/rsif.2020.0639

Received: 9 August 2020
Accepted: 28 October 2020

Subject Category:
Life Sciences–Physics interface

Subject Areas:
computational biology, evolution, biophysics

Keywords:
reinforcement learning, evolutionary game theory, stochastic game, adaptive behaviour, social dilemma

Authors for correspondence:
Ming Cao
e-mail: m.cao@rug.nl
Long Wang
e-mail: longwang@pku.edu.cn

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.5202170.

THE ROYAL SOCIETY PUBLISHING

Learning enables adaptation in cooperation for multi-player stochastic games

Feng Huang[1,2], Ming Cao[2] and Long Wang[1]

[1]Center for Systems and Control, College of Engineering, Peking University, Beijing 100871, People's Republic of China
[2]Center for Data Science and System Complexity, Faculty of Science and Engineering, University of Groningen, Groningen 9747 AG, The Netherlands

FH, 0000-0002-6725-8263; MC, 0000-0001-5472-562X; LW, 0000-0001-5600-8157

Interactions among individuals in natural populations often occur in a dynamically changing environment. Understanding the role of environmental variation in population dynamics has long been a central topic in theoretical ecology and population biology. However, the key question of how individuals, in the middle of challenging social dilemmas (e.g. the 'tragedy of the commons'), modulate their behaviours to adapt to the fluctuation of the environment has not yet been addressed satisfactorily. Using evolutionary game theory, we develop a framework of stochastic games that incorporates the adaptive mechanism of reinforcement learning to investigate whether cooperative behaviours can evolve in the ever-changing group interaction environment. When the action choices of players are just slightly influenced by past reinforcements, we construct an analytical condition to determine whether cooperation can be favoured over defection. Intuitively, this condition reveals why and how the environment can mediate cooperative dilemmas. Under our model architecture, we also compare this learning mechanism with two non-learning decision rules, and we find that learning significantly improves the propensity for cooperation in weak social dilemmas, and, in sharp contrast, hinders cooperation in strong social dilemmas. Our results suggest that in complex social–ecological dilemmas, learning enables the adaptation of individuals to varying environments.

1. Introduction

Throughout the natural world, cooperating through enduring a cost to endow unrelated others with a benefit is evident at almost all levels of biological organisms, from bacteria to primates [1]. This phenomenon is especially true for modern human societies with various institutions and nation-states, in which cooperation is normally regarded as the first choice to cope with some major global challenges, such as curbing global warming [2,3] and governing the commons [4]. However, the mechanism underlying cooperative behaviour has perplexed evolutionary biologists and social economists for a long time [5,6]. Since according to the evolutionary theory of 'survival of the fittest' and the hypothesis of Homo economicus, this costly prosocial behaviour will be definitively selected against and should have evolved to be dominated by selfish act [7].

To explain how cooperation can evolve and be maintained in human societies or other animal groups, a large body of theoretical and experimental models have been put forward based on evolutionary game theory [6,8,9] and social evolution theory [10]. Traditionally, the vast majority of the previous work addressing this cooperative conundrum concentrates on the intriguing paradigm of a two-player game with two strategies, Prisoner's Dilemma [6,11]. Motivated by abundant biological and social scenarios where interactions frequently occur in a group of individuals, its multi-person version, the public goods game, has attracted much attention in recent years [12].

Meanwhile, it also prompts a growing number of researchers to devote to studying multi-player games and multi-strategy games [13–18]. However, these prominent studies implicitly assume, as most of the canonical work does, that the game environment is static and independent of players' actions. In other words, in these models, how players act by choosing game-play strategies only affects the strategic composition in the population, but the game environment itself is not influenced. As a result, a single fixed game is played repeatedly. Of course, this assumption is well grounded, if the timescale of interest (e.g. the time to fixation or extinction of a species) is significantly shorter than that of the environmental change. For most realistic social and ecological systems, however, it seems to be too idealized. Hence, an explicit consideration of environmental change is needed. A prototypical instance is the overgrazing of common pasture lands [19], where the depleted state may force individuals to cooperate and accordingly the common-pool resources will increase, whereas the replete state may induce defection and the common-pool resources will decrease [20,21]. Other examples also exist widely across scales from small-scale microbes to large-scale human societies [22]. A common feature of these examples is the existence of the feedback loop where individual behaviours alter environmental states, and are influenced in turn by the modified environment [20,23].

Although the effect of environmental variations on population dynamics has long been recognized in theoretical ecology and population biology [24–26], it is only recently that there has been a surge of interest in constructing game–environment feedbacks [20,21,23,27,28] to understand the puzzle of cooperation, especially in structured populations [29–31]. Different from the conventional set-up in evolutionary game theory [8,9], the key conceptual innovation of these works is the introduction of multiple games [32,33], evolving games [34,35], dynamical system games [36] or stochastic games [37,38]. By doing so, the players' payoff depends on not only strategic interactions but also the environmental state, and meanwhile, the fluctuation of the environment will be subject to the actions adopted by players. In this sense, the consideration of a dynamic game environment for the evolution of cooperation has as least two significant implications. First, it vastly expands the existing research scope of evolutionary game theory by adding a third dimension (multiple games) to the prior two-dimension space (multiple players and multiple strategies) [33]. That is, this extension generalizes the existing framework to encompass a broader range of scenarios. Second, the new key component, environmental feedbacks [20,23], is integrated seamlessly into the previous theoretical architecture.

While these promising studies primarily focused on pre-specified or pre-programmed behavioural policies to analyse the interdependent dynamics between individual behaviours and environmental variations, the key question of how individuals adjust their behaviours to adapt to the changing environment has not yet been sufficiently addressed. In fact, when confronting complex biotic and abiotic environmental fluctuations, how organisms adaptively modulate their behaviours is of great importance for their long-term survival efforts [25,39]. For example, those plants growing in the lower strata of established canopies can adjust their stem elongation and morphology in response to the spectral distribution of radiation, especially the ratio of red to far-red wavelength bands [40]; in arid regions, bee larvae, as well as angiosperm seeds, strictly comply with a bet-hedging emergence and germination rule such that reproduction activities are only limited to a short period of time following the desert rainy season [41]. Particularly, as an individual-level adaptation, learning through reinforcement is a fundamental cognitive or psychological mechanism used by humans and animals to guide action selections in response to the contingencies provided by the environment [42–44]. Employing the experience gained from historical interactions, individuals always tend to reinforce those actions that will increase the probability of rewarding events and lower the probability of aversive events. Although this learning principle has become a central method in various disciplines, such as artificial intelligence [44,45], neuroscience [43], learning in games [46] and behavioural game theory [47], there is still a lack of theoretical understanding of how it guides individuals to make decisions in order to resolve cooperative dilemmas.

In the present work, we develop a general framework to investigate whether cooperative behaviours can evolve by learning through reinforcement in constantly changing multi-player game environments. To characterize the interplay between players' behaviours and environmental variations, we propose a normative model of multi-player stochastic games, in which the outcome of one's choice relies on not only the opponents' choices but also the current game environment. Moreover, we use a social network to capture the spatial interactions of individuals. Instead of using a pre-specified pattern, every decision-maker in our model learns to choose a behavioural policy by associating each game outcome with reinforcements. By doing so, our model not only considers the environmental feedback, but also incorporates a cognitive or psychological feedback loop (i.e. players' decisions determine their payoffs in the game, and in turn are affected by the payoffs). When selection intensity is so weak that the action choices of players are just slightly influenced by past reinforcements, we derive the analytical condition that allows for cooperation to evolve under the threat of the temptation to defection. Through extensive agent-based simulations, we validate the effectiveness of the closed-form criterion in well-mixed and structured populations. Also, we compare the learning mechanism with two non-learning decision rules, and interestingly, we find that learning markedly improves the propensity for cooperation in weak social dilemmas whereas it hinders cooperation in strong social dilemmas. Furthermore, under non-stationary conditions, we analyse how cooperation coevolves with the environment and the effect of external incentives on the cooperative evolution by agent-based simulations.

2. Model and methods

2.1. Model

We consider a finite population of N individuals living in an evolving physical or social environment. The population structure describing how individuals interact with their neighbours is characterized by a network, where nodes represent individuals and edges indicate interactions. When individuals interact with their neighbours, only two actions, cooperation (C) and defection (D), are available, and initially, every individual is initialized with a random action in the set $\mathcal{A} = \{C, D\}$ with a certain probability. In each time step, one individual is

Table 1. Payoff table of the d-player stochastic game.

no. C co-players	$d-1$	…	j	…	0
C	$a_{d-1}(s)$	…	$a_j(s)$	…	$a_0(s)$
D	$b_{d-1}(s)$	…	$b_j(s)$	…	$b_0(s)$

chosen randomly from the population to be the focal player, and then $d-1$ of its neighbours as co-players are selected at random to form a d-player ($d \geq 2$) stochastic game [37,38]. To ensure that the game can always be organized successfully, we assume that each individual in the population has at least $d-1$ neighbours. Denote the possible number of C players among $d-1$ co-players by the set $\mathcal{J} \triangleq \{0, 1, \ldots, d-1\}$, and possible environmental states by the set $\mathcal{S} \triangleq \{s^1, s^2, \ldots, s^M\}$, where s^i, $i = 1, 2, \ldots, M$, represents the environmental state of type i. Then, depending on the co-players' configuration $j \in \mathcal{J}$ and the environmental state $s \in \mathcal{S}$ in the current round, each player will gain a payoff given in table 1. Players who take action C will get a payoff $a_j(s) \in \mathbb{R}$, whereas those who take action D will get a payoff $b_j(s) \in \mathbb{R}$, where $\mathbb{R}$ represents the set of real numbers. Players update their actions asynchronously; that is, in each time step, only the focal player updates its action, and other individuals still use the actions in the previous round. Furthermore, to prescribe the action update rule, we define the policy $\pi(s, j, a; \theta, \beta): \mathcal{S} \times \mathcal{J} \times \mathcal{A} \to [0, 1]$ with two parameters θ and β to specify the probability that action a is chosen by the focal player when there are j opponents taking action C among $d-1$ co-players in the environmental state $s \in \mathcal{S}$. Therein, $\theta \in \mathbb{R}^L$ is the column vector of L-dimension used for updating the policy by learning through reinforcement, and $\beta \in [0, +\infty)$ is the selection intensity [48], also termed the adaptation rate [49], which captures the effect of past reinforcements on the current action choice.

After each round, players' decisions regarding whether to cooperate or defect in the game interaction will not only influence their immediate payoffs but also the environmental state in the next round. That is to say, the probability of the environmental state in the next round is conditioned on the action chosen by the focal player and the environmental state in the current round. Without loss of generality, we here assume that the dynamics of environmental states $\{s_t\}$ obey an irreducible and aperiodic Markov chain, which thus possesses a unique stationary distribution. Also, from table 1, it is clear that the payoff of each player is a function of the environmental state. Therefore, when the environment transits from one state to another, the type of the (multi-player) normal-form game defined by the payoff table may be altered accordingly.

The emergence of the new environmental state in the next round, apart from influencing the game type, may also trigger players to adjust their behavioural policies. This is because those previously used decision-making schemes may no longer be appropriate in the changed environment. We here consider a canonical learning mechanism, the actor–critic reinforcement learning [42–44], to characterize the individual adaptation to the fluctuating environment. Specifically, after each round, the players' payoffs received from the game interaction will play a role of the incentive signal of the interactive scenario. If one choice gives rise to a higher return in a certain scenario, then it will be reinforced with a higher probability in the future when encountering the same situation again. By contrast, those choices resulting in lower payoffs will be weakened gradually. Technically, this process is achieved via updating the learning parameter θ of the policy after each round (see Methods for more details). In the successive round, the acquired experience will be shared within the population and the updated policy will be reused by the newly chosen focal player to determine which action to be taken. In a similar way, this dynamical process of game formation and policy updating is repeated infinitely (figure 1).

2.2. Methods

2.2.1. Actor–critic reinforcement learning

As the name suggests, the architecture of the actor–critic reinforcement learning consists of two modules. The actor module maintains and learns the action policy. Generally, there are two commonly used forms, ϵ-greedy and Boltzmann exploration [44,45]. Here, we adopt the latter for convenience, and consider the following Boltzmann distribution with a linear combination of features:

$$\pi(s, j, a; \theta, \beta) = \frac{e^{\beta \theta^{\mathrm{T}} \phi_{s,j,a}}}{\sum_{b \in \mathcal{A}} e^{\beta \theta^{\mathrm{T}} \phi_{s,j,b}}}, \quad \forall s \in \mathcal{S}, j \in \mathcal{J}, a \in \mathcal{A}, \tag{2.1}$$

where $\phi_{s,j,a} \in \mathbb{R}^L$ is the column feature vector with the same dimension of θ, which is handcrafted to capture the important features when a focal player takes action a given the environmental state s and the number of C players j among its $d-1$ co-players. Moreover, the dimension of the feature vector will in general be chosen to be much smaller than that of environmental states for the computational efficiency, i.e. $L \ll M$. For the construction of the feature vector, there are many options, such as polynomials, Fourier basis, radial basis functions and artificial neural networks [44]. As mentioned in the Model, β controls the selection intensity, or equivalently the adaptation rate. If $\beta \to 0$, it defines a weak selection and the action choice is only slightly affected by past reinforcements. When $\beta = 0$, in particular, players choose actions with uniform probability. By contrast, if $\beta \to +\infty$, the action with the maximum $\theta^{\mathrm{T}} \phi_{s,j,a}$ will be exclusively selected.

Another module is the critic, which is designed to evaluate the performance of the policy. In general, the long-run expected return of the policy per step, $\rho(\pi)$, will be a good measurement of the policy's performance, which is defined by

$$\rho(\pi) \triangleq \lim_{t \to \infty} \frac{1}{t} \mathbb{E}\{r_1 + r_2 + \cdots + r_t | \pi\}, \tag{2.2}$$

where $r_{t+1} \in \{a_{d-1}(s), \ldots, a_0(s), b_{d-1}(s), \ldots, b_0(s)\}$ is a random variable which denotes the payoff of the focal player at time $t \in \{0, 1, 2, \ldots\}$. In particular, if one denotes the probability that the environmental state at time t is s_t under the policy π when starting from the initial state s_0 by $Pr\{s_t = s \mid s_0, \pi\}$, and the average probability that all possible individuals chosen as the focal player encounter j opponents taking action C among $d-1$ co-players by $p_{\cdot j}$, then $\rho(\pi)$ can be computed by

$$\rho(\pi) = \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{j \in \mathcal{J}} p_{\cdot j} \sum_{a \in \mathcal{A}} \pi(s, j, a; \theta, \beta) \mathcal{R}^a_{s,j}, \tag{2.3}$$

where $d^{\pi}(s) = \lim_{t \to \infty} Pr\{s_t = s | s_0, \pi\}$ is the stationary distribution of environmental states under the policy π; $\mathcal{R}^a_{s,j}$ is the payoff of the focal player when it takes action a given the environmental state s and the number of C players j among its $d-1$ co-players, which is given by

$$\mathcal{R}^a_{s,j} = \begin{cases} a_j(s), & \text{if } a = C; \\ b_j(s), & \text{if } a = D. \end{cases} \tag{2.4}$$

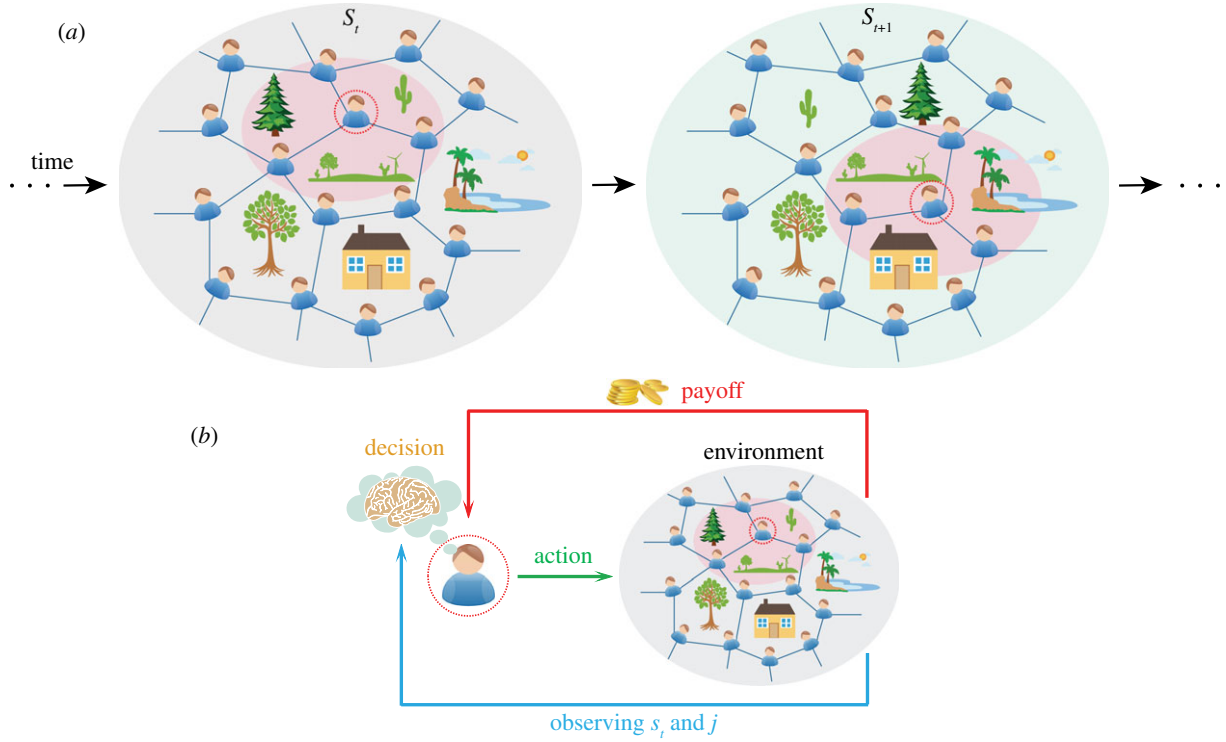

Figure 1. Illustration of evolutionary dynamics for 4-player stochastic games in the structured population. (*a*) At a time step t, a random individual is chosen as the focal player (depicted by the dashed red circle), and then three of its neighbours are selected randomly as co-players to form a 4-player game (because the focal player only has three neighbours, all of them are chosen), which is depicted by the light magenta shaded area. Conditioned on the focal player's action and the environmental state s_t at time t, the environmental state at time $t+1$ will be changed to s_{t+1} with a transition probability. Similarly, a new round of the game will be reorganized at time $t+1$. This process is repeated infinitely. (*b*) At time t, after perceiving the environmental state s_t and the co-players' configuration j, the focal player uses policy π to determine which action to be taken, whereas its co-players still use their previous actions in the past round. At the end of this round, each player will gain a payoff, which will play a role of the feedback signal and will assist the focal player to update its policy.

Moreover, to measure the long-term accumulative performance of the policy, we define a Q-value function,

$$Q^{\pi}(s, j, a) \triangleq \sum_{t=1}^{\infty} \mathbb{E}\{r_t - \rho(\pi) | s_0 = s, j_0 = j, a_0 = a, \pi\}, \ \forall s \in \mathcal{S}, \ j \in \mathcal{J}, a \in \mathcal{A}, \tag{2.5}$$

which is a conditional value depending on the initial environmental state $s_0 = s$, the number of C players $j_0 = j$ among $d-1$ co-players, and action $a_0 = a$ at time $t = 0$. Since the space of the environmental state is usually combinatorial and extremely large in many game scenarios, it is in effect impossible to calculate the Q-value function exactly for every environmental state within finite time with given computational resources [44]. Typically, one effective way to deal with this problem is to find a good approximation of the Q-value function. Let $f_w(s, j, a)$ be the approximation to the Q-value function and satisfy the compatibility condition [50,51]

$$\begin{aligned} f_w(s, j, a) &= w^{\mathrm{T}} \left[\frac{\partial \pi(s, j, a; \theta, \beta)}{\partial \theta} \frac{1}{\pi(s, j, a; \theta, \beta)} \right] \\ &= w^{\mathrm{T}} \left[\phi_{s,j,a} - \sum_{b \in \mathcal{A}} \pi(s, j, b; \theta, \beta) \phi_{s,j,b} \right] \beta, \end{aligned} \tag{2.6}$$

where $w \in \mathbb{R}^L$ is the column vector of weight parameters. To effectively approximate the Q-value function, it is natural to learn $f_w(s, j, a)$ by updating w via the least mean square method under the policy π. After acquiring the approximated measurement of the policy's performance $f_w(s, j, a)$, policy π can be then improved by following the gradient ascent of $\rho(\pi)$. Thus, the full algorithm of the actor–critic reinforcement learning can be given by (see electronic supplementary material SI.1 for details)

$$\left.\begin{aligned} w_{t+1} &= w_t + \alpha_t [r_{t+1} - \bar{R}_t + f_{w_t}(s_{t+1}, j_{t+1}, a_{t+1}) - f_{w_t}(s_t, j_t, a_t)] \frac{\partial f_{w_t}(s_t, j_t, a_t)}{\partial w_t}, \\ \theta_{t+1} &= \theta_t + \gamma_t \frac{\partial \pi(s_t, j_t, a_t; \theta_t, \beta)}{\partial \theta_t} \frac{1}{\pi(s_t, j_t, a_t; \theta_t, \beta)} f_{w_t}(s_t, j_t, a_t), \end{aligned}\right\} \tag{2.7}$$

where $\bar{R}_t$ is the estimation of $\rho(\pi)$, and iterates through $\bar{R}_{t+1} = \bar{R}_t + [r_{t+1} - \bar{R}_t]/(t+1)$ and $\bar{R}_0 = 0$, $t = 0, 1, 2, \ldots$; α_t and γ_t are learning step-sizes which are positive, non-increasing for $\forall t$, and satisfy $\sum_t \alpha_t = \sum_t \gamma_t = \infty$, $\sum_t \alpha_t^2 < \infty$ and $\sum_t \gamma_t^2 < \infty$, and $\gamma_t / \alpha_t \to 0$ for $t \to \infty$. These conditions required for the learning step-sizes guarantee that the policy parameter θ_t is updated at a slower timescale than that of the function approximation w_t, and thus ensure the convergence of the learning rule [51,52].

2.2.2. Evolution of cooperative behaviours

To capture the evolutionary process of cooperation, we first denote the number of C players in the population by n_t at

time t. Since there is only one individual to revise its action per step in our model, all possible changes of n_t in each time step will be limited to increasing by one, decreasing by one, or keeping unchanged. It implies that the evolutionary process of cooperation can be formulated as a Markov chain $\{n_t\}$ defined over the finite state space $\mathcal{N} = \{0, 1, 2, \ldots, N\}$. Meanwhile, the transition probability from $n_t = u \in \mathcal{N}$ to $n_{t+1} = v \in \mathcal{N}$ can be calculated by

$$p_{u,v}(t) = \sum_{s\in\mathcal{S}} Pr\{s_t = s|s_0, \boldsymbol{\pi}\} \sum_{j\in\mathcal{J}} \begin{cases} p_C p_{C,j}\pi(s, j, C; \theta_t, \beta) + p_D p_{D,j}\pi(s, j, D; \theta_t, \beta), & \text{for } v = u; \\ p_C p_{C,j}\pi(s, j, D; \theta_t, \beta), & \text{for } v = u-1; \\ p_D p_{D,j}\pi(s, j, C; \theta_t, \beta), & \text{for } v = u+1; \\ 0, & \text{otherwise,} \end{cases} \tag{2.8}$$

where $p_C = u/N$ (respectively, $p_D = (N-u)/N$) is the probability that an individual who previously took action C (respectively, D) is chosen as the focal player at time t; $p_{C,j}$ (respectively, $p_{D,j}$) is the average probability that players who previously took action C (respectively, D) encounter j opponents taking action C among $d-1$ co-players at time t. It is clear that the Markov chain is non-stationary because the transition probabilities change with time.

To find the average abundance of cooperators in the population, we first note that the actor–critic reinforcement learning converges [50,51] and the environmental dynamics have been described by an irreducible and aperiodic Markov chain. As such, we denote the limiting value of the policy parameter θ_t for $t \to \infty$ by θ^* (a local optimum of $\rho(\pi)$; see electronic supplementary material SI.1 for details), and the unique stationary distribution of environmental states by $d^\pi(s) = \lim_{t\to\infty} Pr\{s_t = s|s_0, \pi\}$. It follows that the probability transition matrix $P(t) = [p_{u,v}(t)]_{(N+1)\times(N+1)}$ will converge to $P^* = [p^*_{u,v}]_{(N+1)\times(N+1)}$ for $t \to \infty$, where

$$p^*_{u,v} = \lim_{t\to\infty} p_{u,v}(t) = \sum_{s\in\mathcal{S}} d^\pi(s) \sum_{j\in\mathcal{J}} \begin{cases} p_C p_{C,j}\pi(s, j, C; \theta^*, \beta) + p_D p_{D,j}\pi(s, j, D; \theta^*, \beta), & \text{for } v = u; \\ p_C p_{C,j}\pi(s, j, D; \theta^*, \beta), & \text{for } v = u-1; \\ p_D p_{D,j}\pi(s, j, C; \theta^*, \beta), & \text{for } v = u+1; \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

Moreover, it is noteworthy that the Markov chain described by the probability transition matrix P^* will be irreducible and aperiodic. This is because based on the matrix P^*, any two states of the Markov chain are accessible to each other and the period of all states is 1. Hence, one can conclude that the non-stationary Markov chain $\{n_t\}$ is strongly ergodic [53,54], and there exists a unique long-run (i.e. stationary) distribution $X = [x_n]_{1\times(N+1)}$, $n \in \mathcal{N}$. Therein, X can be obtained by calculating the left eigenvector corresponding to eigenvalue 1 of the probability transition matrix P^*, i.e. the unique solution to $X(P^* - I) = \mathbf{0}_{N+1}$ and $\sum_{n\in\mathcal{N}} x_n = 1$, where I is the identity matrix with the same dimension of P^* and $\mathbf{0}_{N+1}$ is the row vector with $N+1$ zero entries. When the system has reached the stationary state, the average abundance of C players in the population can be computed by $\langle x_C\rangle = \sum_{n\in\mathcal{N}} (x_n \cdot n/N)$. If $\langle x_C\rangle > 1/2$, it implies that C players are more abundant than D players in the population.

3. Results

3.1. Conditions for the prevalence of cooperation

We first study the condition under which cooperation can be favoured over defection, and restrict our analysis in the limit of weak selection ($\beta \to 0$) given that finding a closed-form solution to this problem for arbitrary selection intensity is usually NP-complete or # P-complete [55]. In the absence of mutations, such a condition can be obtained in general by comparing the fixation probability of cooperation with that of defection [48]. In our model, however, how players update their actions is conducted by the policy with an exploration–exploitation trade-off, which possesses a property similar to the mutation–selection process [56]. Thus, in this case, we need to calculate the average abundance of C players when the population has reached the stationary state, and determine whether it is higher than that of D players [57]. Using all $a_j(s)$ to construct the vector $A = [\mathbf{a}(s^1), \mathbf{a}(s^2), \ldots, \mathbf{a}(s^M)]^T$, and all $b_j(s)$ to construct the vector $B = [\mathbf{b}(s^1), \mathbf{b}(s^2), \ldots, \mathbf{b}(s^M)]^T$, where $\mathbf{a}(s^k) = [a_0(s^k), a_1(s^k), \ldots, a_{d-1}(s^k)]$ and $\mathbf{b}(s^k) = [b_{d-1}(s^k), b_{d-2}(s^k), \ldots, b_0(s^k)]$, $k = 1, 2, \ldots, M$, it follows that under weak selection the average abundance of C players in the stationary state is (see electronic supplementary material SI.2 for details)

$$\langle x_C\rangle = \frac{1}{2} + \frac{1}{N}\left[\sum_{s\in\mathcal{S}} d^\pi(s)\theta^{*\mathrm{T}}\Phi_s(A - B)\right]\beta + o(\beta), \tag{3.1}$$

and thus it is higher than that of D players if and only if

$$\sum_{s\in\mathcal{S}} d^\pi(s)\theta^{*\mathrm{T}}\Phi_s(A - B) > 0, \tag{3.2}$$

where Φ_s is the coefficient matrix corresponding to the environmental state s, and needed to be calculated for the given population structure, but independent of both $a_j(s)$ and $b_j(s)$ for $\forall j \in \mathcal{J}$ and $\forall s \in \mathcal{S}$.

To obtain an explicit formulation of condition (3.2), we further consider two specific population structures, well-mixed populations and structured populations. In the former case, the interactive links of individuals are described by a complete graph, whereas in the latter case, they are described by a regular graph with node degree $d-1$. When the population size is sufficiently large, we find that in the limit of weak selection, condition (3.2) in these two populations reduces to an identical closed form (see electronic supplementary material SI.3 for details):

$$\sum_{s\in\mathcal{S}} d^\pi(s) \sum_{j=0}^{d-1} \binom{d-1}{j} \frac{1}{2^{d+1}} \theta^{*\mathrm{T}}[\phi_{s,j,C} - \phi_{s,j,D}] > 0. \tag{3.3}$$

Through extensive agent-based simulations, we validate the effectiveness of this criterion. As illustrated in figure 2,

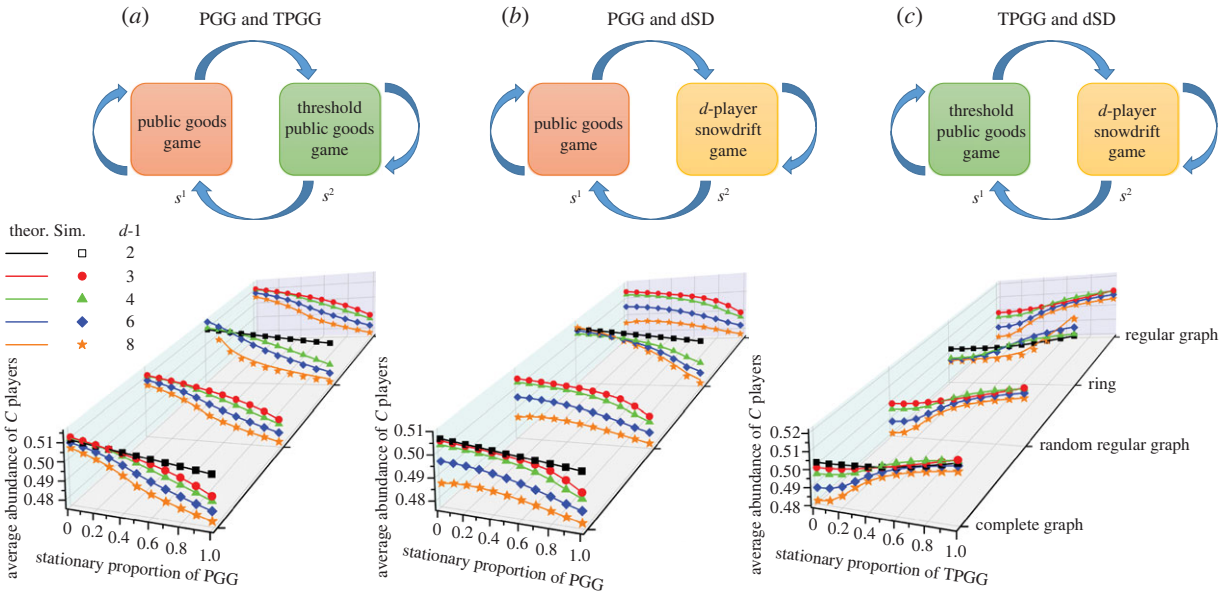

Figure 2. Average abundance of C players in the population as a function of the stationary proportion of different games. In each homogeneous environmental state, s^1 or s^2, one of the three normal-form games, PGG, TPGG and dSD, is played. In the top row, three transition graphs are depicted to describe how the environment transits from one state to another. Corresponding to these three transition graphs, the bottom row shows the average abundance of C players in various population structures, based on theoretical calculations and simulations. All simulations are obtained by averaging 40 network realizations and 10^8 time steps after a transient time of 10^7, and θ is normalized per step to unify the magnitude. The feature vector $\phi_{s,j,a}$ is chosen to be the one-hot vector. Parameter values: $N = 400$, $\beta = 0.01$, $\mathcal{C} = c = 1$, $r_s = 3$ in the PGG while $r_s = 4$ in the TPGG, $\mathcal{B}_s = 12$ in (b) while $\mathcal{B}_s = 4$ in (c), and $T = [d/2] + 1$ ($[\cdot]$ represents the integer part).

we calculate the average abundance of C players in the population with two distinct environmental states, s^1 and s^2, which, for instance, can represent the prosperous state and degraded state of a social–ecological system [20,58], respectively. To specify the type of the (multi-player) normal-form game defined by the payoff table 1 for each given environmental state, in figure 2, we consider that one of the three candidates, the public goods game (PGG) [19], threshold public goods game (TPGG) [3,59] and d-player snowdrift game (dSD) [60], is played in each state. In these three kinds of games, the implication of defection is unambiguous and it means not to contribute. However, in defining cooperation and calculating payoffs, there are some differences. In the PGG, action C means contributing a fixed amount c to the common pool. After a round of donation, the sum of all contributions from the d-player group will be multiplied by a synergy factor $r_s > 1$ and then allotted equally among all members, where r_s depends on the current game environment s. In this case, the payoffs of cooperators and defectors are computed by $a_j(s) = (j+1)r_s c/d - c$ and $b_j(s) = jr_s c/d$, $j \in \mathcal{J}$, respectively. The aforementioned setting is also true for the TPGG, except that there exists a minimum contribution effort, T, for players to receive benefits. More specifically, only when the number of C players in the d-player game is not smaller than T can each player receive a payoff from the common pool; otherwise, everyone gets nothing. It then follows that a C player will receive a payoff $a_j(s) = (j+1)cr_s/d - c$ for $j \geq T-1$ and $a_j(s) = 0$ otherwise, whereas a D player will receive $b_j(s) = jcr_s/d$ for $j \geq T$ and $b_j(s) = 0$ otherwise. Different from the PGG and TPGG, in the dSD, action C means endowing everyone with a fixed payoff $\mathcal{B}_s$ and simultaneously sharing a total cost $\mathcal{C}$ evenly with the other C players, where $\mathcal{B}_s$ depends on the environmental state s. In this case, the payoffs of cooperators and defectors are then changed to $a_j(s) = \mathcal{B}_s - \mathcal{C}/(j+1)$ for $j \in \mathcal{J}$, and $b_j(s) = \mathcal{B}_s$ for $j > 0$ and $b_0(s) = 0$, respectively. As shown in figure 2, the theoretical predictions for the average abundance of C players are highly consistent with simulation results, which suggests that criterion (3.3) is effective for determining whether cooperation can outperform defection.

Moreover, conditions (3.2) and (3.3) offer us an intuitional theoretical interpretation of why the environment can mediate social dilemmas [22]. As shown in figure 2, in an identical scenario, the average abundance of C players is always less than $1/2$ in the homogeneous state where the PGG is played, whereas it is greater than $1/2$ in some homogeneous states where a TPGG or dSD is played. The reason is that the social dilemma in the TPGG and dSD is weaker than that in the PGG. Thus, cooperation in these two kinds of games is easier to evolve. Namely, if the environment is homogeneous, condition (3.2) or (3.3) in the PGG is more difficult to be satisfied in contrast to the TPGG or dSD. Due to the existence of the underlying transition of the environment, however, the population may have some opportunities to extricate itself from those hostile environmental states where defection is dominant (e.g. the state of the PGG). This case is especially likely after some prosocial behaviours have been implemented by players [21,29,58]. As such, the population will spend some time staying in the states where defection is not always favourable (e.g. the TPGG or dSD). Consequently, the changing environment balances the conditions that favour versus undermine cooperation, and meanwhile the social dilemma that the population is confronted with is diluted. Such an observation is also in line with the fact that the final condition of whether cooperation can prevail is a convex combination of those results in each homogeneously environmental state, as shown in conditions (3.2) and (3.3).

3.2. Learning versus non-learning

Here, we exclude the effect of reinforcement learning, and apply our model framework to study two prototypical non-

learning processes of action choices, the smoothed best response [11] and the aspiration-based update [59,61]. For the former, in each time step, the focal player chosen in our model revises its action by comparing the payoff of cooperation with that of defection, and the more profitable action will be adopted. Instead of doing this in a deterministic fashion, in many real-life situations, it is more reasonable to assume that the choice of the best response is achieved smoothly and influenced by noise. One typical form to model this process is the Fermi function [11]:

$$\pi(s, j, a; \beta) = \frac{1}{1 + e^{-\beta[\mathcal{R}^a_{s,j} - \mathcal{R}^b_{s,j}]}}, \ \forall s \in \mathcal{S}, j \in \mathcal{J}, a, b(\neq a) \in \mathcal{A}, \tag{3.4}$$

which specifies the probability for the focal player to choose action $a \in \mathcal{A}$. For the latter, however, the focal player determines whether to switch to a new action by comparing the action's payoff with an internal aspiration level. If the payoff is higher than the aspiration level, the focal player will switch to that action with a higher probability. Otherwise, its action is more likely to keep unchanged. Similarly, the commonly used form to quantify the probability that the focal player switches to the new action $a \in \mathcal{A}$ is still the Fermi function [59,61]:

$$\pi(s, j, a; \beta) = \frac{1}{1 + e^{-\beta[\mathcal{R}^a_{s,j} - \mathcal{E})]}}, \ \forall s \in \mathcal{S}, j \in \mathcal{J}, a \in \mathcal{A}, \tag{3.5}$$

where a constant aspiration level $\mathcal{E}$ is adopted because heterogenous aspirations [61] or time-varying aspirations (see electronic supplementary material SI.4) cannot result in altering the evolutionary outcome under weak selection. Using these two non-learning update rules as the decision-making policy of the focal player, under our model framework, we find that in the limit of weak selection, cooperation is more abundant than defection if and only if

$$\sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{j \in \mathcal{J}} \sigma_j [a_j(s) - b_{d-1-j}(s)] > 0, \tag{3.6}$$

where $\sigma_j, \forall j \in \mathcal{J}$, are some coefficients needed to be calculated for the given population structure, but independent of both $a_j(s)$ and $b_j(s)$. In either well-mixed populations or structured populations, we find that the coefficients are $\sigma_j = \binom{d-1}{j}/2^{d+1}$ for the smoothed best response and $\sigma_j = \binom{d-1}{j}/2^{d+2}$ for the aspiration-based update (see electronic supplementary material SI.4 for details). In particular, if the population consistently stays in a fixed environment, condition (3.6) will reduce to the 'sigma-rule' of multi-player normal-form games [15].

In a population where there are three distinct environmental states and in each state one of the PGG, TPGG and dSD is played, we compare the results obtained by learning through reinforcement with those obtained from the two non-learning updates. As illustrated in figure 3, we calculate the average abundance of C players and the expected payoff of focal players per round for all possible stationary distributions of environmental states. Intriguingly, one can find that learning enables players to adapt to the varying environment. When the population stays in the environment where players are confronted with a weak social dilemma (i.e. the TPGG or dSD will be more likely to be played than the PGG), learning players will have a higher propensity for cooperation than those non-learning players. Meanwhile, they will reap a higher expected payoff per step. By contrast, when the population stays in the environment where the social dilemma is strong (i.e. the PGG will be more likely to be played than the TPGG and dSD), learning players will have a lower propensity for cooperation and accordingly they will get a lower expected payoff per step than non-learning players. Once again, we demonstrate that the analytical results are consistent with the agent-based simulations (see electronic supplementary material, figure S5).

3.3. Evolutionary dynamics under non-stationary conditions

The aforementioned analysis mainly focuses on the stationary population environment, i.e. the dynamics of environmental states have a unique stationary distribution and the payoff structure of the game does not change in time. Here, we relax this set-up to study the evolutionary dynamics of cooperation under two kinds of non-stationary conditions by agent-based simulations.

3.3.1. Non-stationary environmental state distribution

The first case that we are interested in is that the probability distribution of environmental states changes with time. In a population with two environmental states, s^1 and s^2, we denote the average proportion of the time that the environment stays in state s^1 (i.e. the average probability that the environment stays in s^1 per step) by $z \in [0, 1]$. Then, the average fraction of time in state s^2 is $1 - z$. To describe the type of game played in each environmental state, let s^1 be the prosperous state where environmental resources are replete and players are at the risk of the 'tragedy of the commons' (i.e. a PGG is played), whereas s^2 be the degraded state where environmental resources are gradually depleted. In both environmental states, cooperation is an altruistic behaviour that will increase the common-pool resources, whereas defection is a selfish behaviour that will lead the common-pool resources to be consumed. Furthermore, the state of common-pool resources (i.e. the environmental state) will conversely affect individual behaviours. To characterize this feedback relation, we here adopt the difference form of the replicator dynamics with environmental feedbacks [20,23] to describe the evolution of the average time proportion of state s^1:

$$\Delta z(t) = \eta z(t)(1 - z(t))(x_C(t) - \bar{x}_C), \tag{3.7}$$

where η denotes the positive step-size, $x_C(t)$ is the proportion of C players in the population at time t, and $\bar{x}_C$ is the tipping point of the proportion of C players. If the proportion of C players $x_C(t)$ is above the tipping point $\bar{x}_C$, it means that the number of cooperators is competent to sustain the supply of common-pool resources. At the same time, the environment will be more likely to stay in the prosperous state s^1, leading $z(t)$ to increase. Otherwise, cooperators are insufficient and the public resources will be continuously consumed. In this case, $z(t)$ will decrease as the environment will more frequently stay in the degraded state s^2.

We consider that in the prosperous state s^1 players play a PGG. However, in the degraded state s^2, one of four different games, the PGG, IPGG (inverse public goods game, which reverses the payoffs of action C and D in the PGG), dSH (d-player stag hunt game, which is a variant of the TPGG, and whose only difference from the TPGG is that cooperators

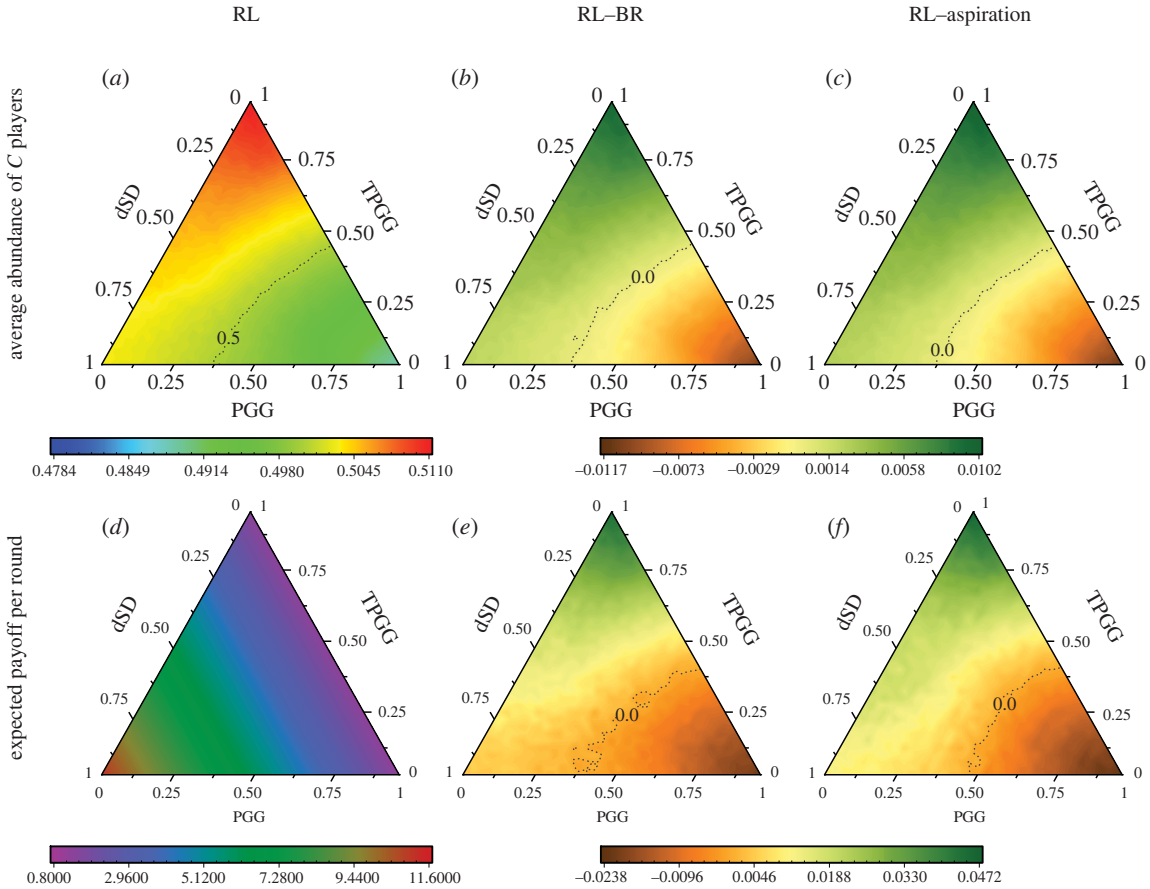

Figure 3. Differences in the average abundance of C players and the expected payoff of players per round between the reinforcement learning (RL) and two non-learning updates. In (*a*,*d*), we show the average abundance of C players and the expected payoff per round when players update actions via the RL, respectively. Taking them as the benchmark, (*b*,*e*) illustrate the differences between the RL and the smoothed best response (BR), while (*c*,*f*) show the gaps between the RL and the aspiration-based rule (aspiration). The population structure is a lattice network (see electronic supplementary material, figures S1–S4, for other population structures with different network degrees). Parameter values: $N = 400$, $d = 5$, $\beta = 0.01$, $\mathcal{C} = c = 1$, $T = [d/2] + 1$, $\mathcal{B}_s = 12$, $r_s = 3$ in the PGG while $r_s = 4$ in the TPGG.

always entail a cost c even if $j < T$), and dSD, is played. The reason that we select these four types of games is twofold. On the one hand, they are commonly used to mimic the essence of a vast number of real-life group interactions [12]; on the other hand, they encompass all possible evolutionary behaviours for the frequency-dependent selection between C and D under the classic replicator dynamics [9]: D dominance, C dominance, bistability, and coexistence (figure 4). Through agent-based simulations, in figure 4, we show the co-evolutionary dynamics of cooperation and the environment under moderate selection intensity. Depending on the game type and the value of the tipping point $\bar{x}_C$, the population emerges various dynamic behaviours. Particularly, although our model is stochastic and incorporates the effect of environment and learning, we can still observe those dominance, bistability and coexistence behaviours analogously obtained under the deterministic replicator dynamics. In addition, when replicator dynamics predict that cooperation will be the dominant choice in the degraded state s^2, our results show some persistent oscillations between cooperation and the environment (panel I in figure 4).

3.3.2. External incentives

Another interesting case is the existence of external incentives, which will undermine the stationarity of the payoff structure of the game. Like two sides of a coin, reward and punishment are two diametrically opposed external incentives for sustaining human cooperation [63,64]. The former is a type of positive incentives where players who cooperate will get an additional bonus, while the latter is a kind of negative incentives where those who defect will be sanctioned and need to pay a fine. At a certain moment during the evolution of cooperation, we separately implement punishment and reward, or jointly enforce them to all players in the population with four environmental states. One can observe that both punishment and reward are effective tools in promoting cooperation, even if the game environment may change (figure 5).

4. Discussion

In natural populations, the biotic and abiotic environment that organisms are exposed to varies persistently in time and space. To win the struggle for survival in this uncertain world, organisms have to timely adjust their behaviours in response to the fluctuation of their living environments [25,39]. For the longstanding conundrum of how cooperation can evolve, however, the majority of the existing evolutionary interpretations has been devoted to understanding the static interactive scenarios [1,6]. Therefore, when individual interactions, especially involving multiple players at a time, occur in the changing environment, determining whether cooperation can evolve will become fairly tricky. Here, we

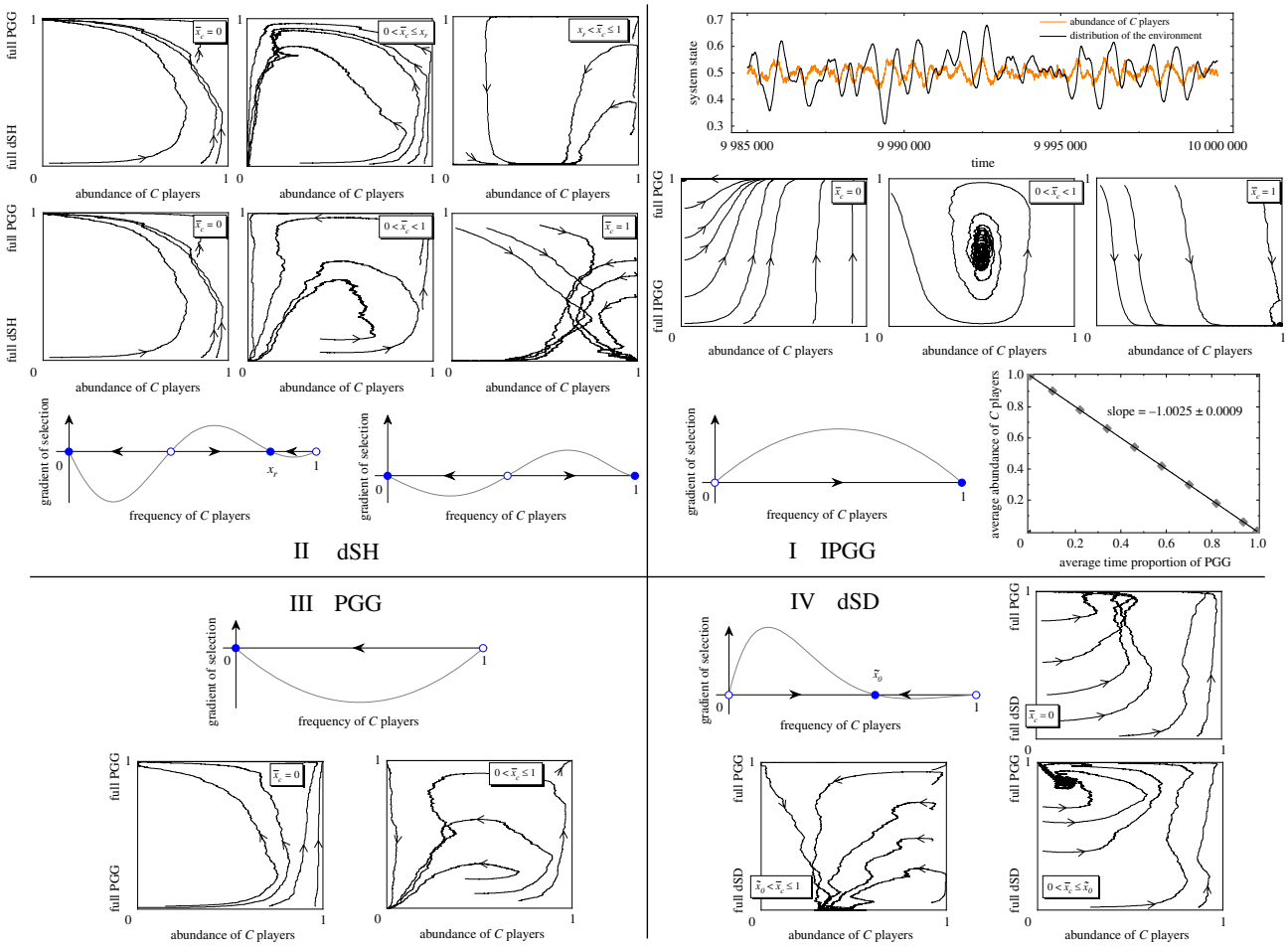

Figure 4. Co-evolutionary dynamics of cooperation and the environment under moderate selection intensity. From panel I to panel IV, the PGG is fixed to be played in state s^1, while in state s^2, the IPGG, dSH, PGG and dSD are played, respectively. Under replicator dynamics [12,60,62], the gradients of selection in these four games are shown in each panel, respectively. Blue solid circles are used to depict stable equilibria, while open blue circles are used to depict unstable equilibria. The direction of evolution is indicated by arrows. The phase graphs in each panel show the co-evolutionary dynamics of the time proportion of the PGG and the average proportion of C players for different value intervals of the tipping point $\bar{x}_C$. Corresponding to the value interval $0 < \bar{x}_C < 1$, the first row in panel I shows the persistent oscillations of cooperation and the environment. The bottom right sub-figure in panel I shows the linear relation between the average abundance of C players and the average time proportion of the PGG, which suggests that condition (3.3) is still valid for relatively moderate selection intensity. The first row in panel II uses the parameter condition [62] under which there is a stable and an unstable interior equilibrium for the dSH under replicator dynamics (the bottom left), whereas the second row uses that under which there is a unique interior unstable equilibrium (the bottom right). The population structure is a complete graph. Parameter values: $N = 400$, $d = 5$, $\beta = 2$, $\mathcal{C} = c = 1$, $\mathcal{B}_s = 12$, $r_s = 3$ for all panels, except in panel II, where $r_s = 4$ and $T = 3$ for the first row, and $r_s = 5.01$ and $T = 4$ for the second row.

developed a general model framework by introducing the adaptation mechanism of reinforcement learning to investigate how cooperation can evolve in the constantly changing multi-player game environment. Our model not only considers the interplay between players' behaviours and environmental variations, but also incorporates a cognitive or psychological feedback loop where players' choices determine the game outcome, and in turn are affected by it. Such a setting is, to some extent, analogous to the human decision in the context of the hybrid human–machine cooperation [65], a key research theme in the emerging interdisciplinary field of machine behaviour [66], in which humans can use algorithms to make decisions and subsequently the training of the same algorithms is affected by those decisions.

The importance of environmental variations in population dynamics has long been recognized in theoretical ecology and population biology [24–26]. In a realistic social or ecological system, individual behaviours and environmental variations are inevitably coupled together [24,25]. By consuming, transforming, or producing common-pool resources, for example, organisms are enabled to alter their living environments, and consequently, such modification may consequentially be detrimental or beneficial to their survival [22]. Our analytical condition for determining whether cooperation can be favoured over defection indeed provides us a plausible theoretical explanation for this phenomenon. If mutual actions of individuals lead the environment to transit from a preferable state where cooperation is more profitable to a hostile one where defection is more dominant, cooperation will be suppressed. By contrast, cooperation will flourish if the transition order is reversed. In particular, if the population has access to switching among multiple environmental states, the environment will play the role of intermediates in social interactions and the final outcome of whether cooperation can evolve will be the synthesis of results in each environmental state. Such an observation is different from the recent findings where game transitions can result in a more favourable outcome for cooperation even if all individual games favour defection [21,29]. One important reason for this is that we do not follow the

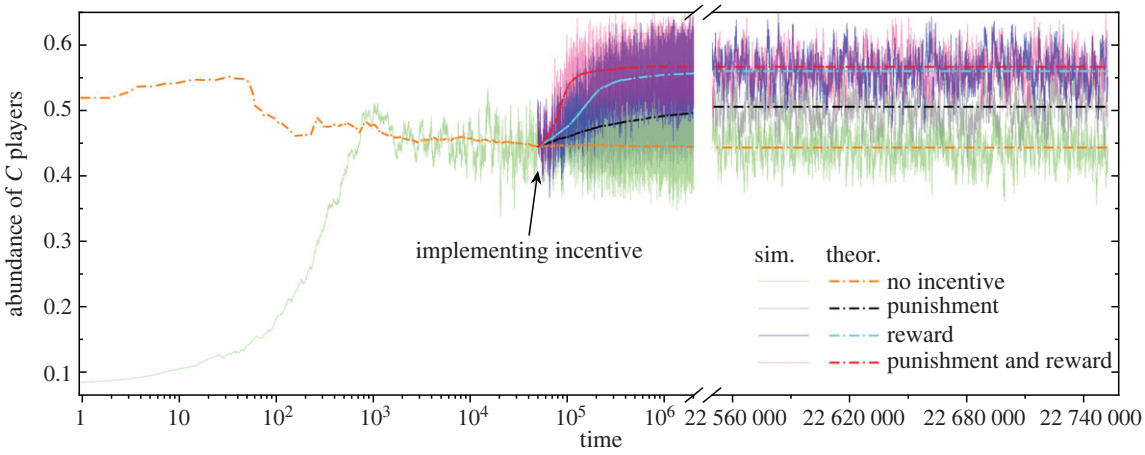

Figure 5. Evolution of cooperation under the influence of external incentives. Light solid lines indicate simulations whereas dash dot lines are theoretical results. During the evolution, we separately implement punishment (the fine is 0.4) and reward (the bonus is 0.65), or jointly enforce them in a population where the IPGG, dSH, PGG and dSD are played in each state with probability distribution (0.05, 0.05, 0.85, 0.05), respectively. The population structure is a lattice network. Parameter values: $N = 400$, $d = 5$, $\beta = 0.05$, $\mathcal{C} = c = 1$ for all games, except $r_s = 3$ in the PGG and TPGG, $r_s = 5$ and $T = [d/2] + 2$ in the dSH, and $\mathcal{B}_s = 12$ in the dSD.

scheme to explicitly assign a specific rule to prescribe the update of environmental states (i.e. model-specific methods), but rather simply assuming the ergodicity for the environmental dynamics (i.e. model-free methods). Thus, in this sense, our model is general and can be applied to a large variety of environmental dynamic processes.

Moreover, compared with the existing studies on the evolution of cooperation in the changing environment [20,21,23,27–29], another striking difference is that, apart from the environmental feedback, our model introduces the learning mechanism of reinforcement. Since, when the environment changes, the previous decision-making scheme adopted by individuals may fail to work, they must learn how to adjust their behaviours in response to the contingencies given by the environment, in order to obtain a higher fitness. Such a scenario is also closely related to some recent work across disciplines, including complexity science [49,67–70], artificial intelligence [44,56,71], evolutionary biology [72,73] and neuroscience [43]. However, their dominant attention has been paid to learning dynamics, the deterministic limit of the learning process, the design of new learning algorithms in games, or neural computations. In comparison, our model is discrete and stochastic, and focuses on multi-player stochastic games. In particular, our analysis for the game system is systematic and encompasses a variety of factors, such as group interactions, spatial structures and environmental variations. In addition, our work may offer some new insight into the interface between reinforcement learning and evolutionary game theory from the perspective of function approximation [44,50], because most existing progress in combining tools from these two fields to explore the interaction of multiple agents is based on value-based methods [49,56,70,71].

In the present work, one of the main limitations is that the strategic update is restricted to the asynchronous type and the learning experience is required to be shared among individuals. Although such a setting is appropriate in those scenarios where individuals modify their strategies independently, and typical in economics applications and for overlapping generations [11], it has been suggested that the unanimous satisfactory decisions reached by all asynchronous update individuals cannot always be guaranteed by synchronous updates [74]. In particular, if individuals are able to communicate with each other via a network or leverage the perceived information to model and infer the choices of others [45,47], the asynchronous update will suffer from some difficulties. Thus, further work on synchronously strategic revisions is worth exploring in the future. Of course, such an extension will also be full of challenges, because updating strategies concurrently for multiple agents will inevitably give rise to some complications, such as the curse of dimensionality, requirement for coordination, non-stationarity and exploration–exploitation trade-off [45]. Moreover, some further efforts should be invested in the partial observability of the Markov environmental states and relaxing the perfect environmental information required in our model to the unobservable or unpredictable type [75].

Data accessibility. This article has no additional data.

Authors' contributions. F.H., M.C. and L.W. participated in the design of the study and drafted the manuscript.

Competing interests. We declare we have no competing interest.

Funding. This work was supported by the National Natural Science Foundation of China (grant nos. 61751301 and 61533001). F.H. acknowledges the support from China Scholarship Council (grant no. 201906010075). M.C. was supported in part by the European Research Council (grant no. ERC-CoG-771687) and The Netherlands Organization for Scientific Research (grant no. NWO-vidi-14134).

Acknowledgements. The simulations were performed on the High-performance Computing Platform of Peking University.

References

1. West SA, Griffin AS, Gardner A. 2007 Evolutionary explanations for cooperation. *Curr. Biol.* **17**, R661–R672. (doi:10.1016/j.cub.2007.06.004)
2. Gardiner SM, Caney S, Jamieson D, Shue H (eds). 2010 *Climate ethics: essential readings*. Oxford, UK: Oxford University Press.
3. Milinski M, Sommerfeld RD, Krambeck H-J, Reed FA, Marotzke J. 2008 The collective-risk social dilemma and the prevention of simulated

dangerous climate change. *Proc. Natl Acad. Sci. USA* **105**, 2291–2294. (doi:10.1073/pnas.0709546105)

4. Ostrom E. 1990 *Governing the commons: the evolution of institutions for collective action*. Cambridge, UK: Cambridge University Press.
5. Colman AM. 2006 The puzzle of cooperation. *Nature* **440**, 744–745. (doi:10.1038/440744b)
6. Nowak MA. 2006 Five rules for the evolution of cooperation. *Science* **314**, 1560–1563. (doi:10.1126/science.1133755)
7. Dawkins R. 2016 *The selfish gene*. Oxford, UK: Oxford University Press.
8. Smith JM. 1982 *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.
9. Hofbauer J, Sigmund K. 1998 *Evolutionary games and population dynamics*. Cambridge, UK: Cambridge University Press.
10. Hamilton WD. 1964 The genetical evolution of social behaviour I and II. *J. Theor. Biol.* **7**, 1–52. (doi:10.1016/0022-5193(64)90038-4)
11. Szabó G, Fath G. 2007 Evolutionary games on graphs. *Phys. Rep.* **446**, 97–216. (doi:10.1016/j.physrep.2007.04.004)
12. Archetti M, Scheuring I. 2012 Game theory of public goods in one-shot social dilemmas without assortment. *J. Theor. Biol.* **299**, 9–20. (doi:10.1016/j.jtbi.2011.06.018)
13. Gokhale CS, Traulsen A. 2010 Evolutionary games in the multiverse. *Proc. Natl Acad. Sci. USA* **107**, 5500–5504. (doi:10.1073/pnas.0912214107)
14. Tarnita CE, Wage N, Nowak MA. 2011 Multiple strategies in structured populations. *Proc. Natl Acad. Sci. USA* **108**, 2334–2337. (doi:10.1073/pnas.1016008108)
15. Wu B, Traulsen A, Gokhale CS. 2013 Dynamic properties of evolutionary multi-player games in finite populations. *Games* **4**, 182–199. (doi:10.3390/g4020182)
16. Pena J, Wu B, Traulsen A. 2016 Ordering structured populations in multiplayer cooperation games. *J. R. Soc. Interface* **13**, 20150881. (doi:10.1098/rsif.2015.0881)
17. McAvoy A, Hauert C. 2016 Structure coefficients and strategy selection in multiplayer games. *J. Math. Biol.* **72**, 203–238. (doi:10.1007/s00285-015-0882-3)
18. Huang F, Chen X, Wang L. 2019 Evolutionary dynamics of networked multi-person games: mixing opponent-aware and opponent-independent strategy decisions. *New J. Phys.* **21**, 063013. (doi:10.1088/1367-2630/ab241b)
19. Hardin G. 1968 The tragedy of the commons. *Science* **162**, 1243–1248. (doi:10.1126/science.162.3859.1243)
20. Weitz JS, Eksin C, Paarporn K, Brown SP, Ratcliff WC. 2016 An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proc. Natl Acad. Sci. USA* **113**, E7518–E7525. (doi:10.1073/pnas.1604096113)
21. Hilbe C, Šimsa Š, Chatterjee K, Nowak MA. 2018 Evolution of cooperation in stochastic games. *Nature* **559**, 246–249. (doi:10.1038/s41586-018-0277-x)
22. Estrela S, Libby E, Van Cleve J, Débarre F, Deforet M, Harcombe WR, Peña J, Brown SP, Hochberg ME. 2019 Environmentally mediated social dilemmas. *Trends Ecol. Evol.* **34**, 6–18. (doi:10.1016/j.tree.2018.10.004)
23. Tilman AR, Plotkin JB, Akçay E. 2020 Evolutionary games with environmental feedbacks. *Nat. Commun.* **11**, 915. (doi:10.1038/s41467-020-14531-6)
24. MacArthur R. 1970 Species packing and competitive equilibrium for many species. *Theor. Popul. Biol.* **1**, 1–11. (doi:10.1016/0040-5809(70)90039-0)
25. Levins R. 1968 *Evolution in changing environments: some theoretical explorations*. Princeton, NJ: Princeton University Press.
26. Rosenberg NA. 2020 Fifty years of theoretical population biology. *Theor. Popul. Biol.* **133**, 1–12. (doi:10.1016/j.tpb.2020.04.001)
27. Chen X, Szolnoki A. 2018 Punishment and inspection for governing the commons in a feedback-evolving game. *PLoS Comput. Biol.* **14**, e1006347. (doi:10.1371/journal.pcbi.1006347)
28. Hauert C, Saade C, McAvoy A. 2019 Asymmetric evolutionary games with environmental feedback. *J. Theor. Biol.* **462**, 347–360. (doi:10.1016/j.jtbi.2018.11.019)
29. Su Q, McAvoy A, Wang L, Nowak MA. 2019 Evolutionary dynamics with game transitions. *Proc. Natl Acad. Sci. USA* **116**, 25 398–25 404. (doi:10.1073/pnas.1908936116)
30. Szolnoki A, Chen X. 2018 Environmental feedback drives cooperation in spatial social dilemmas. *Europhys. Lett.* **120**, 58001. (doi:10.1209/0295-5075/120/58001)
31. Szolnoki A, Perc M. 2019 Seasonal payoff variations and the evolution of cooperation in social dilemmas. *Sci. Rep.* **9**, 12575. (doi:10.1038/s41598-019-49075-3)
32. Hashimoto K. 2006 Unpredictability induced by unfocused games in evolutionary game dynamics. *J. Theor. Biol.* **241**, 669–675. (doi:10.1016/j.jtbi.2006.01.003)
33. Venkateswaran VR, Gokhale CS. 2019 Evolutionary dynamics of complex multiple games. *Proc. R. Soc. B* **286**, 20190900. (doi:10.1098/rspb.2019.0900)
34. Ashcroft P, Altrock PM, Galla T. 2014 Fixation in finite populations evolving in fluctuating environments. *J. R. Soc. Interface* **11**, 20140663. (doi:10.1098/rsif.2014.0663)
35. Stewart AJ, Plotkin JB. 2014 Collapse of cooperation in evolving games. *Proc. Natl Acad. Sci. USA* **111**, 17 558–17 563. (doi:10.1073/pnas.1408618111)
36. Akiyama E, Kaneko K. 2000 Dynamical systems game theory and dynamics of games. *Physica D* **147**, 221–258. (doi:10.1016/S0167-2789(00)00157-3)
37. Shapley LS. 1953 Stochastic games. *Proc. Natl Acad. Sci. USA* **39**, 1095–1100. (doi:10.1073/pnas.39.10.1953)
38. Neyman A, Sorin S (eds). 2003 *Stochastic games and applications*. Dordrecht, The Netherlands: Kluwer Academic Press.
39. Meyers LA, Bull JJ. 2002 Fighting change with change: adaptive variation in an uncertain world. *Trends Ecol. Evol.* **17**, 551–557. (doi:10.1016/S0169-5347(02)02633-2)
40. Ballaré CL, Scopel AL, Sánchez RA. 1990 Far-red radiation reflected from adjacent leaves: an early signal of competition in plant canopies. *Science* **247**, 329–332. (doi:10.1126/science.247.4940.329)
41. Danforth BN. 1999 Emergence dynamics and bet hedging in a desert bee, perdita portalis. *Proc. R. Soc. B* **266**, 1985–1994. (doi:10.1098/rspb.1999.0876)
42. Thorndike EL. 1911 *Animal intelligence: experimental studies*. New York, NY: Macmillan.
43. Niv Y. 2009 Reinforcement learning in the brain. *J. Math. Psychol.* **53**, 139–154. (doi:10.1016/j.jmp.2008.12.005)
44. Sutton RS, Barto AG. 2018 *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.
45. Busoniu L, Babuska R, De Schutter B. 2008 A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Syst. Man Cybernet. C* **38**, 156–172. (doi:10.1109/TSMCC.2007.913919)
46. Fudenberg D, Levine D. 1998 *The theory of learning in games*. Cambridge, MA: MIT Press.
47. Camerer CF. 2011 *Behavioral game theory: experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
48. Nowak MA, Sasaki A, Taylor C, Fudenberg D. 2004 Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650. (doi:10.1038/nature02414)
49. Sato Y, Akiyama E, Crutchfield JP. 2005 Stability and diversity in collective adaptation. *Physica D* **210**, 21–57. (doi:10.1016/j.physd.2005.06.031)
50. Sutton RS, McAllester DA, Singh SP, Mansour Y. 1999 Policy gradient methods for reinforcement learning with function approximation. In *Proc. 12th Int. Conf. on Neural Information Processing Systems*, pp. 1057–1063. Cambridge, MA: MIT Press.
51. Konda VR, Tsitsiklis JN. 1999 Actor-critic algorithms. In *Proc. 12th Int. Conf. on Neural Information Processing Systems*, pp. 1008–1014. Cambridge, MA: MIT Press.
52. Borkar VS. 1997 Stochastic approximation with two time scales. *Syst. Control Lett.* **29**, 291–294. (doi:10.1016/S0167-6911(97)90015-3)
53. Isaacson DL, Madsen RW. 1976 *Markov chains theory and applications*. New York, NY: John Wiley & Sons.
54. Bowerman BL. 1974 Nonstationary Markov decision processes and related topics in nonstationary Markov chains. PhD thesis, Iowa State University.
55. Ibsen-Jensen R, Chatterjee K, Nowak MA. 2015 Computational complexity of ecological and evolutionary spatial dynamics. *Proc. Natl Acad. Sci. USA* **112**, 15 636–15 641. (doi:10.1073/pnas.1511366112)
56. Tuyls K, Verbeeck K, Lenaerts T. 2003 A selection-mutation model for q-learning in multi-agent systems. In *Proc. 2nd Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2003)*,

pp. 693–700. New York, NY: ACM. (doi:10.1145/860575.860687)

57. Tarnita CE, Ohtsuki H, Antal T, Fu F, Nowak MA. 2009 Strategy selection in structured populations. *J. Theor. Biol.* **259**, 570–581. (doi:10.1016/j.jtbi.2009.03.035)
58. Barfuss W, Donges JF, Vasconcelos VV, Kurths J, Levin SA. 2020 Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proc. Natl Acad. Sci. USA* **117**, 12 915–12 922. (doi:10.1073/pnas.1916545117)
59. Du J, Wu B, Altrock PM, Wang L. 2014 Aspiration dynamics of multi-player games in finite populations. *J. R. Soc. Interface* **11**, 20140077. (doi:10.1098/rsif.2014.0077)
60. Souza MO, Pacheco JM, Santos FC. 2009 Evolution of cooperation under n-person snowdrift games. *J. Theor. Biol.* **260**, 581–588. (doi:10.1016/j.jtbi.2009.07.010)
61. Wu B, Zhou L. 2018 Individualised aspiration dynamics: calculation by proofs. *PLoS Comput. Biol.* **14**, e1006035. (doi:10.1371/journal.pcbi.1006035)
62. Pacheco JM, Santos FC, Souza MO, Skyrms B. 2009 Evolutionary dynamics of collective action in n-person stag hunt dilemmas. *Proc. R. Soc. B* **276**, 315–321. (doi:10.1098/rspb.2008.1126)
63. Fehr E, Fischbacher U. 2003 The nature of human altruism. *Nature* **425**, 785–791. (doi:10.1038/nature02043)
64. Perc M, Jordan JJ, Rand DG, Wang Z, Boccaletti S, Szolnoki A. 2017 Statistical physics of human cooperation. *Phys. Rep.* **687**, 1–51. (doi:10.1016/j.physrep.2017.05.004)
65. Crandall JW, Oudah M, Ishowo-Oloko F, Abdallah S, Bonnefon JF, Cebrian M, Shariff A, Goodrich MA, Rahwan I. 2018 Cooperating with machines. *Nat. Commun.* **9**, 233. (doi:10.1038/s41467-017-02597-8)
66. Rahwan I *et al.* 2019 Machine behaviour. *Nature* **568**, 477–486. (doi:10.1038/s41586-019-1138-y)
67. Macy MW, Flache A. 2002 Learning dynamics in social dilemmas. *Proc. Natl Acad. Sci. USA* **99** (Suppl. 3), 7229–7236. (doi:10.1073/pnas.092080099)
68. Sato Y, Akiyama E, Farmer JD. 2002 Chaos in learning a simple two-person game. *Proc. Natl Acad. Sci. USA* **99**, 4748–4751. (doi:10.1073/pnas.032086299)
69. Galla T, Farmer JD. 2013 Complex dynamics in learning complicated games. *Proc. Natl Acad. Sci. USA* **110**, 1232–1236. (doi:10.1073/pnas.1109672110)
70. Barfuss W, Donges JF, Kurths J. 2019 Deterministic limit of temporal difference reinforcement learning for stochastic games. *Phys. Rev. E* **99**, 043305. (doi:10.1103/PhysRevE.99.043305)
71. Bloembergen D, Tuyls K, Hennes D, Kaisers M. 2015 Evolutionary dynamics of multi-agent learning: a survey. *J. Artif. Intell. Res.* **53**, 659–697. (doi:10.1613/jair.4818)
72. Dridi S, Lehmann L. 2014 On learning dynamics underlying the evolution of learning rules. *Theor. Popul. Biol.* **91**, 20–36. (doi:10.1016/j.tpb.2013.09.003)
73. Dridi S, Akçay E. 2018 Learning to cooperate: the evolution of social rewards in repeated interactions. *Am. Nat.* **191**, 58–73. (doi:10.1086/694822)
74. Ramazi P, Riehl J, Cao M. 2016 Networks of conforming or nonconforming individuals tend to reach satisfactory decisions. *Proc. Natl Acad. Sci. USA* **113**, 12 985–12 990. (doi:10.1073/pnas.1610244113)
75. Kaelbling LP, Littman ML, Cassandra AR. 1998 Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134. (doi:10.1016/S0004-3702(98)00023-X)