PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Reinforcement Learning Explains Conditional Cooperation and Its Moody Cousin

Takahiro Ezaki[1,2,3,4], Yutaka Horita[3,4], Masanori Takezawa[5,6], Naoki Masuda[7]*

1 Research Center for Advanced Science and Technology, The University of Tokyo, Meguro-ku, Tokyo, Japan, **2** Japan Society for the Promotion of Science, Kojimachi, Chiyoda-ku, Tokyo, Japan, **3** National Institute of Informatics, Hitotsubashi, Chiyoda-ku, Tokyo, Japan, **4** JST, ERATO, Kawarabayashi Large Graph Project, c/o Global Research Center for Big Data Mathematics, NII, Chiyoda-ku, Tokyo, Japan, **5** Department of Behavioral Science, Hokkaido University, Kita-ku, Sapporo, Japan, **6** Center for Experimental Research in Social Sciences, Hokkaido University, Kita-ku, Sapporo, Japan, **7** Department of Engineering Mathematics, University of Bristol, Clifton, Bristol, United Kingdom

* naoki.masuda@bristol.ac.uk

OPEN ACCESS

Citation: Ezaki T, Horita Y, Takezawa M, Masuda N (2016) Reinforcement Learning Explains Conditional Cooperation and Its Moody Cousin. PLoS Comput Biol 12(7): e1005034. doi:10.1371/journal.pcbi.1005034

Editor: Natalia L. Komarova, University of California, Irvine, UNITED STATES

Received: March 21, 2016

Accepted: June 27, 2016

Published: July 20, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by: Japan Society for the Promotion of Science (Grant No. 13J05086 to TE and 15K13111 and 25285176 to MT). https://www.jsps.go.jp/english/index.html; and Japan Science and Technology Agency, Exploratory Research for Advanced Technology, Kawarabayashi Large Graph Project to MT and NM. http://www.jst.go.jp/erato/kawarabayashi/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Direct reciprocity, or repeated interaction, is a main mechanism to sustain cooperation under social dilemmas involving two individuals. For larger groups and networks, which are probably more relevant to understanding and engineering our society, experiments employing repeated multiplayer social dilemma games have suggested that humans often show conditional cooperation behavior and its moody variant. Mechanisms underlying these behaviors largely remain unclear. Here we provide a proximate account for this behavior by showing that individuals adopting a type of reinforcement learning, called aspiration learning, phenomenologically behave as conditional cooperator. By definition, individuals are satisfied if and only if the obtained payoff is larger than a fixed aspiration level. They reinforce actions that have resulted in satisfactory outcomes and anti-reinforce those yielding unsatisfactory outcomes. The results obtained in the present study are general in that they explain extant experimental results obtained for both so-called moody and non-moody conditional cooperation, prisoner's dilemma and public goods games, and well-mixed groups and networks. Different from the previous theory, individuals are assumed to have no access to information about what other individuals are doing such that they cannot explicitly use conditional cooperation rules. In this sense, myopic aspiration learning in which the unconditional propensity of cooperation is modulated in every discrete time step explains conditional behavior of humans. Aspiration learners showing (moody) conditional cooperation obeyed a noisy GRIM-like strategy. This is different from the Pavlov, a reinforcement learning strategy promoting mutual cooperation in two-player situations.

Author Summary

Laboratory experiments using human participants have shown that, in groups or contact networks, humans often behave as conditional cooperator or its moody variant. Although conditional cooperation in dyadic interaction is well understood, mechanisms underlying

Competing Interests: The authors have declared that no competing interests exist.

these behaviors in group or networks beyond a pair of individuals largely remain unclear. In this study, we show that players adopting a type of reinforcement learning exhibit these conditional cooperation behaviors. The results are general in the sense that the model explains experimental results to date obtained in various situations. It explains moody conditional cooperation, which is a recently discovered behavioral trait of humans, in addition to traditional conditional cooperation. It also explains experimental results obtained with both the prisoner's dilemma and public goods games and with different population structure. Crucially, our model assumes that individuals do not have access to information about what other individuals are doing such that they cannot explicitly condition their behavior on how many others have previously cooperated. Thus, our results provide a proximate and unified understanding of these experimentally observed patterns.

Introduction

Humans very often cooperate with each other when free-riding on others' efforts is ostensibly lucrative. Among various mechanisms enabling cooperation in social dilemma situations, direct reciprocity, i.e., repeated interaction between a pair of individuals, is widespread. If individuals will repeatedly interact, they are motivated to keep on cooperation because no cooperation would invite retaliation by the peer in the succeeding interactions [1, 2]. Past theoretical research using the two-player prisoner's dilemma game (PDG) identified tit-for-tat (TFT) [2], generous TFT [3], a win-stay lose-shift strategy often called Pavlov [4–6] as representative strong competitors in the repeated two-player PDG.

Direct reciprocity in larger groups corresponds to the individual's action rule collectively called the conditional cooperation (CC), a multiplayer variant of TFT. By definition, an individual employing CC would cooperate if a large amount of cooperation has been made by other group members. In the present study, we study a reinforcement learning model. Depending on the parameter values, the outcome of the learning process shows CC patterns and their variant that have been observed in behavioral experiments.

In fact, the following evidence suggests that the concept and relevance of CC are much more nuanced than in the case of dyadic interactions, calling for examinations. First, early theoretical studies have concluded that CC in the multiplayer PDG is unstable as the group size increases [7, 8]. In addition, CC assumed in these and follow-up studies is a threshold behavior (i.e., players cooperate when the number of peers that have cooperated the last time exceeds a prescribed threshold). However, CC patterns and their variants observed in the extant experiments are gradual rather than a threshold behavior [9–17].

Second, the public goods game (PGG) models social dilemmas occurring in a group beyond a pair of individuals. In the repeated PGG, CC has been observed in laboratory experiments with human participants [9–12, 17] and in real society [18]. By definition, an individual adopting CC increases the amount of cooperation when others have made large contributions the last time. CC in the repeated PGG with two or more players is theoretically stable under some conditions [19–24]. However, these conditions are not generous and how they connect to the experimental results is not clear. This situation contrasts to that of the two-player, discrete-action PDG mentioned previously, where conditions under which direct reciprocity occurs and strategies enabling them are well characterized [2, 3, 5].

Third, recent experiments have discovered that humans show moody conditional cooperation (MCC) behavior in the repeated PDG game played on contact networks [13–16]. MCC is defined as follows. MCC is the same as CC if the player has cooperated the last time. If the

player has defected the last time, a player adopting MCC decides on the action without taking into account what the neighbors in the contact network have done previously. In this sense, the player's action rule is moody. The genesis of MCC is not well understood. First, evolutionary dynamics do not promote MCC behavior [15, 25]. Second, non-evolutionary numerical simulations assuming MCC do not intend to explain why MCC emerges or is stable [14, 15, 26]. Third, a numerical study employing reinforcement learning [25] has MCC behavior built in into the model in the sense that MCC occurs whenever cooperation is sustained (see Discussion for more).

In this article, we provide an account for experimentally observed CC and MCC patterns using a family of reinforcement learning called the aspiration learning [27–36]. In reinforcement learning, players satisfice themselves rather than maximize the payoff in the sense that a player increases and decreases the likelihood of the behavior that has yielded a large and small reward, respectively. In aspiration learning, players are satisfied if and only if the obtained payoff is larger than a threshold. Because the probability to select the behavior, such as cooperation, is dynamically updated in every discrete time step, aspiration learning is different from a conditional strategy in general.

Our main conclusion that reinforcement learning explains CC and MCC resembles that of a previous study [25]. However, the present study is radically different from Ref. [25] in the following aspects. First, as stated above, MCC behavior is an assumed mode of the model proposed in Ref. [25]. In the present model, players myopically adjust the unconditional probability of cooperation depending on the previous action and reward, as in previous aspiration learning models [28, 29, 31, 32, 37–39]. Second, the present model is also simpler, even without assuming players to be aware of the amount of cooperation carried out nearby or to explicitly implement conditional strategies.

Model

We place a player obeying the reinforcement learning rule on each node of the square lattice with 10×10 nodes with periodic boundary conditions. However, the following results do not require particular network structure (Fig A in S1 Text). Each player is involved in the two-player PDG against each of the four neighbors on the network. The game is also interpreted as a PGG played in the group composed of the player and all neighbors submitting binary decisions [40]. The game is repeated over $t_{\max}$ rounds. We set $t_{\max} = 25$ unless otherwise stated.

Each player selects either to cooperate (C) or defect (D) in each round (Fig 1A). The submitted action (i.e., C or D) is used consistently against all the neighbors. In other words, a player is not allowed to cooperate with one neighbor and defect against another neighbor in the same round. If both players in a pair cooperate, both players gain payoff $R = 3$. If both defect, both gain $P = 1$. If a player cooperates and the other player defects, the defector exploits the cooperator such that the cooperator and defector gain $S = 0$ and $T = 5$, respectively.

Each player is assumed to update the intended probability to cooperate, p_t, according to the Bush-Mosteller (BM) model of reinforcement learning [27–29, 32, 39] as follows:

$$p_t = \begin{cases} p_{t-1} + (1 - p_{t-1})s_{t-1} & (a_{t-1} = \mathrm{C}, s_{t-1} \geq 0), \\ p_{t-1} + p_{t-1}s_{t-1} & (a_{t-1} = \mathrm{C}, s_{t-1} < 0), \\ p_{t-1} - p_{t-1}s_{t-1} & (a_{t-1} = \mathrm{D}, s_{t-1} \geq 0), \\ p_{t-1} - (1 - p_{t-1})s_{t-1} & (a_{t-1} = \mathrm{D}, s_{t-1} < 0), \end{cases} \quad (1)$$

where a_{t-1} is the action in the $(t-1)$th round, and s_{t-1} is the stimulus that drives learning $(-1 < s_{t-1} < 1)$. The current action is reinforced and suppressed if $s_{t-1} > 0$ and $s_{t-1} < 0$,

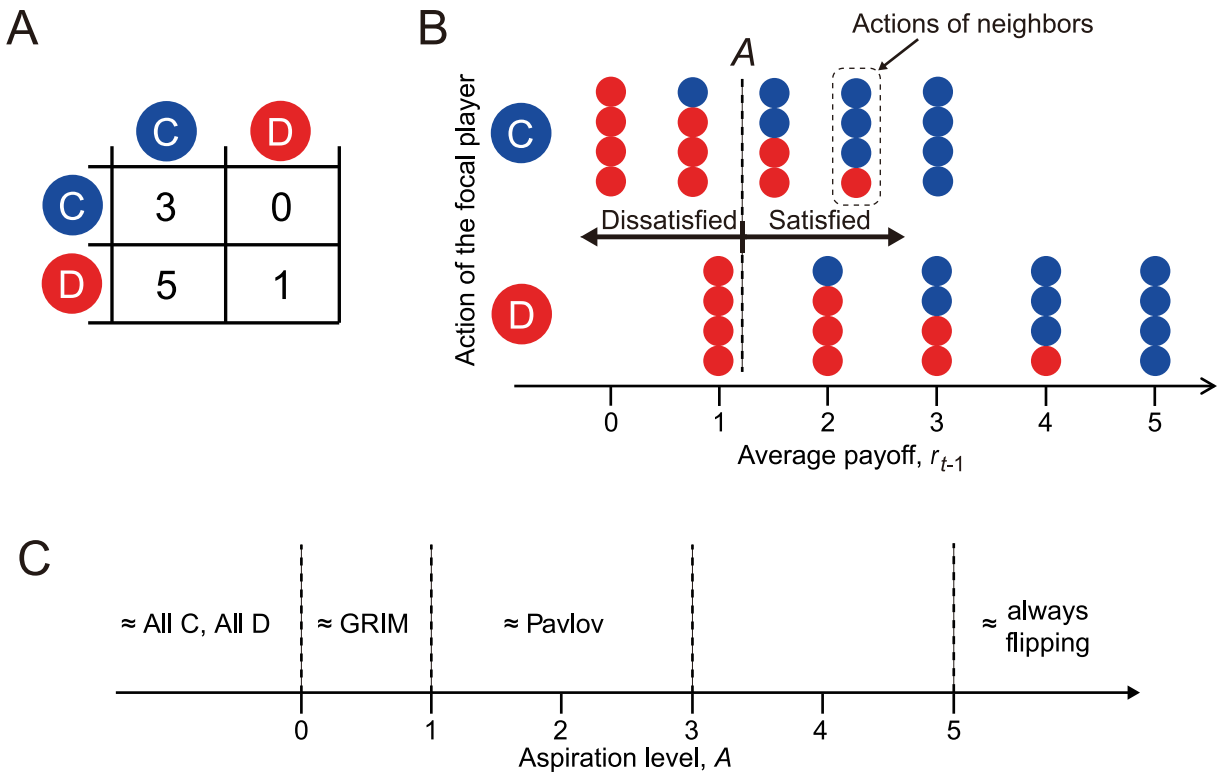

Fig 1. Behavior of the aspiration learner in the repeated PD game. (A) Payoff matrix. The payoff values for the row player are shown. (B) Concept of satisficing in the aspiration-based reinforcement learning model. The payoff values shown on the horizontal axis are those for the focal player. (C) Relationship between the aspiration level, A, and the approximate (un)conditional strategy, given the payoff matrix shown in (A).

doi:10.1371/journal.pcbi.1005034.g001

respectively. For example, the first line on the right-hand side of Eq (1) states that the player increases the probability to cooperate if it has cooperated and been satisfied in the previous round. The multiplicative factor $(1 - p_{t-1})$ is imposed to respect the constraint $p_t < 1$.

The stimulus is defined by

$$s_{t-1} = \tanh[\beta(r_{t-1} - A)], \tag{2}$$

where r_{t-1} is the payoff to the player in round $t - 1$, averaged over the four neighboring players, A is the aspiration level, and $\beta(> 0)$ controls the sensitivity of s_{t-1} to $r_{t-1} - A$ [39]. The player is satisfied and dissatisfied if $r_{t-1} - A > 0$ (i.e., $s_{t-1} > 0$) and $r_{t-1} - A < 0$ (i.e., $s_{t-1} < 0$), respectively (Fig 1B). The so-called Pavlov strategy corresponds to $\beta = \infty$ and $P < A < R$ [4, 5] (Fig 1C). The so-called GRIM strategy, which starts with cooperation and turns into permanent defection (if without noise) once the player is defected [2, 41], corresponds to $\beta = \infty$ and $S < A < R$ [38]. When $\beta < \infty$, which we assume, the behavior realized by the BM model is not an exact conditional strategy such as Pavlov or GRIM, but an approximate one. Unlike some previous studies in which A adaptively changes over time [32, 37–39], we assume that A is fixed.

In each round, each player is assumed to misimplement the decision with probability ϵ [5, 6, 39]. Therefore, the actual probability to cooperate in round t is given by $\tilde{p}_t \equiv p_t(1 - \epsilon) + (1 - p_t)\epsilon$. We set $\epsilon = 0.2$ and the initial probability of cooperation $p_1 = 0.5$ unless otherwise stated.

Results

Prisoner's dilemma game

For $A = 0.5$ and $A = 1.5$, the realized probability of cooperation, $\tilde{p}_t$, averaged over the players and simulations is shown in Fig 2A up to 100 rounds. Due to a relatively large initial probability of cooperation, $p_1 = 0.5$, $\tilde{p}_t$ drops within the first ≈20 rounds and stays at the same level afterwards for both A values. This pattern is roughly consistent with behavioral results obtained in laboratory experiments [13–16, 42].

For a range of the two main parameters, the sensitivity of the stimulus to the reward (i.e., β) and the aspiration level setting the satisfaction threshold for players (i.e., A), $\tilde{p}_t$ averaged over the first 25 rounds is shown in Fig 2B. The figure indicates that cooperation is frequent when β is large, which is consistent with the previous results [39], and when A is less than ≈1. The probability of cooperation is also relatively large when A is larger than ≈2. In this situation, defection leads to an unsatisfactory outcome unless at least two out of the four neighbors cooperate (Fig 1B). Because this does not happen often, a player would frequently switch between defection and cooperation, leading to $\tilde{p}_t \approx 0.4$.

The results shown in Fig 2B were largely unchanged when we varied $t_{\max}$ and ϵ (Fig B in S1 Text).

The probability of cooperation, $\tilde{p}_t$, is plotted against the fraction of cooperating neighbors in the previous round, denoted by f_C, for $A = 0.5$ and two values of β in Fig 3A and 3B. The results not conditioned on the action of the player in the previous round are shown by the circles. The player is more likely to cooperate when more neighbors cooperate, consistent with CC patterns reported in experiments with the PDG on the square lattice [42]. CC is particularly pronounced at a large value of β (Fig 3B as compared to Fig 3A).

The relationship between $\tilde{p}_t$ and f_C conditioned on the last action of the focal player, denoted by a_{t-1}, is shown by the triangles and squares. We observe clear MCC patterns, particularly for a large β. In other words, players that have previously cooperated (i.e., $a_{t-1} = C$) show CC, whereas the probability of cooperation stays constant or mildly decreases as f_C increases

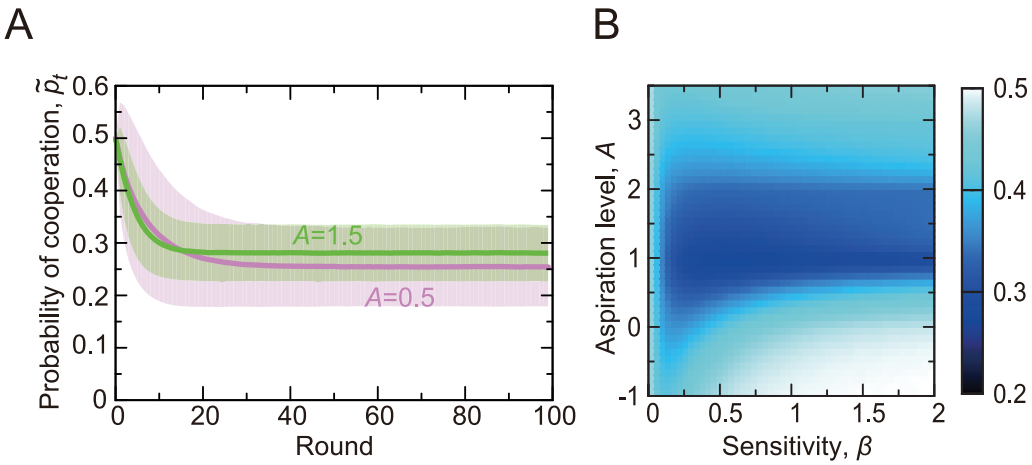

Fig 2. Probability of cooperation in the repeated PDG game on the square lattice having 10 × 10 nodes. (A) Mean time courses of the actual probability of cooperation, $\tilde{p}_t$. The lines represent the actual probability of cooperation averaged over the 10^2 players and 10^3 simulations. We set $\beta = 0.2$ and $A = 0.5$. The shaded regions represent the error bar calculated as one standard deviation. (B) Probability of cooperation for various values of the sensitivity of the stimulus to the reward, β, and the aspiration level, A. The shown values are averages over the 10^2 players, the first $t_{\max} = 25$ rounds, and 10^3 simulations.

doi:10.1371/journal.pcbi.1005034.g002

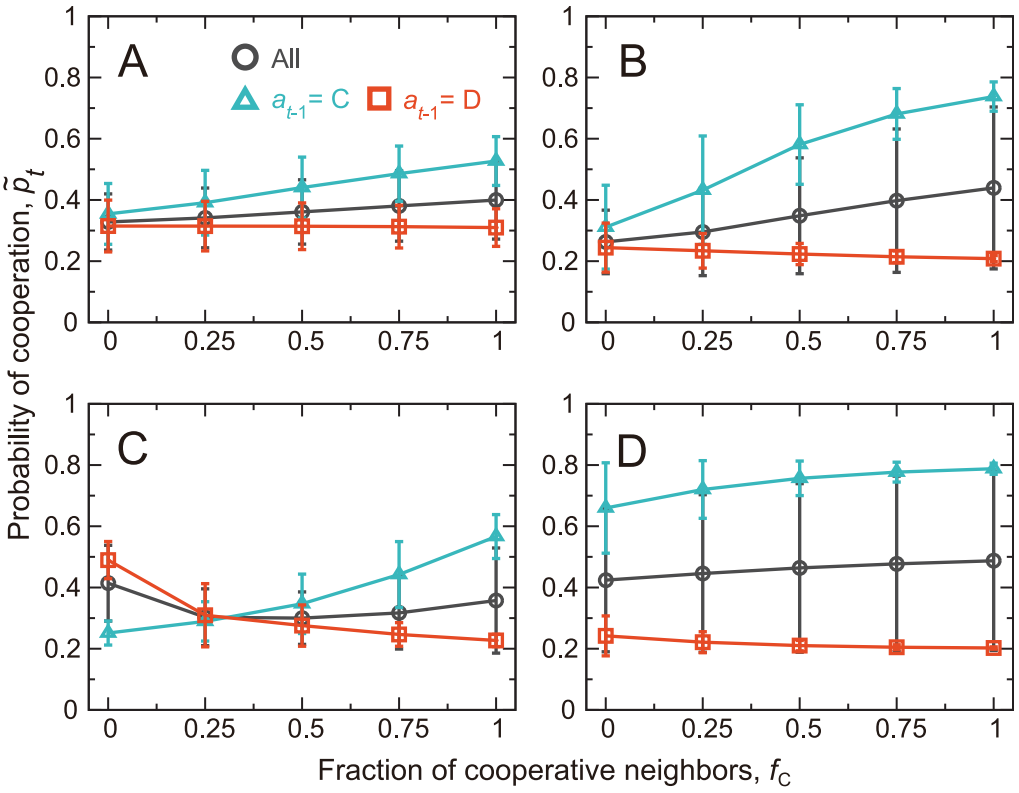

Fig 3. CC and MCC in the repeated PDG on the square lattice. The actual probability of cooperation, $\tilde{p}_t$, is plotted against the fraction of cooperative neighbors in the previous round, f_C. The error bars represent the mean ± standard deviation calculated on the basis of all players, t_{max} = 25 rounds, and 10^3 simulations. The circles represent the results not conditioned on a_{t-1}. The triangles and the squares represent the results conditioned on a_{t-1} = C and a_{t-1} = D, respectively. We set (A) β = 0.1 and A = 0.5, (B) β = 0.4 and A = 0.5, (C) β = 0.4 and A = 2.0, and (D) β = 0.4 and A = −1.0.

doi:10.1371/journal.pcbi.1005034.g003

when the player has previously defected (i.e., a_{t-1} = D). These MCC patterns are consistent with the extant experimental results [13–16].

In the experiments, MCC has also been observed for different population structure such as the scale-free network [16] and a dynamically changing network [14]. We carried out numerical simulations on the regular random graph (i.e., random graph in which all nodes have the same degree, or the number of neighbors) with degree four and the well-mixed group of five players in which each player had four partners. The results remained qualitatively the same as those for the square lattice, suggesting robustness of the present numerical results with respect to the network structure (Fig A in S1 Text). Spatial or network reciprocity is not needed for the present model to show MCC patterns.

A different aspiration level, A, produces different patterns. CC and MCC patterns are lost when we set A = 2 (Fig 3C), with which the dependence of $\tilde{p}_t$ on f_C is small, and $\tilde{p}_t$ when no neighbor has cooperated in the previous round (i.e., f_C = 0) is larger for a_{t-1} = D (squares in Fig 3C) than for a_{t-1} = C (triangles). The latter pattern in particular contradicts the previous behavioral results [13–16]. CC and MCC patterns are mostly lost for A = −1 as well (Fig 3D). With A = −1, the BM player is satisfied by any outcome such that any action is reinforced except for the action implementation error. Therefore, the behavior is insensitive to the reward, or to f_C.

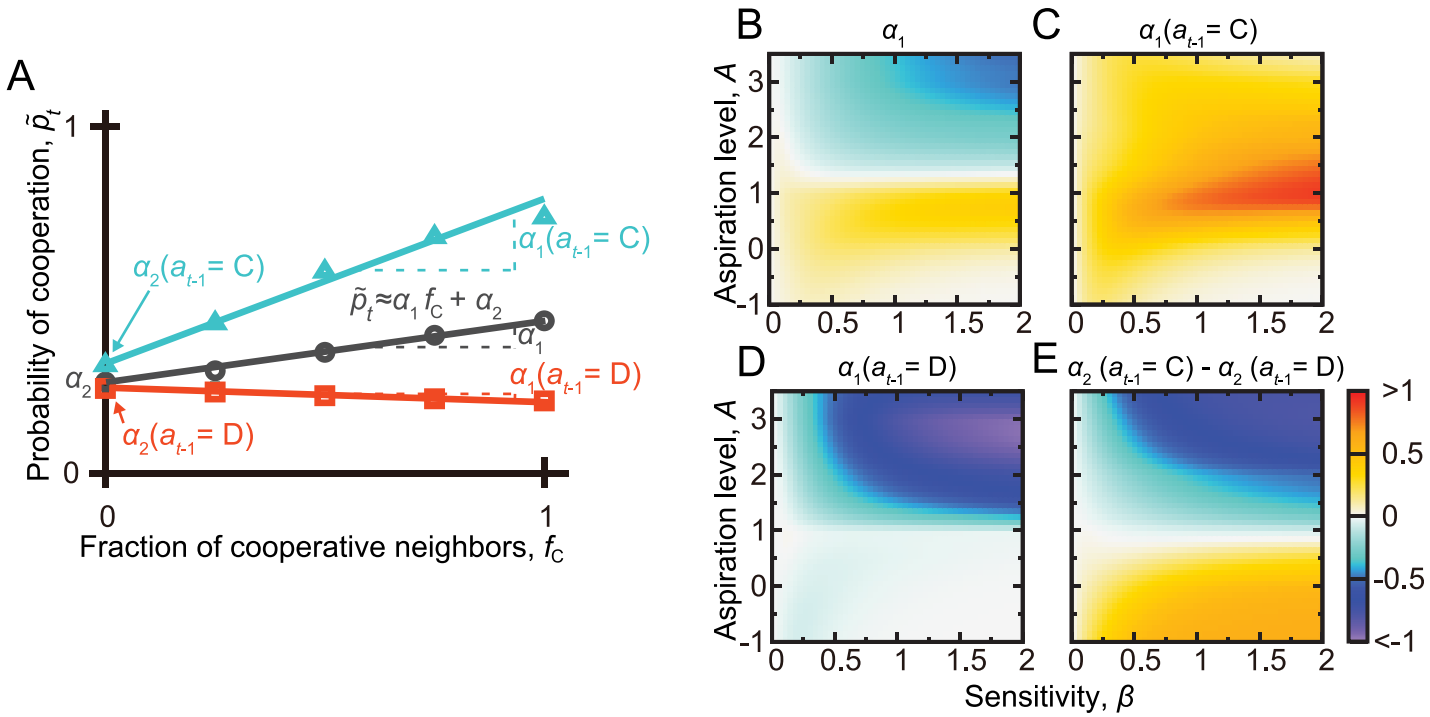

Fig 4. Search of CC and MCC patterns in the repeated PDG on the square lattice. (A) Schematic of the linear fit, $\tilde{p}_t \approx \alpha_1 f_C + \alpha_2$. (B) Slope α_1 of the linear fit when not conditioned on the focal player's previous action, a_{t-1}. (C) α_1 when conditioned on a_{t-1} = C. (D) α_1 when conditioned on a_{t-1} = D. (E) Difference between the intercept, α_2, obtained from the linear fit conditioned on a_{t-1} = C and that conditioned on a_{t-1} = D. For each combination of the β and A values, a linear fit was obtained by the least-squares method on the basis of the 10^2 players, t_{max} = 25 rounds, and 10^3 simulations, yielding 2.5×10^6 samples in total.

doi:10.1371/journal.pcbi.1005034.g004

To assess the robustness of the results, we scanned a region in the $\beta - A$ parameter space. For each combination of β and A values, we performed linear least-square fits to the relationship between the mean $\tilde{p}_t$ and f_C, estimating $\tilde{p}_t \approx \alpha_1 f_C + \alpha_2$ (Fig 4A). CC is supported if the obtained slope α_1 is positive when unconditioned on a_{t-1} (circles in Fig 3). MCC is supported if α_1 is positive when a_{t-1} = C (triangles in Fig 3) and negative or close to zero when a_{t-1} = D (squares in Fig 3). Intercept α_2 is equal to the value of $\tilde{p}_t$ when no neighbor has cooperated in the previous round. The behavioral results suggest that α_2 is larger when conditioned on a_{t-1} = C than on a_{t-1} = D [13–16].

Fig 4B indicates that the slope α_1 unconditioned on a_{t-1} is positive, producing CC, when $A \leq 1$ and β is larger than ≈0.25. However, α_1 is less positive when A is extremely small, i.e., smaller than ≈0. When conditioned on a_{t-1} = C, α_1 is positive, consistent with the MCC patterns, except when β is larger than ≈0.5 and A is smaller than ≈0 (Fig 4C). When conditioned on a_{t-1} = D, α_1 is close to zero when $A \leq 1$ and substantially negative when $A \geq 1$ (Fig 4D). The difference in the value of α_2, the intercept of the linear fit, between the cases a_{t-1} = C and a_{t-1} = D is shown in Fig 4E. The figure indicates that this value is non-negative, consistent with the experimental results, only when $A < 1$. To conclude, CC and MCC patterns consistent with the behavioral results are produced when $0 < A < 1$ and β is not too small. We also confirmed that a different implementation of the BM model [32] produced CC and MCC patterns when $A < 1$ (Fig C in S1 Text).

The BM model with $P < A < R$, i.e., $1 < A < 3$, corresponds to the Pavlov strategy, which is a strong competitor and facilitator of cooperation in the repeated PDG [4, 5]. Our results do not indicate that the Pavlov strategy explains CC and MCC patterns. In fact, the BM model

with $S < A < P$ (i.e., $0 < A < 1$), which is a noisy GRIM-like reinforcement learning, robustly produces CC and MCC patterns. It should be noted that, a noisy GRIM strategy without reinforcement learning components does not produce CC and MCC patterns (Fig D in S1 Text). This result suggests an active role of reinforcement learning rather than merely conditional strategies such as the noisy GRIM.

Public goods game

CC behavior has been commonly observed for humans engaged in the repeated PGG in which participants make a graded amount of contribution [9–11, 43–45]. It should be noted that the player's action is binary in the PDG. In accordance with the setting of previous experiments [10], we consider the following repeated PGG in this section. We assume that four players form a group and repeatedly play the game. In each round, each player receives one monetary unit and determines the amount of contribution to a common pool, denoted by $a_t \in [0, 1]$. The sum of the contribution over the four players is multiplied by 1.6 and equally redistributed to them. Therefore, the payoff to a player is equal to $1 - a_t + 0.4(a_t + \sum_{j=1}^{3} \tilde{a}_{j,t})$, where $\tilde{a}_{j,t}$ is the contribution by the jth other group member in round t. The Nash equilibrium is given by no contribution by anybody, i.e., $a_t = \tilde{a}_{j,t} = 0$ $(1 \leq j \leq 3)$.

We simulated the repeated PGG in which players implemented a variant of the BM model (see Materials and Methods). Crucially, we introduced a threshold contribution value X above which the action was regarded to be cooperative. In other words, an amount of contribution $a_t \geq X$ and $a_t < X$ are defined to be cooperation and defection, respectively. Binarization of the action is necessary for determining the behavior to be reinforced and that to be anti-reinforced.

In Fig 5, the contribution by a player, a_t, averaged over the players, rounds, and simulations is plotted against the average contribution by the other group members, which is again denoted by $f_C (0 \leq f_C \leq 1)$. We observe CC behavior for this parameter set when $X = 0.3$ and 0.4 (circles in Fig 5A and 5B, respectively). CC patterns are weak for $X = 0.5$ (Fig 5C). The average contribution by a player as a function of f_C and the action of the focal player in the previous round is shown by the triangles and squares in Fig 5A–5C. We find MCC patterns. CC and MCC shown in Fig 5 are robustly observed if β is larger than ≈ 0.2, $A \leq 1$, and $0.1 \leq X \leq 0.4$ (Fig 5D–5G and Fig E in S1 Text).

Directional learning is a reinforcement learning rule often applied to behavioral data in the PGG [46, 47] and the PDG [48]. By definition, a directional learner keeps increasing (decreasing) the contribution if an increase (decrease) in the contribution in the previous round has yielded a large reward. In a broad parameter region, we did not find CC or MCC behavior with players obeying the directional learning rule (Fig F in S1 Text). The present BM model is simpler and more accurate in explaining the experimental results in terms of CC and MCC patterns than directional learning is.

Presence of free riders

So far, we have assumed that all players are aspiration learners. Empirically, strategies depend on individuals in the repeated PDG [13, 15, 16] and PGG [9, 10, 18]. In particular, a substantial portion of participants in the repeated PGG, varying between 2.5% and 33% depending on experiments, is free rider, i.e., unconditional defector [9, 43, 49, 50]. Therefore, we performed simulations when BM players and unconditional defectors were mixed. We found that the CC and MCC patterns measured for the learning players did not considerably alter in both PDG and PGG when up to half the players were assumed to be unconditional defectors (Fig G in S1 Text).

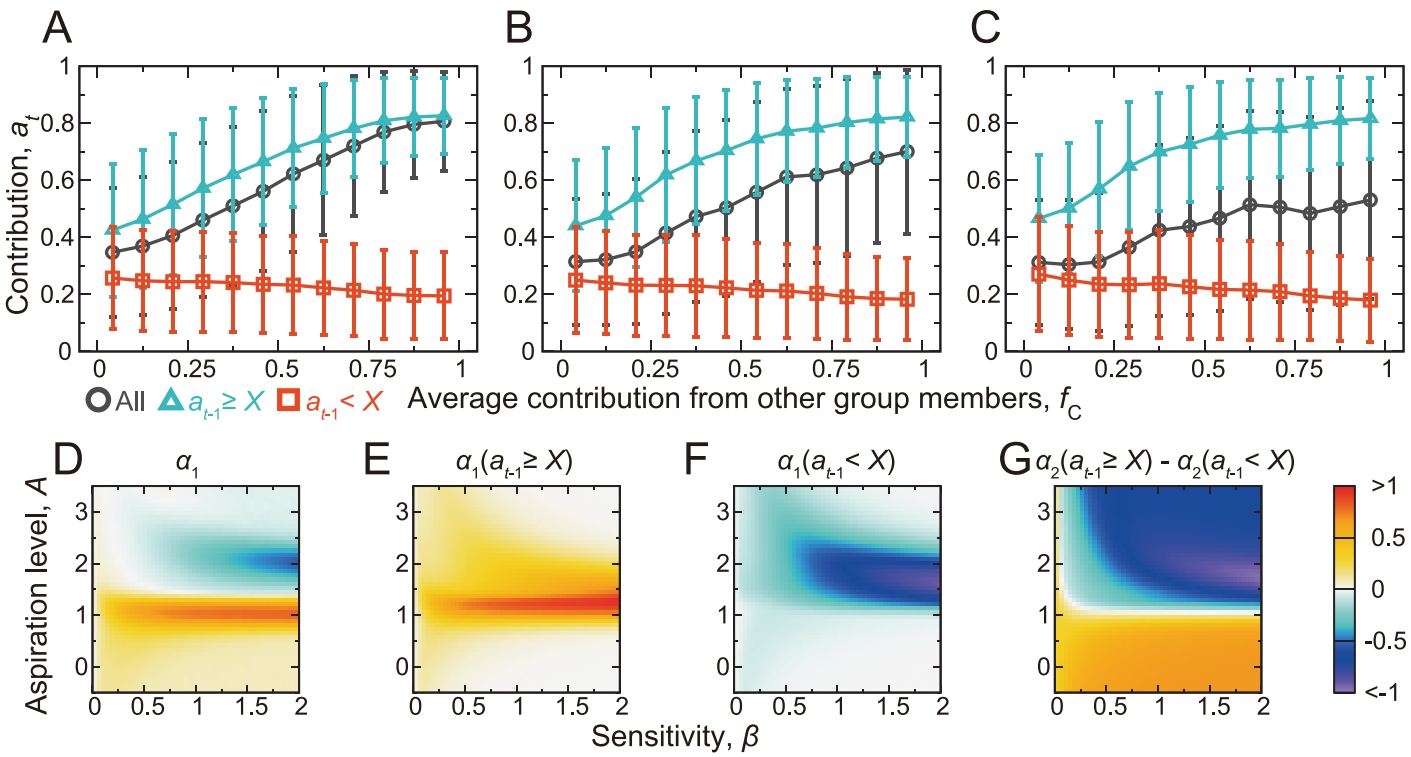

Fig 5. CC and MCC patterns in the repeated PGG in a group of four players. (A)–(C) Contribution by a player (i.e., a_t) conditioned on the average contribution by the other group members in the previous round (i.e., f_C). We set $\beta = 0.4$ and $A = 0.9$. (A) $X = 0.3$, (B) $X = 0.4$, and (B) $X = 0.5$. The circles represent the results not conditioned on a_{t-1}. The triangles and the squares represent the results conditioned on $a_{t-1} \geq X$ and $a_{t-1} < X$, respectively. (D) Slope α_1 of the linear fit, $a_t \approx \alpha_1 f_C + \alpha_2$, when not conditioned on a_{t-1}. (E) α_1 when conditioned on $a_{t-1} \geq X$. (F) α_1 when conditioned on $a_{t-1} < X$. (G) Difference between α_2 obtained from the linear fit conditioned on $a_{t-1} \geq X$ and that conditioned on $a_{t-1} < X$. The mean and standard deviation in (A)–(C) and the linear fit used in (D)–(G) were calculated on the basis of the four players, $t_{max} = 25$ rounds, and 2.5×10^4 simulations, yielding 2.5×10^6 samples in total.

doi:10.1371/journal.pcbi.1005034.g005

Discussion

We have provided compelling numerical evidence that the BM model, a relatively simple aspiration-based reinforcement learning model that has been employed in various decision making tasks [27–29, 31–39], explains CC and MCC patterns. On one hand, aspiration learning has offered a proximate mechanism for cooperation [28, 29, 31, 32, 37–39]. On the other hand, conditional cooperation in the repeated PGG [9–11, 43–45] and its moody variant in the repeated PDG on networks [13–16] have been consistently observed. Here we provided a connection between aspiration learning and conditional cooperation. Our choice of the parameter values including the number of rounds, the size of the group or neighborhood, and the payoff values, supports the comparison of the present numerical data with the results of behavioral experiments.

We are not the first to provide this link. Cimini and Sánchez have shown that MCC emerges from a BM model [25]. The current results significantly depart from theirs and are fundamentally new as follows.

First, MCC is built in into their model in the sense that every outcome except for a population of unconditional defectors implies MCC patterns. In their model, the linear relationship $p_t = \alpha_1 f_C + \alpha_2$ after the focal player's cooperation, where p_t is the probability of cooperation and f_C is the fraction of cooperation in the neighborhood in the previous round, adaptively changes according to the BM model dynamics. In fact, α_1 and α_2 are simultaneously updated

under a constraint and take a common value after a transient (S1 Text), consistent with their numerical results (Fig 2 in [25]). This relationship yields $p_t = \alpha_1(f_C + 1)$, implying MCC whenever $\alpha_1 > 0$. When $\alpha_1 = 0$, we obtain $p_t = 0$, i.e., unconditional defection. In contrast, players in our model directly adapt the unconditional probability of cooperation without knowing f_C such that there is no room for players to explicitly learn the MCC rule. Therefore, our approach is inherently bottom-up.

Second, our model is cognitively less taxing than the Cimini-Sánchez model. In their model, a player refers to f_C and updates the action rule based on its own actions in the last two rounds. Depending on the action that the player has submitted in the second last round, the parameters in one of the two subrules ((p, r) or q in [25]) are updated. In contrast, as already mentioned, players do not refer to f_C in our model. They only refer to their own reward and action in the previous round. A player simply increases or decreases the unconditional probability of cooperation in the next round depending on the amount of satisfaction, as assumed in the previous experimental [28] and theoretical [29, 32, 37–39] studies applying aspiration-based reinforcement learning models to social dilemma games.

In Ref. [25], the Pavlov rather than GRIM rule produced MCC patterns. Our results were the opposite. With Pavlov, CC behavior is lost in our simulations (Figs 4B and 5D). In addition, a Pavlov player cooperates more often after it has defected than cooperated in the last round (Figs 4E and 5G), qualitatively contradicting the experimental results. This inconsistency with Pavlov persists even if we use the Macy-Flache reinforcement learning model as in [25] (Fig C in S1 Text). MCC is intuitively associated with GRIM, not Pavlov, for the following reason. Consider the two-person PDG for simplicity and a player obeying MCC. The player and has obtained payoff R (by mutual cooperation; $f_C = 1$), the player would cooperate in the next round. If the same MCC player has obtained payoff S (by the player's unilateral cooperation; $f_C = 0$), the player would defect in the next round. If the player has obtained payoff P or T (by the player's defection, i.e., $a_{t-1} = \mathrm{D}$), the player would next submit a_t (= C or D) independently of the previously obtained payoff (i.e., P or T). If $a_t = \mathrm{C}$, the player has flipped the action because $a_{t-1} = \mathrm{D}$. This MCC behavior is not realizable by the aspiration learning because it requires $S, P, T < A < R$, which contradicts the payoff of the PDG, i.e., $S < P < R < T$. If $a_t = \mathrm{D}$, the player has not flipped the action. This MCC behavior is realizable by a value of A verifying $S < A < R, P, T$, which is the GRIM.

The GRIM is not exploited by an unconditional defector. In contrast, the Pavlov is exploited by an unconditional defector every other round because Pavlov players flip between cooperation and defection. In experiments, a substantial fraction of participants unconditionally defects [9, 43, 49, 50]. The parameters of the aspiration learning may have evolved such that humans behave like noisy GRIM to protect themselves against exploitation by unconditional defectors. It should be noted that the mere GRIM strategy, corresponding to $\beta = \infty$ and $S < A < P$ in our model, does not produce MCC patterns (Fig D in S1 Text). Therefore, an involvement of reinforcement learning seems to be crucial in explaining the behavioral results, at least within the framework of the present model.

Our numerical results indicated MCC in the PGG. Past laboratory experiments using the PGG focused on CC, not MCC, to the best of our knowledge. As pointed out in previous literature [16], examining the possibility of MCC patterns in the repeated PGG with experimental data warrants future research. Conversely, applying the BM model and examining the relevance of noisy GRIM in the existing and new experimental data may be fruitful exercises.

The results were insensitive to the population structure (Fig A in S1 Text). This is in a stark contrast with a range of results in evolutionary games on networks, which generally say that the population structure is a major determinant of evolutionary game dynamics, in particular, the frequency of cooperation [51–53]. The discrepancy suggests that, under social dilemma

games in laboratory experiments, humans may behave differently from the assumptions of evolutionary dynamics. In fact, regular lattices [54] and scale-free networks [16] do not enhance cooperation in behavioral experiments, which is contrary to the prediction of the evolutionary game theory. In addition, human strategy updating can considerably deviate from those corresponding to major evolutionary rules [42]. Aspiration learning provides an attractive alternative to evolutionary rules in approximating human behavior in social dilemma situations and beyond.

Materials and Methods

BM model for the PGG

Unlike in the PDG, the action is continuous in the PGG such that the behavior to be reinforced or anti-reinforced is not obvious. Therefore, we modify the BM model for the PDG in the following two aspects. First, we define p_t as the expected contribution that the player makes in round t. We draw the actual contribution a_t from the truncated Gaussian distribution whose mean and standard deviation are equal to p_t and 0.2, respectively. If a_t falls outside the interval [0, 1], we discard it and redraw a_t until it falls within [0, 1]. Second, we introduce a threshold contribution value X, distinct from A, used for regarding the action to be either cooperative or defective.

We update p_t as follows:

$$p_t = \begin{cases} p_{t-1} + (1 - p_{t-1})s_{t-1} & (a_{t-1} \geq X \text{ and } s_{t-1} \geq 0), \\ p_{t-1} + p_{t-1}s_{t-1} & (a_{t-1} \geq X \text{ and } s_{t-1} < 0), \\ p_{t-1} - p_{t-1}s_{t-1} & (a_{t-1} < X \text{ and } s_{t-1} \geq 0), \\ p_{t-1} - (1 - p_{t-1})s_{t-1} & (a_{t-1} < X \text{ and } s_{t-1} < 0). \end{cases} \quad (3)$$

For example, the first line on the right-hand side of Eq (3) states that, if the player has made a large contribution (hence regarded to be C) and it has been rewarding, the player will increase the expected contribution in the next round. The stimulus, s_{t-1}, is defined by Eq (2).

In the numerical simulations, we draw the initial condition, p_1, from the uniform density on [0, 1], independently for different players.

Supporting Information

S1 Text. Supporting Information for: Reinforcement Learning Explains Conditional Cooperation and Its Moody Cousin.
(PDF)

Acknowledgments

We acknowledge Hisashi Ohtsuki and Shinsuke Suzuki for valuable comments on the manuscript.

Author Contributions

Conceived and designed the experiments: NM. Performed the experiments: TE. Analyzed the data: TE. Wrote the paper: TE YH MT NM.

References

1. Trivers RL. The evolution of reciprocal altruism. Q Rev Biol. 1971; 46:35–57. doi: 10.1086/406755

2. Axelrod R. The Evolution of Cooperation. New York: Basic Books; 1984.
3. Nowak MA, May RM. Evolutionary games and spatial chaos. Nature. 1992; 359:826–829. doi: 10.1038/359826a0
4. Kraines D, Kraines V. Learning to cooperate with Pavlov: An adaptive strategy for the iterated prisoner's dilemma with noise. Theory Decis. 1993; 35:107–150. doi: 10.1007/BF01074955
5. Nowak MA, Sigmund K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature. 1993; 364:56–58. doi: 10.1038/364056a0 PMID: 8316296
6. Nowak MA, Sigmund K, El-Sedy E. Automata, repeated games and noise. J Math Biol. 1995; 33:703–722. doi: 10.1007/BF00184645
7. Boyd R, Richerson PJ. The evolution of reciprocity in sizable groups. J Theor Biol. 1988; 132:337–356. doi: 10.1016/S0022-5193(88)80219-4 PMID: 3226132
8. Joshi NV. Evolution of cooperation by reciprocation within structured demes. J Genet. 1987; 66:69–84. doi: 10.1007/BF02934456
9. Keser C, van Winden F. Conditional cooperation and voluntary contributions to public goods. Scand J Econ. 2000; 102:23–39. doi: 10.1111/1467-9442.00182
10. Fischbacher U, Gächter S, Fehr E. Are people conditionally cooperative? Evidence from a public goods experiment. Econ Lett. 2001; 71:397–404. doi: 10.1016/S0165-1765(01)00394-9
11. Fehr E, Fischbacher U. Social norms and human cooperation. Trends Cogn Sci. 2004; 8:185–190. doi: 10.1016/j.tics.2004.02.007 PMID: 15050515
12. Gächter S. Conditional cooperation: Behavioral regularities from the lab and the field and their policy implications. In: Frey BS, Stutzer A, editors. Economics and Psychology: A Promising New Cross-disciplinary Field. Cambridge: MIT Press; 2007. pp. 19–50.
13. Grujić J, Fosco C, Araujo L, Cuesta JA, Sánchez A. Social experiments in the mesoscale: Humans playing a spatial prisoner's dilemma. PLOS ONE. 2010; 5:e13749. doi: 10.1371/journal.pone.0013749 PMID: 21103058
14. Grujić J, Röhl T, Semmann D, Milinski M, Traulsen A. Consistent strategy updating in spatial and non-spatial behavioral experiments does not promote cooperation in social networks. PLOS ONE. 2012; 7: e47718. doi: 10.1371/journal.pone.0047718 PMID: 23185242
15. Grujić J, et al. A comparative analysis of spatial Prisoner's Dilemma experiments: Conditional cooperation and payoff irrelevance. Sci Rep. 2014; 4:4615. doi: 10.1038/srep04615 PMID: 24722557
16. Gracia-Lázaro C, et al. Heterogeneous networks do not promote cooperation when humans play a Prisoner's Dilemma. Proc Natl Acad Sci USA. 2012; 109:12922–12926. doi: 10.1073/pnas.1206681109 PMID: 22773811
17. Fowler JH, Christakis NA. Cooperative behavior cascades in human social networks. Proc Natl Acad Sci USA. 2010; 107:5334–5338. doi: 10.1073/pnas.0913149107 PMID: 20212120
18. Rustagi D, Engel S, Kosfeld M. Conditional cooperation and costly monitoring explain success in forest commons management. Science. 2010; 330:961–965. doi: 10.1126/science.1193649 PMID: 21071668
19. Wahl LM, Nowak MA. The continuous prisoner's dilemma: I. Linear reactive strategies. J Theor Biol. 1999; 200:307–321. doi: 10.1006/jtbi.1999.0996 PMID: 10527720
20. Doebeli M, Hauert C. Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. Ecol Lett. 2005; 8:748–766. doi: 10.1111/j.1461-0248.2005.00773.x
21. André JB, Day T. Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. J Theor Biol. 2007; 247:11–22. doi: 10.1016/j.jtbi.2007.02.007 PMID: 17397874
22. Le S, Boyd R. Evolutionary dynamics of the continuous iterated Prisoner's Dilemma. J Theor Biol. 2007; 245:258–267. doi: 10.1016/j.jtbi.2006.09.016 PMID: 17125798
23. Takezawa M, Price ME. Revisiting "The revolution of reciprocity in sizable groups": Continuous reciprocity in the repeated *n*-person prisoner's dilemma. J Theor Biol. 2010; 264:188–196. doi: 10.1016/j.jtbi.2010.01.028 PMID: 20144622
24. Guttman JM. On the evolution of conditional cooperation. Eur J Polit Econ. 2013; 30:15–34. doi: 10.1016/j.ejpoleco.2012.11.003
25. Cimini G, Sánchez A. Learning dynamics explains human behaviour in Prisoner's Dilemma on networks. J R Soc Interface. 2014; 11:20131186. doi: 10.1098/rsif.2013.1186 PMID: 24554577
26. Gracia-Lázaro C, Cuesta JA, Sánchez A, Moreno Y. Human behavior in Prisoner's Dilemma experiments suppresses network reciprocity. Sci Rep. 2012; 2:325. doi: 10.1038/srep00325 PMID: 22439103
27. Bush RR, Mosteller F. Stochastic Models for Learning. New York: Wiley; 1955.

28. Rapoport A, Chammab AM. Prisoner's Dilemma: A Study in Conflict and Cooperation. Ann Arbor: University of Michigan Press; 1965.
29. Macy MW. Learning to cooperate: Stochastic and tacit collusion in social exchange. Am J Sociol. 1991; 97:808–843. doi: 10.1086/229821
30. Fudenberg D, Levine DK. The Theory of Learning in Games. Cambridge: MIT Press; 1998.
31. Bendor J, Mookherjee D, Ray D. Aspiration-based reinforcement learning in repeated interaction games: An overview. Int Game Theory Rev. 2001; 3:159–174. doi: 10.1142/S0219198901000348
32. Macy MW, Flache A. Learning dynamics in social dilemmas. Proc Natl Acad Sci USA. 2002; 99:7229–7236. doi: 10.1073/pnas.092080099 PMID: 12011402
33. Bendor J, Diermeier D, Ting M. A behavioral model of turnout. Am Polit Sci Rev. 2003; 97:261–280. doi: 10.1017/S0003055403000662
34. Duffy J. Agent-based models and human subject experiments. In: Tesfatsion L, Judd KL, editors. Handbook of Computational Economics. Amsterdam: North-Holland; 2006. pp. 949–1011.
35. Fowler JH. Habitual voting and behavioral turnout. J Polit. 2006; 68:335–344. doi: 10.1111/j.1468-2508.2006.00410.x
36. Rische JL, Komarova NL. Regularization of languages by adults and children: A mathematical framework. Cogn Psychol. 2016; 84:1–30. doi: 10.1016/j.cogpsych.2015.10.001 PMID: 26580218
37. Karandikar R, Mookherjee D, Ray D, Vega-Redondo F. Evolving aspirations and cooperation. J Econ Theory. 1998; 80:292–331. doi: 10.1006/jeth.1997.2379
38. Posch M, Pichler A, Sigmund K. The efficiency of adapting aspiration levels. Proc R Soc B. 1999; 266:1427–1435. doi: 10.1098/rspb.1999.0797
39. Masuda N, Nakamura M. Numerical analysis of a reinforcement learning model with the dynamic aspiration level in the iterated Prisoner's dilemma. J Theor Biol. 2011; 278:55–62. doi: 10.1016/j.jtbi.2011.03.005 PMID: 21397610
40. Pacheco JM, Santos FC, Souza MO, Skyrms B. Evolutionary dynamics of collective action in *N*-person stag hunt dilemmas. Proc R Soc B. 2009; 276:315–321. doi: 10.1098/rspb.2008.1126 PMID: 18812288
41. Friedman JW. Non-cooperative equilibrium for supergames. Rev Econ Stud. 1971; 38:1–12. doi: 10.2307/2296617
42. Traulsen A, Semmann D, Sommerfeld RD, Krambeck HJ, Milinski M. Human strategy updating in evolutionary games. Proc Natl Acad Sci USA. 2010; 107:2962–2966. doi: 10.1073/pnas.0912515107 PMID: 20142470
43. Kurzban R, Houser D. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. Proc Natl Acad Sci USA. 2005; 102:1803–1807. doi: 10.1073/pnas.0408759102 PMID: 15665099
44. Herrmann B, Thöni C. Measuring conditional cooperation: A replication study in Russia. Exp Econ. 2009; 12:87–92. doi: 10.1007/s10683-008-9197-1
45. Chaudhuri A. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. Exp Econ. 2011; 14:47–83. doi: 10.1007/s10683-010-9257-1
46. Burton-Chellew MN, Nax HH, West SA. Payoff-based learning explains the decline in cooperation in public goods games. Proc R Soc B. 2015; 282:20142678. doi: 10.1098/rspb.2014.2678 PMID: 25589609
47. Nax HH, Perc M. Directional learning and the provisioning of public goods. Sci Rep. 2015; 5:8010. doi: 10.1038/srep08010 PMID: 25619192
48. Selten R, Stoecker R. End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach. J Econ Behav Organ. 1986; 7:47–70. doi: 10.1016/0167-2681(86)90021-1
49. Kurzban R, Houser D. Individual differences in cooperation in a circular public goods game. Eur J Pers. 2001; 15:37–52. doi: 10.1002/per.420
50. Fischbacher U, Gächter S. Social preference, beliefs and the dynamics of free riding in public goods experiments. Am Econ Rev. 2010; 100:541–556. doi: 10.1257/aer.100.1.541
51. Nowak MA. Evolutionary Dynamics. Cambridge: Harvard University Press; 2006.
52. Szabó G, Fáth G. Evolutionary games on graphs. Phys Rep. 2007; 446:97–216. doi: 10.1016/j.physrep.2007.04.004
53. Perc M, Gómez-Gardeñes J, Szolnoki A, Floría LM, Moreno Y. Evolutionary dynamics of group interactions on structured populations: A review. J R Soc Interface. 2013; 10:20120997. doi: 10.1098/rsif.2012.0997 PMID: 23303223
54. Kirchkamp O, Nagel R. Naive learning and cooperation in network experiments. Games Econ Behav. 2007; 58:269–292. doi: 10.1016/j.geb.2006.04.002