

Electronic supplementary material

Evolution of reciprocity with limited payoff memory

Nikoleta E. Glynatsi, Alex McAvoy, Christian Hilbe

This document provides further details on our methods and derivations, and it contains additional simulation results. Section 1 summarizes the model. In particular, we provide further details on our implementation of the evolutionary dynamics, and our use of the rare-mutation limit. In Section 2, we derive analytical results for the various settings we consider. These settings differ in what kind of payoff information individuals take into account when updating their strategies. In the perfect-memory setting, individuals take into account all their interactions against all co-players. In the limited-memory setting, they only consider the very last round of their very last interaction. In addition, we describe several model extensions in which the amount of information taken into account is in between these two extremes. Finally, Section 3 presents further simulation results. In particular, we confirm that our main results continue to hold (i) when mutations are no longer rare, and (ii) when players use memory-one strategies instead of reactive strategies.

1 Description of the model

Summary of the model. As described in the main text, we study cooperative behavior in a population of size N , with N being even. The dynamics unfold on two time scales. The short time scale describes the game dynamics. Here the N individuals are randomly matched to form $N/2$ pairs to interact in a repeated prisoner's dilemma. Each round, individuals can choose whether to cooperate (C) or defect (D). In the most general setting, the resulting one-shot payoffs can be summarized by the payoff matrix

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \left(\begin{array}{cc} R & S \\ T & P \end{array} \right). \end{array} \quad (1)$$

Here, R is interpreted as the reward for mutual cooperation, S is the sucker's payoff, T is the temptation, and P is the punishment payoff [1]. Throughout this work, we parametrize these payoffs as $R = b - c$, $S = -c$, $T = b$, and $P = 0$, where b and c are the benefit and cost of cooperation, respectively, with $b > c > 0$. After each round, players learn their co-player's previous action. Then the game continues for another round with

probability δ . Players make their decisions whether to cooperate in any given round based on their reactive strategies $\mathbf{s} = (y, p, q)$. The entry y determines a player's first-round cooperation probability. The other entries p and q determine the player's cooperation probability in all subsequent rounds. The probability is p if the co-player cooperated in the previous round, and it is q if the co-player defected. On this short time scale, the players' strategies are fixed, and players are consecutively matched to play many repeated games with randomly changing interaction partners.

The long time scale describes the evolutionary dynamics. Here, players are allowed to update their strategies based on the payoffs they yield. We model these strategy updates with a pairwise comparison process [2]. This process assumes that at regular time intervals, one player is randomly selected from the population. We refer to this player as the 'learner' (L). The learner is then given an opportunity to update its strategy. There are two possibilities for how this update may occur. With probability μ , the player's strategy mutates randomly. In that case, the player's new strategy is drawn uniformly from the space of all reactive strategies $[0, 1]^3$. With probability $1 - \mu$, the player compares itself with a neighbor. In that case, the player randomly picks another individual from the population (referred to as the 'role model', RM). The learner adopts the role model's strategy with a probability ρ , given by

$$\rho(\pi_L, \pi_{RM}) = \frac{1}{1 + e^{-\beta(\pi_{RM} - \pi_L)}}. \quad (2)$$

The parameter β is the selection strength. It determines how important payoff differences are for the learner's decision to imitate the role model. The variables π_{RM} and π_L refer to the relevant payoffs of the role model and the learner, respectively. The exact value of these payoffs depend on the players' memory. We say players have *perfect memory* when π_{RM} and π_L are given by the players' expected payoffs (across all rounds and across all possible co-players). We say players have *limited memory* when π_{RM} and π_L are given by the players' realized payoff in the very last round of the game with their very last interaction partner. In addition, we consider several model extensions in which individuals have memory capacities in between these two extremes. We provide a detailed description of these different settings and the resulting payoffs in Section 2.

Evolutionary simulations for the rare-mutation limit. To simulate the evolutionary dynamics of the pairwise comparison process, it is sometimes useful to assume that mutations are rare, $\mu \rightarrow 0$. In that case, whenever a mutant strategy appears, it either fixes in the population or goes extinct before the next mutant appears. As a result, at any given time there are at most two different strategies present in the population [3–5]. This assumptions makes computations more efficient, and it makes some of the results easier to interpret. In the following, we describe our implementation of the process in the rare-mutation limit in more detail.

Initially, the process starts with a population where all members use the same strategy (referred to as the resident strategy, R). Then one individual adopts a mutant strategy selected uniformly at random from the set

Algorithm 1: Evolutionary process in the limit of rare mutations

```
 $N \leftarrow$  population size;  
resident  $\leftarrow$  starting resident;  
while  $t < \text{maximum number of steps}$  do  
    mutant  $\leftarrow$  random strategy;  
    fixation probability  $\leftarrow \varphi_M$ ;  
    if  $\varphi_M > \text{random}: i \rightarrow [0, 1]$  then  
        | resident  $\leftarrow$  mutant;  
    end  
end
```

of feasible strategies. The fixation probability φ_M of the mutant strategy can be calculated explicitly [6],

$$\varphi_M = \frac{1}{1 + \sum_{i=1}^{N-1} \prod_{k=1}^i \frac{\lambda_k^-}{\lambda_k^+}}. \quad (3)$$

Here, the index k corresponds to the current number of players with the mutant strategy (mutants). The variables λ_k^-, λ_k^+ are the probabilities that the number of mutants decreases or increases within a single updating step. These probabilities depend on the probability that a mutant and a resident are chosen as the learner and the resident, respectively. In addition, they depend on the respective switching probability ρ , as described by Eq. (2). We specify the exact values of λ_k^-, λ_k^+ for each memory-setting in the next section.

Depending on the fixation probability φ_M , the mutant strategy either fixes (becomes the new resident) or goes extinct. Afterwards, another random mutant strategy is introduced into the population. We iterate this elementary population updating process for a large number of mutant strategies. At each step, we record the current resident strategy and the resulting average cooperation rate.

We consider this limit of rare-mutation throughout the main text. The respective process is summarised by Algorithm 1. In Section 3, we present additional simulation results to show that our qualitative results continue to hold when the mutation rate is strictly bounded away from zero.

2 Analytical results

In the following, we discuss our different memory settings in more detail. We discuss six cases explicitly. In these cases, updating occurs (i) based on average payoffs based on all interactions (perfect memory), (ii) based on the last round of one interaction (limited memory), (iii) based on the last round of two interactions, (iv) based on the last two rounds of one interaction, (v) based on the last two rounds of two interactions, and (vi) based on the expected payoff of one interaction. In each case, we consider the case that there are only two strategies present in the population (a resident and a mutant strategy). We first derive how likely it is that a learner (of any type) assigns a given payoff π_L to itself, and a payoff of π_{RM} to the role model. This allows us to derive explicit expressions for λ_k^-/λ_k^+ , and hence for the mutant's fixation probability according to Eq. (3). Based on these expressions we can characterize under which conditions cooperation is stochastically stable.

2.1 Perfect payoff memory

Computing the ratio λ_k^-/λ_k^+ . The case of perfect payoff memory corresponds to the classical case considered in the previous literature. Here, individuals update their strategies based on the expected payoffs, based on all rounds and all possible interaction partners. When players use reactive strategies (or more generally, strategies with finite memory), these expected payoffs can be computed explicitly, based on a Markov chain approach [7]. To this end, consider two players with strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$, respectively. In each round t of the game, player 1 may get one of the four possible payoffs R , S , T , or P , as described by the general payoff matrix (1). Let $\mathbf{v}(t) = (v_R(t), v_S(t), v_T(t), v_P(t))$ denote the respective probability distribution of observing one of these four outcomes. This probability distribution can be computed recursively. Using the shortcut notation $\bar{z} = 1 - z$ for any $z \in [0, 1]$, we get for the initial round

$$\mathbf{v}_0 := \mathbf{v}(0) = (y_1 y_2, y_1 \bar{y}_2, \bar{y}_1 y_2, \bar{y}_1 \bar{y}_2). \quad (4)$$

Given $\mathbf{v}(t)$, we can compute $\mathbf{v}(t+1)$ as

$$\mathbf{v}(t+1) = \mathbf{v}(t) \cdot M, \quad (5)$$

where M is the transition matrix of the process,

$$M = \begin{bmatrix} p_1 p_2 & p_1 \bar{p}_2 & \bar{p}_1 p_2 & \bar{p}_1 \bar{p}_2 \\ q_1 p_2 & q_1 \bar{p}_2 & \bar{q}_1 p_2 & \bar{q}_1 \bar{p}_2 \\ p_1 q_2 & p_1 \bar{q}_2 & \bar{p}_1 q_2 & \bar{p}_1 \bar{q}_2 \\ q_1 q_2 & q_1 \bar{q}_2 & \bar{q}_1 q_2 & \bar{q}_1 \bar{q}_2 \end{bmatrix} \quad (6)$$

Based on this recursion, we can compute how often player 1 receives one of the four payoffs R, S, T, P on average (across all possible realizations of games among the two players). This average distribution \mathbf{v} is

$$\mathbf{v} := (1-\delta) \sum_{t=0}^{\infty} \delta^t \mathbf{v}(t) = (1-\delta) \mathbf{v}_0 \sum_{t=0}^{\infty} \delta^t M^t = (1-\delta) \mathbf{v}_0 (I_4 - \delta M)^{-1}, \quad (7)$$

where I_4 is the 4×4 identity matrix. Based on this general formula, the four entries of $\mathbf{v} = (v_R, v_S, v_T, v_P)$ can be computed explicitly. Using the auxiliary notation $r_i := p_i - q_i$, we obtain

$$\begin{aligned} v_R &= (1-\delta) \frac{y_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left(q_1 + r_1 ((1-\delta)y_2 + \delta q_2) \right) \left(q_2 + r_2 ((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\ v_S &= (1-\delta) \frac{y_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left(q_1 + r_1 ((1-\delta)y_2 + \delta q_2) \right) \left(\bar{q}_2 - r_2 ((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\ v_T &= (1-\delta) \frac{\bar{y}_1 y_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left(\bar{q}_1 - r_1 ((1-\delta)y_2 + \delta p_2) \right) \left(q_2 + r_2 ((1-\delta)y_1 + \delta q_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}, \\ v_P &= (1-\delta) \frac{\bar{y}_1 \bar{y}_2}{1-\delta^2 r_1 r_2} + \delta \frac{\left(\bar{q}_1 - r_1 ((1-\delta)y_2 + \delta p_2) \right) \left(\bar{q}_2 - r_2 ((1-\delta)y_1 + \delta p_1) \right)}{(1-\delta r_1 r_2)(1-\delta^2 r_1 r_2)}. \end{aligned} \quad (8)$$

Using this distribution \mathbf{v} , we compute the first player's expected payoff as the weighted average

$$\pi(s_1, s_2) = R v_R + S v_S + T v_T + P v_P. \quad (9)$$

After these preparations, consider now a population with k mutants and $N - k$ residents, whose strategies we denote by $s_M = (y_M, p_M, q_M)$ and $s_R = (y_R, p_R, q_R)$, respectively. Assuming that population members are matched randomly (or equivalently, that they interact with all other population members), the resulting expected payoffs of residents and mutants are

$$\begin{aligned} \pi_R(k) &= \frac{N-k-1}{N-1} \cdot \pi(s_R, s_R) + \frac{k}{N-1} \cdot \pi(s_R, s_M), \\ \pi_M(k) &= \frac{N-k}{N-1} \cdot \pi(s_M, s_R) + \frac{k-1}{N-1} \cdot \pi(s_M, s_M). \end{aligned} \quad (10)$$

The number of mutants in the population decreases in a single time step if a mutant is chosen to be the learner and adopts the strategy of a resident. Similarly, it increases if a resident is the learner and adopts the strategy

of a mutant. The respective transition probabilities are

$$\lambda_k^- = \frac{N-k}{N} \frac{k}{N-1} \rho(\pi_M(k), \pi_R(k)) \quad \text{and} \quad \lambda_k^+ = \frac{k}{N} \frac{N-k}{N-1} \rho(\pi_R(k), \pi_M(k)).$$

For ρ as defined by Eq. (2), the ratio of these two transition probabilities simplifies to

$$\frac{\lambda_k^-}{\lambda_k^+} = \frac{\rho(\pi_M(k), \pi_R(k))}{\rho(\pi_R(k), \pi_M(k))} = e^{-\beta(\pi_M(k) - \pi_R(k))}. \quad (11)$$

Based on these ratios for each k , we can compute the mutant's fixation probability by Eq. (3).

Stochastic stability of cooperation. As an application of this formalism, we can compute when cooperation is stochastically stable in the perfect-information setting. To this end, suppose there is only a single mutant, $k = 1$. The residents adopt Generous Tit-for-Tat, $\text{GTFT} = (1, 1, q)$ and the mutant adopts ALLD $= (0, 0, 0)$. When two GTFT players interact, the resulting average distribution according to Eq. (8) simplifies to

$$\mathbf{v}(\text{GTFT}, \text{GTFT}) = (1, 0, 0, 0).$$

On the other hand, if ALLD interacts with GTFT, the respective probabilities become,

$$\mathbf{v}(\text{ALLD}, \text{GTFT}) = (0, 0, 1 - \delta + \delta q, \delta(1 - q)).$$

Based on Eq. (10), we can compute the strategies' expected payoffs as

$$\pi_{\text{GTFT}} = \frac{N-2}{N-1}(b-c) - \frac{1}{N-1}(1-\delta+\delta q)c \quad \text{and} \quad \pi_{\text{ALLD}} = (1-\delta+\delta q)b.$$

As a consequence, we can calculate the corresponding ratio of transition probabilities according to Eq. (11),

$$\frac{\lambda_1^-}{\lambda_1^+} = e^{-\beta((1-\delta+\delta q)(b+\frac{c}{N-1}) - \frac{N-2}{N-1}(b-c))}$$

By definition, cooperation is stochastically stable if this ratio exceeds one, which is equivalent to

$$q < 1 - \frac{1}{\delta} \cdot \frac{b + (N-1)c}{(N-1)b + c}. \quad (12)$$

For such a strategy to be feasible we require $q > 0$, which implies $\delta > (b + (N-1)c)/((N-1)b + c)$. In particular, in the limit of large populations $N \rightarrow \infty$, we obtain that cooperation is stochastically stable if $q < 1 - c/(\delta b)$. The minimum continuation probability for such a strategy to exist is $\delta > c/b$. In this way, we recover the classical conditions for cooperation to be feasible under direct reciprocity [8–10].

2.2 Limited payoff memory

Computing the distribution of last-round payoffs. The case of perfect payoff memory is straightforward to handle; here, every player gets the expected payoff with certainty. In comparison, computing transition probabilities for the case of limited payoff memory is more elaborate. Here, we need to consider the different possible outcomes that both the learner and role model may have experienced in their very last interaction. A further complication arises when the learner's last interaction partner happens to be the role model. In that case, the learner's and the role model's last payoff will be correlated (e.g., if the learner got the sucker's payoff of S , the role model's payoff is T with certainty). To treat the case of limited memory analytically, let $\mathcal{U} = \{R, S, T, P\}$ be the set of possible one-shot payoffs. Since the game ends after round T with probability $\delta^T (1 - \delta)$, if two players use reactive strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$, then, by definition, the probability that the first player receives a payoff of $u \in \mathcal{U}$ in the final round of the game is given by Eq. (8). To make the strategies explicit, we denote this probability by $v_u(s_1, s_2)$.

Computing the ratio $\lambda_k^- / \lambda_k^+$. After these preparations, let us again consider the corresponding population setup, with $N - k$ residents with strategy $s_R = (y_R, p_R, q_R)$ and k mutants with strategy $s_M = (y_M, p_M, q_M)$. At each step of the evolutionary process we choose a learner and a role model. The learner compares the performance of its strategy by comparing its own last one-shot payoff with the last one-shot payoff of the role model. In the following, we assume either the learner or the role model is a resident and that the other player is a mutant (otherwise it is certain that the number of mutants does not change). There are two major cases to consider.

1. The learner and the role model have their last respective interaction with each other. This happens with probability $1/(N - 1)$. In that case, there are four possible cases for their joint final payoffs, $(u_R, u_M) \in \mathcal{U}_F := \{(R, R), (S, T), (T, S), (P, P)\}$. These outcomes follow the distribution $\mathbf{v}(s_R, s_M)$ (Eq. (8)).
2. The learner's last interaction was not with the role model, with probability $(N - 2)/(N - 1)$. In this case, there are four different subcases, depending on whether the resident's last interaction partner was a mutant or a resident, and depending on whether the mutant's last interaction partner was a mutant or a resident. As a result, the resident's last one-shot payoff is distributed according to $\mathbf{v}(s_R, s_M)$ or $\mathbf{v}(s_R, s_R)$; the mutant's last payoff is distributed according to $\mathbf{v}(s_M, s_M)$ or $\mathbf{v}(s_M, s_R)$, respectively.

Let $x(u_R, u_M)$ denote the probability that the resident and the mutant received the payoff u_R and u_M in their

respective last interaction. By taking into account the above two cases, we can compute this probability as

$$\begin{aligned}
x(u_R, u_M) &= \frac{1}{N-1} \cdot v_{u_R}(s_R, s_M) \cdot 1_{(u_R, u_M) \in \mathcal{U}_F} \\
&+ \frac{N-2}{N-1} \cdot \left[\frac{k-1}{N-2} \frac{k-2}{N-3} \cdot v_{u_R}(s_R, s_M) v_{u_M}(s_M, s_M) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} \cdot v_{u_R}(s_R, s_M) v_{u_M}(s_M, s_R) \right. \\
&\quad \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} \cdot v_{u_R}(s_R, s_R) v_{u_M}(s_M, s_M) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} \cdot v_{u_R}(s_R, s_R) v_{u_M}(s_M, s_R) \right].
\end{aligned} \tag{13}$$

The first term on the right side corresponds to the case that the learner and the role model happened to be matched directly for their last interaction. In that case, only those payoff pairs can occur that are feasible in a direct interaction. That is, it needs to be the case that $(u_R, u_M) \in \mathcal{U}_F$, as represented by the respective indicator function. The probability that the number of mutants increase or decreases by one is now given by

$$\begin{aligned}
\lambda_k^+ &= \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_R, u_M \in \mathcal{U}} x(u_R, u_M) \cdot \rho(u_R, u_M), \\
\lambda_k^- &= \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_R, u_M \in \mathcal{U}} x(u_R, u_M) \cdot \rho(u_M, u_R).
\end{aligned} \tag{14}$$

In this expression, the prefactor $(N-k)k/N^2$ gives the probability that the two players (learner and role model) have different strategies. The sum corresponds to the total probability that the learner adopts the role model's strategy, by summing up over all possible payoffs u_R and u_M that the two players may have received in their respective last rounds.

Stochastic stability of cooperation. To illustrate this formalism, we again use it to characterize the stability of cooperation. There is $k=1$ mutant with strategy ALLD. The remaining players use the resident strategy GTFT. When two residents interact, it follows by Eq. (8) that the outcome of the final round is distributed according to the distribution

$$\left(v_R(\text{GTFT}, \text{GTFT}), v_S(\text{GTFT}, \text{GTFT}), v_T(\text{GTFT}, \text{GTFT}), v_P(\text{GTFT}, \text{GTFT}) \right) = (1, 0, 0, 0).$$

Similarly, if the mutant interacts with a resident, the final round's outcome is distributed according to

$$\left(v_R(\text{ALLD}, \text{GTFT}), v_S(\text{ALLD}, \text{GTFT}), v_T(\text{ALLD}, \text{GTFT}), v_P(\text{ALLD}, \text{GTFT}) \right) = (0, 0, 1-\delta+\delta q, \delta(1-q)).$$

Based on these, we compute the probability $x(u_R, u_M)$ that the payoff of a randomly chosen GTFT player

is u_R and that the payoff of the ALLD player is u_M , with $u_R, u_M \in \mathcal{U}$. We obtain

$$\begin{aligned} x(R, T) &= \frac{N-2}{N-1} \cdot (1 - \delta + \delta q) & x(R, P) &= \frac{N-2}{N-1} \cdot \delta(1 - q) \\ x(S, T) &= \frac{1}{N-1} \cdot (1 - \delta + \delta q) & x(P, P) &= \frac{1}{N-1} \cdot \delta(1 - q) \\ x(u_R, u_M) &= 0 \text{ for all other payoff pairs } (u_R, u_M). \end{aligned}$$

Based on these expressions, we calculate the ratio of transition probabilities as

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{\frac{N-2}{N-1} \left(\frac{1-\delta+\delta q}{1+e^{\beta c}} + \frac{\delta(1-q)}{1+e^{-\beta(b-c)}} \right) + \frac{1}{N-1} \left(\frac{1-\delta+\delta q}{1+e^{\beta(b+c)}} + \frac{\delta(1-q)}{2} \right)}{\frac{N-2}{N-1} \left(\frac{1-\delta+\delta q}{1+e^{-\beta c}} + \frac{\delta(1-q)}{1+e^{\beta(b-c)}} \right) + \frac{1}{N-1} \left(\frac{1-\delta+\delta q}{1+e^{-\beta(b+c)}} + \frac{\delta(1-q)}{2} \right)}.$$

Cooperation is stochastically stable if $\lambda_1^-/\lambda_1^+ > 1$. While one can solve this inequality for q , the resulting condition is somewhat lengthy. To obtain a more interpretable condition, we consider the limit of strong selection $\beta \rightarrow \infty$ and large populations $N \rightarrow \infty$. In that case, because $b > c > 0$, the above ratio simplifies to

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{\delta(1-q)}{1-\delta+\delta q}. \quad (15)$$

This ratio exceeds one if $q < 1 - 1/(2\delta)$. For such a strategy to be feasible, we require $q > 0$, which in turn implies $\delta > 1/2$. Moreover, in the special case that games are infinitely repeated, $\delta \rightarrow 1$, we conclude that cooperation is stochastically stable if $q < 1/2$. (For $q = 1/2$, the payoff of the ALLD player is $T > R$ for half of the time, and it is $P < R$ for the other half. The probability that the number of mutants increase by one equals the probability that the mutant goes extinct.)

2.3 Updates based on the final round of two repeated games

Motivation. In the limited-memory setting, individuals take into account the outcome of one round, and of one interaction. This setting can be generalized such that an individual considers m rounds and of n interactions. Here, we discuss the case that the update depends on the last round of n interactions.

At each step of the evolutionary process, we consider the role model's and the learner's last n matches. We need to define the probability that for each of the matches they are paired with a mutant, with a resident or with each other. We assume that each pair is unique, such that the learner and the role model can be matched at most once, which is a reasonable assumption in large populations. The case of $n = 1$ corresponds to the previous setting of limited memory. There, we have seen that there are five possible combinations to consider. As we increase n , the number of possible combinations increases non-linearly; for a graphical illustration, see Fig. 1. In the following, we study the case of $n = 2$, such that the learner takes into account the players' final payoffs of two interactions.



Supplementary Figure 1: Possible pairs when the learner takes into account n interactions against different co-players. In this diagram, (s_i, s_j) represents a possible pairing of two players, lines indicate possible cases, and the fractions represent the respective probabilities for each case. We consider n stages in which the population members are consecutively paired with each other to play a repeated game. In the first stage, we need to consider five possible cases, as in Section 2.2. One case arises if the learner and the role model are matched directly. This happens with probability $1/(N-1)$. If they are not matched directly, both can be paired with a mutant, with a resident, or one is paired with a mutant whilst the other is paired with a resident. The second stage is similar. However, here we need to take into account that the learner and the role model can only be matched directly if they have not already interacted during the previous stage. The process continues until the n -th stage.

Computing the ratio λ_k^-/λ_k^+ . As before, we assume that either the learner or the role model is a resident, and that the other player involved in the pairwise comparison is a mutant. For $n = 2$, there are 24 possible combinations to consider. At the first stage there are five possible combinations. These are the same as in the previous setting. The learner either interacts directly with the role model or we need to take into account whether the resident interacts with a resident or a mutant, and whether the mutant interacts with a resident or a mutant. In case the learner and the role model did not directly interact with each other during the first stage, there are again five possible combinations in the second stage. Otherwise, if there already was a direct interaction, there are four combinations. Hence, there are $4 \cdot 5 + 4 = 24$ combinations in total.

We assume that player the resident receives the payoff u_{R1} with their first interaction and u_{R2} with their second. Similarly, the mutant receives payoffs u_{M1} and u_{M2} . Let $x(u_{R1}u_{R1}, u_{M1}u_{M2})$ denote the probability that the resident and the mutant received payoffs u_{R1}, u_{R2} and u_{M1}, u_{M2} in the last round of their respective

last two interactions. This probability can be computed as follows,

$$\begin{aligned}
x(u_{R1}u_{R2}, u_{M1}u_{M2}) &= \frac{1}{N-1} \cdot v_{u_{R1}}(s_R, s_M) \cdot 1_{(u_{R1}, u_{M1}) \in \mathcal{U}_F} \cdot A \\
&+ \frac{N-1}{N-2} \left[\frac{(k-1)(k-2)}{(N-2)(N-3)} v_{u_{R1}}(s_R, s_M) v_{u_{M1}}(s_M, s_M) \left(\frac{v_{u_{R2}}(s_R, s_M)}{N-2} 1_{(u_{R2}, u_{M2}) \in \mathcal{U}_F} + \frac{N-3}{N-2} [B_1 + B_2 + B_3 + B_4] \right) \right. \\
&+ \frac{(k-1)(N-k-1)}{(N-2)(N-3)} v_{u_{R1}}(s_R, s_M) v_{u_{M1}}(s_M, s_R) \left(\frac{v_{u_{R2}}(s_R, s_M)}{N-2} 1_{(u_{R2}, u_{M2}) \in \mathcal{U}_F} + \frac{N-3}{N-2} [C_1 + C_2 + C_3 + C_4] \right) \\
&+ \frac{(k-1)(N-k-1)}{(N-2)(N-3)} v_{u_{R1}}(s_R, s_R) v_{u_{M1}}(s_M, s_M) \left(\frac{v_{u_{R2}}(s_R, s_M)}{N-2} 1_{(u_{R2}, u_{M2}) \in \mathcal{U}_F} + \frac{N-3}{N-2} [D_1 + D_2 + D_3 + D_4] \right) \\
&\left. + \frac{(N-k-2)(N-k-1)}{(N-2)(N-3)} v_{u_{R1}}(s_R, s_R) v_{u_{M1}}(s_M, s_R) \left(\frac{v_{u_{R2}}(s_R, s_M)}{N-2} 1_{(u_{R2}, u_{M2}) \in \mathcal{U}_F} + \frac{N-3}{N-2} [E_1 + E_2 + E_3 + E_4] \right) \right]. \tag{16}
\end{aligned}$$

The first line on the right side corresponds to the case that the learner and the role model are matched during the first stage. The factor A considers the different pairings that can then occur in the second stage,

$$\begin{aligned}
A &= \frac{N-2}{N-1} \left[\frac{k-1}{N-2} \frac{k-2}{N-3} v_{u_{M1}}(s_R, s_M) v_{u_{M2}}(s_M, s_M) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} v_{u_{M1}}(s_R, s_M) v_{u_{M2}}(s_M, s_R) \right. \\
&\quad \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} v_{u_{M1}}(s_R, s_R) v_{u_{M2}}(s_M, s_M) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} v_{u_{M1}}(s_R, s_R) v_{u_{M2}}(s_M, s_R) \right].
\end{aligned}$$

The second line of Eq. (16) corresponds to the case that during the first stage, both the resident and the mutant interact with a mutant. In the second stage, they may interact with each other. If they do not, then again either both of them interact with a mutant, one of them interacts with a resident whereas the other interacts with a mutant, or both interact with a resident. These cases are captured by the variables

$$\begin{aligned}
B_1 &= \frac{(k-2)(k-3)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_M), & B_2 &= \frac{(k-2)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_M), \\
B_3 &= \frac{(k-2)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_R), & B_4 &= \frac{(N-k-2)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_R).
\end{aligned}$$

The third line of Eq. (16) corresponds to the case that during the first stage, the resident interacts with a mutant whereas the mutant interacts with a resident. The variables for the pairings in the second stage are

$$\begin{aligned}
C_1 &= \frac{(k-3)(k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_M), & C_2 &= \frac{(k-1)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_M), \\
C_3 &= \frac{(k-2)(N-k-2)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_R), & C_4 &= \frac{(N-k-2)^2}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_R).
\end{aligned}$$

The fourth line represents the case that during the first stage, the resident interacts with another resident whereas the mutant interacts with a mutant. The respective variables for the second-stage pairings are

$$\begin{aligned}
D_1 &= \frac{(k-2)^2}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_M), & D_2 &= \frac{(k-2)(N-k-2)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_M), \\
D_3 &= \frac{(k-1)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_R), & D_4 &= \frac{(N-k-3)(N-k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_R).
\end{aligned}$$

Finally, the last line of Eq. (16) corresponds to the case that during the first stage, both the resident and the mutant interact with a resident. The variables for the second stage are

$$\begin{aligned} E_1 &= \frac{(k-2)(k-1)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_M), & E_2 &= \frac{(k-1)(N-k-2)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_M), \\ E_3 &= \frac{(k-1)(N-k-2)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_M) v_{u_{M2}}(s_M, s_R), & E_4 &= \frac{(N-k-3)(N-k-2)}{(N-3)(N-4)} v_{u_{R2}}(s_R, s_R) v_{u_{M2}}(s_M, s_R). \end{aligned}$$

As a result, the transition probabilities for the number of mutants is now given by

$$\begin{aligned} \lambda_k^+ &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{u_{R1}, u_{R2}, u_{M1}, u_{M2}} x(u_{R1} u_{R2}, u_{M1} u_{M2}) \cdot \rho \left(\frac{u_{R1} + u_{R2}}{2}, \frac{u_{M1} + u_{M2}}{2} \right), \\ \lambda_k^- &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{u_{R1}, u_{R2}, u_{M1}, u_{M2}} x(u_{R1} u_{R2}, u_{M1} u_{M2}) \cdot \rho \left(\frac{u_{M1} + u_{M2}}{2}, \frac{u_{R1} + u_{R2}}{2} \right). \end{aligned} \quad (17)$$

Note that we assume players to use their average payoff across their two interactions for comparisons.

Stochastic stability of cooperation. As before, we can use this formalism to compute when cooperation is stochastically stable. For a single ALLD mutant in a GTFT resident population, we obtain the following non-zero probabilities $x(u_{R1} u_{R2}, u_{M1}, u_{M2})$

$$\begin{aligned} x(RR, TT) &= \frac{N-3}{N-1} (\delta q - \delta + 1)^2, & x(RR, TP) &= \frac{N-3}{N-1} \delta (1-q) (\delta q - \delta + 1), \\ x(RR, PT) &= \frac{N-3}{N-1} \delta (1-q) (\delta q - \delta + 1), & x(RR, PP) &= \frac{N-3}{N-1} \delta^2 (1-q)^2, \\ x(RS, TT) &= \frac{1}{N-1} (\delta q - \delta + 1)^2, & x(RS, PT) &= \frac{1}{N-1} \delta (1-q) (\delta q - \delta + 1), \\ x(RP, TP) &= \frac{1}{N-1} \delta (1-q) (\delta q - \delta + 1), & x(RP, PP) &= \frac{1}{N-1} \delta^2 (1-q)^2, \\ x(SR, TT) &= \frac{1}{N-1} (\delta q - \delta + 1)^2, & x(SR, TP) &= \frac{1}{N-1} \delta (1-q) (\delta q - \delta + 1), \\ x(PR, PT) &= \frac{1}{N-1} \delta (1-q) (\delta q - \delta + 1), & x(PR, PP) &= \frac{1}{N-1} \delta^2 (1-q)^2. \end{aligned}$$

The resulting ratio of transition probabilities is

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{\frac{N-3}{N-1} \left(\frac{\delta^2(1-q)^2}{1+e^{-\beta(b-c)}} + \frac{2\delta(1-q)(\delta q - \delta + 1)}{1+e^{-\beta(b/2-c)}} + \frac{(\delta q - \delta + 1)^2}{1+e^{\beta c}} \right) + \frac{2}{N-1} \left(\frac{\delta^2(1-q)^2}{1+e^{-\beta(b-c)/2}} + \frac{\delta(1-q)(\delta q - \delta + 1)}{1+e^{\beta c/2}} + \frac{\delta(1-q)(\delta q - \delta + 1)}{1+e^{\beta c}} + \frac{(\delta q - \delta + 1)^2}{1+e^{\beta(b/2+c)}} \right)}{\frac{N-3}{N-1} \left(\frac{\delta^2(1-q)^2}{1+e^{\beta(b-c)}} + \frac{2\delta(1-q)(\delta q - \delta + 1)}{1+e^{\beta(b/2-c)}} + \frac{(\delta q - \delta + 1)^2}{1+e^{-\beta c}} \right) + \frac{2}{N-1} \left(\frac{\delta^2(1-q)^2}{1+e^{\beta(b-c)/2}} + \frac{\delta(1-q)(\delta q - \delta + 1)}{1+e^{-\beta c/2}} + \frac{\delta(1-q)(\delta q - \delta + 1)}{1+e^{-\beta c}} + \frac{(\delta q - \delta + 1)^2}{1+e^{-\beta(b/2+c)}} \right)}. \quad (18)$$

Again this expression simplifies considerably in the limit of strong selection, $\beta \rightarrow \infty$, and of large popula-

tions, $N \rightarrow \infty$. In that case, we obtain the following three cases

$$\frac{\lambda_1^-}{\lambda_1^+} = \begin{cases} \frac{1-(1-\delta(1-q))^2}{(1-\delta(1-q))^2} & \text{if } \frac{b}{2} > c, \\ \frac{\delta(1-q)}{1-\delta(1-q)} & \text{if } \frac{b}{2} = c, \\ \frac{\delta^2(1-q)^2}{1-\delta^2(1-q)^2} & \text{if } \frac{b}{2} < c. \end{cases} \quad (19)$$

We note that unlike the case of limited payoff memory, here the exact magnitude of the benefit and the cost of cooperation do have some effect on the transition probabilities. By requiring that the above ratio exceeds one, we obtain the following conditions for stochastic stability of cooperation,

$$\begin{cases} q < \frac{\delta-1+\sqrt{2}/2}{\delta} & \text{if } \frac{b}{2} > c, \\ q < 1 - \frac{1}{2\delta} & \text{if } \frac{b}{2} = c, \\ q < \frac{\delta-\sqrt{2}/2}{\delta} & \text{if } \frac{b}{2} < c. \end{cases} \quad (20)$$

Only when $b/2 = c$ we obtain the same expression as in the limited-memory case, Eq. (15). The requirement $q \geq 0$ implies that the minimum continuation probability is $\delta > 1 - \sqrt{2}/2 \approx 0.293$ in the first case, $\delta > \frac{1}{2}$ in the second, and $\delta > \sqrt{2}/2 \approx 0.707$ in the third case. Conversely, if the game is infinitely repeated, $\delta \rightarrow 1$, stochastic stability of cooperation requires $q < \sqrt{2}/2 \approx 0.707$ in the first case, $q < 1/2$ in the second, and $q < 1 - \sqrt{2}/2 \approx 0.293$ in the third case.

2.4 Updates based on the last two rounds of one repeated game

General setup. As our next setting, we consider a setup in which players update their strategies based on the last two rounds of their last game. Here, additional complications may arise because some games may only last for one round if $\delta < 1$. In the following, we abstract from these complications. We shall assume that learners only update their strategies when both the learner's and the role model's last game had at least two rounds. If either the learner or the role model only engaged in a one-round game in their previous interaction, we assume the learner decides to keep its strategy. As a result, in the following, we accordingly re-interpret some of our quantities. For example, λ_k^+ is no longer the (unconditional) probability that the number of mutants increases by one after a learner and a role model have been chosen. Instead, we interpret it as the (conditional) probability that the number of mutants increases, given that the learner and the role model happened to interact at least for two rounds during their respective last interaction. This new interpretation of transition probabilities affects the time scale of fixation. After all, we ignore all time steps in which no update occurred because a player's last game only lasted for a single round. However, our re-interpretation does not affect the ratio λ_k^-/λ_k^+ , which is the key quantity we consider.

Computing the distribution of outcomes in the penultimate and in the final round of a game. To compute the ratio of transition probabilities, again we need to compute how likely it is that the resident and the mutant received any possible combination of two payoffs during the last two rounds. The probability that the game lasts for at least two rounds is δ , since that is the probability of continuing from round zero (the initial round) to round one. The distribution of actions in the penultimate round, conditioned on there being at least two rounds in the game, is then

$$\frac{1}{\delta} \sum_{t=1}^{\infty} \delta^t (1 - \delta) \mathbf{v}_0 M^{t-1} = (1 - \delta) \mathbf{v}_0 \sum_{t=0}^{\infty} \delta^t M^t, \quad (21)$$

which is identical to the expression for \mathbf{v} in Eq. (7). Thus, the probability that the first player receives the payoff $u \in \mathcal{U}$ in the penultimate round of the game, conditional on the event that the game lasts at least two rounds, is again identical to the player's average probability v_u to receive payoff u across all rounds of the game, as specified by Eq. (8).

We can now derive a probability distribution for the first player's payoff in the last two rounds of a repeated game. To this end, consider the 4×4 transition matrix $M(s_1, s_2) = (m_{u,u'})$ according to Eq. (6). Instead of the usual indexing of the four rows by numbers, $i \in \{1, 2, 3, 4\}$, here we label the four rows of this matrix by the first player's payoffs in the previous round, $u \in \{R, S, T, P\}$. Similarly, we label the four columns of this matrix by the resulting payoffs to player 1 in the next round. For example, m_{ST} corresponds to the second row and third column of matrix M . So by Eq. (6), $m_{ST} = \bar{q}_1 p_2 = (1 - q_1) p_2$. Based on this notation, we can describe the probability $w_{uu'}(s_1, s_2)$ that player 1 obtains a payoff of u in the penultimate round (probability v_u) and a payoff of u' in the last round (probability $m_{u,u'}$), giving $w_{uu'}(s_1, s_2) = v_u \cdot m_{u,u'}$.

Computing the ratio $\lambda_k^- / \lambda_k^+$. To compute the ratio of transition probabilities, let us again take a population perspective. We consider k mutants with strategy $s_M = (y_M, p_M, q_M)$ and $N - k$ residents with strategy $s_R = (y_R, p_R, q_R)$. Without loss of generality, we assume that either the learner or the role model is a resident, and that the respective other player is a mutant. Let $x(u_R u'_R, u_M u'_M)$ be the probability that the two players received payoffs u and u' in the last two rounds of their last repeated game. We can compute this probability based on the same logic as in the limited-memory setting (Section 2.2). This yields

$$\begin{aligned} x(u_R u'_R, u_M u'_M) &= \frac{1}{N-1} \cdot w_{u_R u'_R}(s_R, s_M) \cdot 1_{(u_R, u'_R) \in \mathcal{U}_F} \cdot 1_{(u'_R, u'_M) \in \mathcal{U}_F} \\ &+ \frac{N-2}{N-1} \cdot \left[\frac{k-1}{N-2} \frac{k-2}{N-3} \cdot w_{u_R u'_R}(s_R, s_M) w_{u_M u'_M}(s_M, s_M) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} \cdot w_{u_R u'_R}(s_R, s_M) w_{u_M u'_M}(s_M, s_R) \right. \\ &\quad \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} \cdot w_{u_R u'_R}(s_R, s_R) w_{u_M u'_M}(s_M, s_M) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} \cdot w_{u_R u'_R}(s_R, s_R) w_{u_M u'_M}(s_M, s_R) \right]. \end{aligned}$$

Overall, we obtain the following formula for the transition probabilities

$$\begin{aligned}\lambda_k^+ &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{u_R, u'_R, u_M, u'_M} x(u_R u'_R, u_M u'_M) \cdot \rho\left(\frac{u_R + u'_R}{2}, \frac{u_M + u'_M}{2}\right), \\ \lambda_k^- &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{u_R, u'_R, u_M, u'_M} x(u_R u'_R, u_M u'_M) \cdot \rho\left(\frac{u_M + u'_M}{2}, \frac{u_R + u'_R}{2}\right).\end{aligned}\quad (22)$$

Stochastic stability of cooperation. We once again calculate how easily a single ALLD mutant can invade a resident population of GTFT players. When two residents interact, we obtain the following probability distribution for the outcome of the two last rounds,

$$\begin{aligned}w_{RR}(\text{GTFT}, \text{GTFT}) &= 1, \\ w_{uu'}(\text{GTFT}, \text{GTFT}) &= 0 \text{ for all other } u, u' \in \mathcal{U}.\end{aligned}$$

For interactions between the mutant and a resident, we obtain

$$\begin{aligned}w_{TT}(\text{ALLD}, \text{GTFT}) &= q(\delta q - \delta + 1), & w_{TP}(\text{ALLD}, \text{GTFT}) &= (1 - q)(\delta q - \delta + 1), \\ w_{PT}(\text{ALLD}, \text{GTFT}) &= \delta q(1 - q), & w_{PP}(\text{ALLD}, \text{GTFT}) &= \delta(1 - q)^2, \\ w_{uu'}(\text{ALLD}, \text{GTFT}) &= 0 \text{ for all other } u, u' \in \mathcal{U}.\end{aligned}$$

As a consequence, we obtain the following probabilities $x(u_R u'_R, u_M u'_M)$ that the payoffs of a randomly chosen GTFT resident are u_R and u'_R , whereas the payoffs of the mutant are u_M and u'_M :

$$\begin{aligned}x(RR, TT) &= \frac{N-2}{N-1} q(\delta q - \delta + 1), & x(RR, TP) &= \frac{N-2}{N-1} (1 - q)(\delta q - \delta + 1), \\ x(RR, PT) &= \frac{N-2}{N-1} \delta q(1 - q), & x(RR, PP) &= \frac{N-2}{N-1} \delta(1 - q)^2, \\ x(SS, TT) &= \frac{1}{N-1} q(\delta q - \delta + 1), & x(SP, TP) &= \frac{1}{N-1} (1 - q)(\delta q - \delta + 1), \\ x(PS, PT) &= \frac{1}{N-1} \delta q(1 - q), & x(PP, PP) &= \frac{1}{N-1} \delta(1 - q)^2, \\ x(u_R u'_R, u_M u'_M) &= 0 \text{ for all other payoff combinations.}\end{aligned}$$

Based on these payoff probabilities, we compute the ratio

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{\frac{N-2}{N-1} \left(\frac{\delta(1-q)^2}{1+e^{-\beta(b-c)}} + \frac{q(\delta q - \delta + 1)}{1+e^{\beta c}} + \frac{(1-q)(2\delta q - \delta + 1)}{1+e^{-\beta(b/2-c)}} \right) + \frac{1}{N-1} \left(\frac{\delta(1-q)^2}{2} + \frac{q(\delta q - \delta + 1)}{1+e^{\beta(b+c)}} + \frac{(1-q)(2\delta q - \delta + 1)}{1+e^{\beta(b+c)/2}} \right)}{\frac{N-2}{N-1} \left(\frac{\delta(1-q)^2}{1+e^{\beta(b-c)}} + \frac{q(\delta q - \delta + 1)}{1+e^{-\beta c}} + \frac{(1-q)(2\delta q - \delta + 1)}{1+e^{\beta(b/2-c)}} \right) + \frac{1}{N-1} \left(\frac{\delta(1-q)^2}{2} + \frac{q(\delta q - \delta + 1)}{1+e^{-\beta(b+c)}} + \frac{(1-q)(2\delta q - \delta + 1)}{1+e^{-\beta(b+c)/2}} \right)}.\quad (23)$$

In the limit of strong selection $\beta \rightarrow \infty$ and large populations $N \rightarrow \infty$, this ratio simplifies to

$$\frac{\lambda_1^-}{\lambda_1^+} = \begin{cases} \frac{(1-q)(1+\delta q)}{q(\delta q - \delta + 1)} & \text{if } \frac{b}{2} > c, \\ \frac{(1+\delta)(1-q)}{\delta q - \delta + q + 1} & \text{if } \frac{b}{2} = c, \\ \frac{\delta(1-q)^2}{1 - \delta(1-q)^2} & \text{if } \frac{b}{2} < c. \end{cases} \quad (24)$$

For stochastic stability of cooperation we require this ratio to be larger than one. This yields the conditions

$$\begin{cases} q < \frac{1}{1 - \delta + \sqrt{1 + \delta^2}} & \text{if } \frac{b}{2} > c, \\ q < \frac{\delta}{1 + \delta} & \text{if } \frac{b}{2} = c, \\ q < 1 - \frac{1}{\sqrt{2\delta}} & \text{if } \frac{b}{2} < c. \end{cases} \quad (25)$$

Interestingly, these conditions differ from the respective conditions in the previous section, Eq. (20). Moreover, the requirement $q \geq 0$ imposes no constraints in the first two cases; it imposes the constraint $\delta > 1/2$ in the last case. However, in the limit of long games, $\delta \rightarrow 1$ we recover the same conditions as in the previous section. Stochastic stability of cooperation requires $q < \sqrt{2}/2 \approx 0.707$ in the first case, $q < 1/2$ in the second, and $q < 1 - \sqrt{2}/2 \approx 0.293$ in the third case.

2.5 Updates based on the last two rounds of two repeated games

As our final extension, we consider a combination of the previous two settings, discussed in Sections 2.3 and 2.4. Now, individuals take into account two interactions with different co-players, and from each interaction, they take into account the last two rounds. The probability that the number of mutants increases or decreases by one is now given by

$$\begin{aligned} \lambda_k^+ &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{\substack{u_{R1}, u'_{R1}, u_{R2}, u'_{R2} \\ u_{M1}, u'_{M1}, u_{M2}, u'_{M2}}} x(u_{R1} u'_{R1} u_{R2} u'_{R2}, u_{M1} u'_{M1} u_{M2} u'_{M2}) \cdot \rho\left(\frac{u_{R1} + u'_{R1} + u_{R2} + u'_{R2}}{4}, \frac{u_{M1} + u'_{M1} + u_{M2} + u'_{M2}}{4}\right), \\ \lambda_k^- &= \frac{N-k}{N} \cdot \frac{k}{N} \sum_{\substack{u_{R1}, u'_{R1}, u_{R2}, u'_{R2} \\ u_{M1}, u'_{M1}, u_{M2}, u'_{M2}}} x(u_{R1} u'_{R1} u_{R2} u'_{R2}, u_{M1} u'_{M1} u_{M2} u'_{M2}) \cdot \rho\left(\frac{u_{M1} + u'_{M1} + u_{M2} + u'_{M2}}{4}, \frac{u_{R1} + u'_{R1} + u_{R2} + u'_{R2}}{4}\right). \end{aligned} \quad (26)$$

Though we do not pursue any further analytical exploration of this case, in the next section we present simulation results for this setting, as well as for the other settings.

2.6 Updates based on expected payoff of one repeated game

Motivation.

Computing the ratio λ_k^-/λ_k^+ . We revisit the population setup, consisting of k mutants with the strategy $s_M = (y_M, p_M, q_M)$ and $N - k$ residents with strategies $s_R = (y_R, p_R, q_R)$. In each step of the evolutionary process, we select a learner and a role model. The learner or the role model can be either a resident, and the other player is a mutant. We consider that the learner and the model update based on their most recent interaction. As mentioned previously, two major cases must be considered:

1. The learner and the role model have had their most recent interaction with each other.
2. The learner's last interaction was not with the role model.

In the latter case, there are four different subcases based on whether the resident's last interaction partner was a mutant or a resident and whether the mutant's last interaction partner was a mutant or a resident.

We assume that the players' payoff received from their last interaction is their expected payoff given by Eq. (9). Thus, a resident interacting with a resident receives the payoff $\pi(s_R, s_R)$, and $\pi(s_R, s_M)$ when interacting with a mutant. Similarly, a mutant interacting with a resident receives the payoff $\pi(s_M, s_R)$ and $\pi(s_M, s_M)$ when interacting with a mutant.

After these preparations, we can define the probability that the number of mutants either increases or decreases by one.

$$\begin{aligned} \lambda^+ = & \frac{1}{N-1} \cdot \rho(\pi(s_R, s_M), \pi(s_M, s_R)) \\ & + \frac{N-2}{N-1} \cdot \left[\frac{k-1}{N-2} \frac{k-2}{N-3} \cdot \rho(\pi(s_R, s_M), \pi(s_M, s_M)) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} \cdot \rho(\pi(s_R, s_M), \pi(s_M, s_R)) \right. \\ & \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} \cdot \rho(\pi(s_R, s_R), \pi(s_M, s_M)) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} \cdot \rho(\pi(s_R, s_R), \pi(s_M, s_R)) \right], \end{aligned}$$

$$\begin{aligned} \lambda^- = & \frac{1}{N-1} \cdot \rho(\pi(s_M, s_R), \pi(s_R, s_M)) \\ & + \frac{N-2}{N-1} \cdot \left[\frac{k-1}{N-2} \frac{k-2}{N-3} \cdot \rho(\pi(s_M, s_M), \pi(s_R, s_M)) + \frac{k-1}{N-2} \frac{N-k-1}{N-3} \cdot \rho(\pi(s_M, s_R), \pi(s_R, s_M)) \right. \\ & \left. + \frac{N-k-1}{N-2} \frac{k-1}{N-3} \cdot \rho(\pi(s_M, s_M), \pi(s_R, s_R)) + \frac{N-k-1}{N-2} \frac{N-k-2}{N-3} \cdot \rho(\pi(s_M, s_R), \pi(s_R, s_R)) \right]. \end{aligned}$$

Stochastic stability of cooperation. We once again calculate how easily a single ALLD mutant can invade into a resident population of GTFT players. When resident interacts with another resident or with a mutant,

the expected payoffs based on Eq. (9) of the interactions are,

$$\pi(s_R, s_R) = b - c, \quad \pi(s_R, s_M) = -c(\delta(q - 1) + 1).$$

The expected payoff of the mutant, are

$$\pi(s_M, s_R) = b(\delta(q - 1) + 1), \quad \pi(s_M, s_M) = 0.$$

Based on these expected payoffs, we compute the ratio

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{\frac{1}{N-1} \frac{1}{(1+e^{\beta(b(\delta q - \delta) + c)})} + \frac{N-2}{N-1} \frac{1}{(1+e^{-\beta(-b(\delta q - \delta + 1) + b - c)})}}{\frac{1}{N-1} \frac{1}{(1+e^{-\beta(b(\delta q - \delta) + c)})} + \frac{N-2}{N-1} \frac{1}{(1+e^{-\beta(b(\delta q - \delta + 1) - b + c)})}}. \quad (27)$$

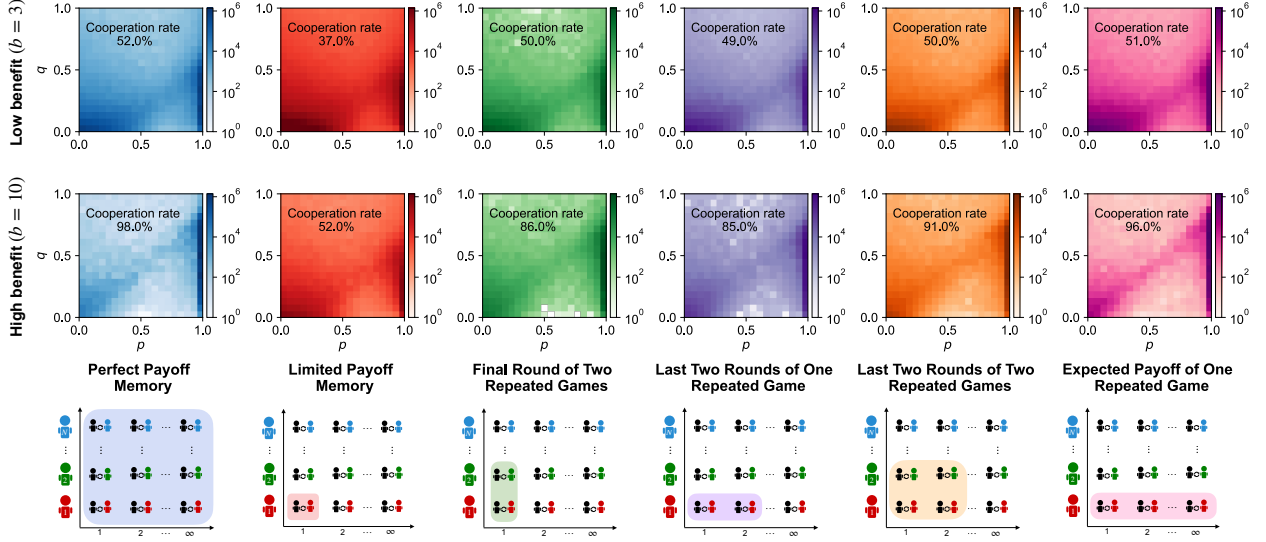
In the limit of large populations $N \rightarrow \infty$, this ratio simplifies to

$$\frac{\lambda_1^-}{\lambda_1^+} = \frac{1 + e^{-\beta(b(\delta q - \delta) + c)}}{1 + e^{\beta(b(\delta q - \delta) + c)}}. \quad (28)$$

For stochastic stability of cooperation we require this ratio to be larger than one, which is equivalent to

$$q < 1 - \frac{c}{\delta b} \quad (29)$$

Interestingly, this conditions is the same as the condition for stochastic stability in the case of perfect memory, see Eq. (12).

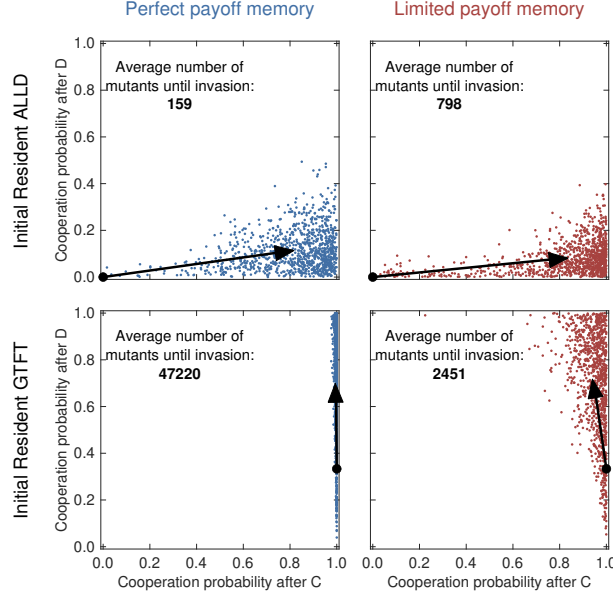


Supplementary Figure 2: Evolutionary dynamics of our six different settings. From left to right, we present results on the following cases: individuals update their strategies based on (i) the expected payoffs (perfect memory); (ii) the last-round payoff of one interaction (limited memory); (iii) the last round payoff of two interactions; (iv) the payoffs of the last two rounds of one interaction; (v) the payoff of the last two rounds of two interactions and (vi) the expected payoff of one interaction. Simulations are run either for a comparably high benefit of cooperation ($b/c = 10$), or for a low benefit ($b/c = 3$). We run each simulation for $T = 10^7$ time steps. For each time step we record the current resident population (y, p, q) . The graphs show how often the resident population chooses each combination (p, q) ; because $\delta \approx 1$, we do not show the distribution of the player's first-round cooperation probability y .

3 Further simulation results

3.1 Further simulation results for reactive strategies

Strategies favored by selection. To gain some insights into the strategies evolving in each of the settings described in the previous section, we simulate the evolutionary process according to Algorithm 1. Each time step, we record the current resident population (y, p, q) . The results of these simulations are shown in Supplementary Fig. 2. Qualitatively, all settings lead to similar dynamics as in the case of perfect and limited memory described in the main text. In particular, populations are clustered around ALLD, and around a strip of conditionally cooperative strategies, $\text{GTFT} = (1, 1, q)$. In all settings, we observe that along this strip, strategies are most abundant when they satisfy the condition for stochastic stability. For perfect memory, this means the resident population predominantly adopts a strategy with $q \leq 1 - \frac{c}{b} = 0.9$ for high benefit and $q \leq 1 - \frac{c}{b} = 0.67$ for low benefit. For limited memory, the generosity of a conditional cooperator satisfies $q < \frac{1}{2}$ for both values of benefit. In the next two cases, when players either remember the last round of two interactions or the last two rounds of one interaction, stochastic stability requires $q < \frac{\sqrt{2}}{2} \approx 0.701$. In the final case, we observe that the generosity of a conditional cooperator is in line with the condition for stochastic stability in the perfect memory case: $q < 1 - \frac{c}{b} = 0.9$ for high benefit and $q < 1 - \frac{c}{b} = 0.67$ for



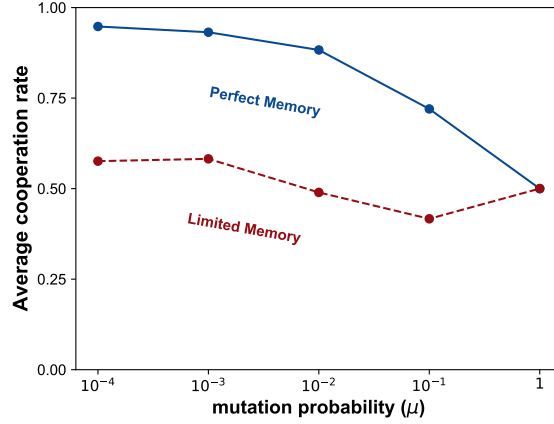
Supplementary Figure 3: Invasion analysis for perfect and limited payoff memory. We consider two residents, $ALLD = (0, 0, 0)$ and $GTFT = (1, 1, 1/3)$ for both perfect and limited payoff memory. In each case, we simulate the pairwise comparison process 10^3 times and record the first mutant strategy that reaches fixation (colored dots). Black arrows indicate the arithmetic mean of the successful mutant strategies. In addition, we record how many mutant strategies it takes on average to successfully invade a given resident. We observe that under perfect memory, defectors are more prone to being invaded, whereas conditional cooperators are more resilient to invasion. Parameters are the same as in Fig. 1 of the main text, with $b = 10$, $c = 1$, $N = 100$, $\beta = 1$, $\delta = 0.999$.

low benefit. In all cases, the simulation results are well aligned with these theoretical predictions.

Invasion dynamics. To gain some further into the dynamics over time, and into the differences between perfect and limited payoff memory, we perform an invasion analysis, see Supplementary Fig. 3. Here, we consider two possible resident strategies. Either populations use $ALLD = (0, 0, 0)$ (top row) or they start out with $GTFT = (1, 1, 1/3)$ (bottom row). In each case, we simulate the process until a mutant reaches fixation, for both perfect memory (left) and limited memory (right). We iterate this simulation 10^3 times. Colored dots indicate the position of successful mutants, and arrows indicate the average trajectory of evolution. Moreover, we also record how long it takes on average until the first mutant strategy takes over. This plot shows that $ALLD$ populations are typically invaded by mutants similar to Tit-for-Tat, with $p > 0$ and $q \approx 0$. In contrast, $GTFT$ populations are typically invaded by other $GTFT$ mutants, but with higher levels of generosity. A further analysis of the fixation time can shed some light on why perfect memory leads to higher average cooperation rates: Compared to limited memory, it takes fewer mutants to successfully invade into $ALLD$, and more mutants to invade into $GTFT$.

Impact of mutations. All of our previous simulations are run under the assumption that mutations are

rare, $\mu \rightarrow 0$. This allows us to use the formula (3) for a mutant's fixation probability. While this limit allows for more efficient computation procedures, it rules out the possibility that two or three strategies might stably coexist in a population [11]. To explore the effect of positive mutation rates, we perform five independent runs of the pairwise process described in Section 1 for different values of μ . At each time step, we record the average cooperation rate of the resident population. The results for different mutation rates are shown in Fig. 4. The cooperation rate is higher when players have perfect memory (compared to limited memory), regardless of the mutation rate. Only as the mutation rate approaches one, and the processes become fully random, does the cooperation rate approach $1/2$ in both simulations.



Supplementary Figure 4: Evolutionary dynamics results for perfect and limited memory for different mutation values. We perform five independent simulations. Simulations are run for $T = 4 \times 10^7$ time steps for each parameter. In each time step, we introduce a new mutant with a probability μ , and we then select two random players to serve as the role model and the learner. The learner adopts the strategy of the role model with a probability $\rho(\pi_L, \pi_{RM})$ where the payoffs depend on the relevant setting. We plot the average cooperation rate within the resident population for each value of μ . The results show that our previous results for rare mutations hold more generally: While limited memory allows the evolution of moderate cooperation levels, perfect memory is always more conducive to the evolution of cooperation. Parameters are the same as before, $N = 100, c = 1, b = 10, \beta = 1, \delta = 0.999$.

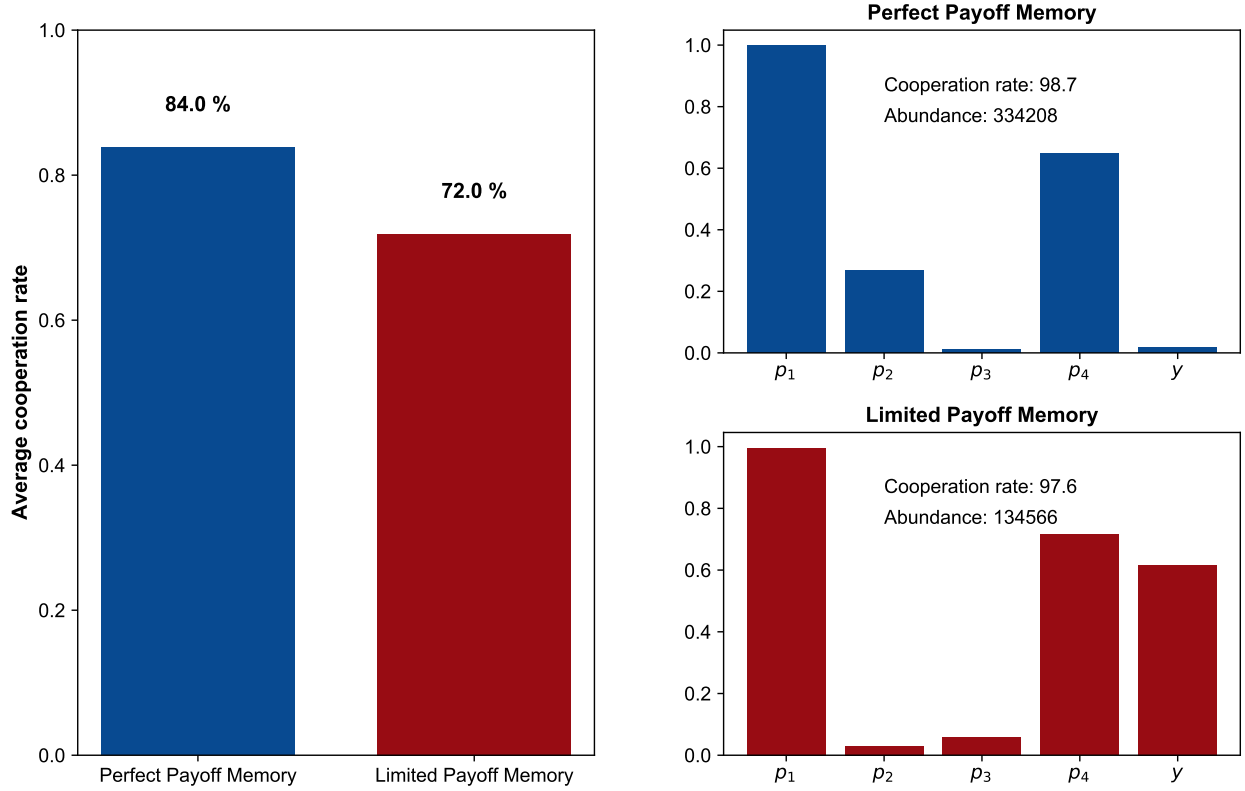
3.2 Simulations for memory-one strategies

So far, we have assumed that individuals adopt reactive strategies. To demonstrate that our results hold for more complex strategy sets, in the following we present results when players can choose among all memory-one strategies [7]. Players with these strategies consider the full outcome of the previous round to decide on their next action. There are four possible outcome in each round; $(C, C), (C, D), (D, C), (D, D)$. A memory-one strategy s can thus be written as a five-dimensional vector $s = (y, p_1, p_2, p_3, p_4)$. As with reactive strategies, the entry y is the probability that the strategy opens with a cooperation. The other entries p_1, p_2, p_3, p_4 are the probabilities that the strategy cooperates in all subsequent rounds, depending on the outcome of the previous round. Every reactive strategy (y, p, q) can be written as a memory-one strategy,

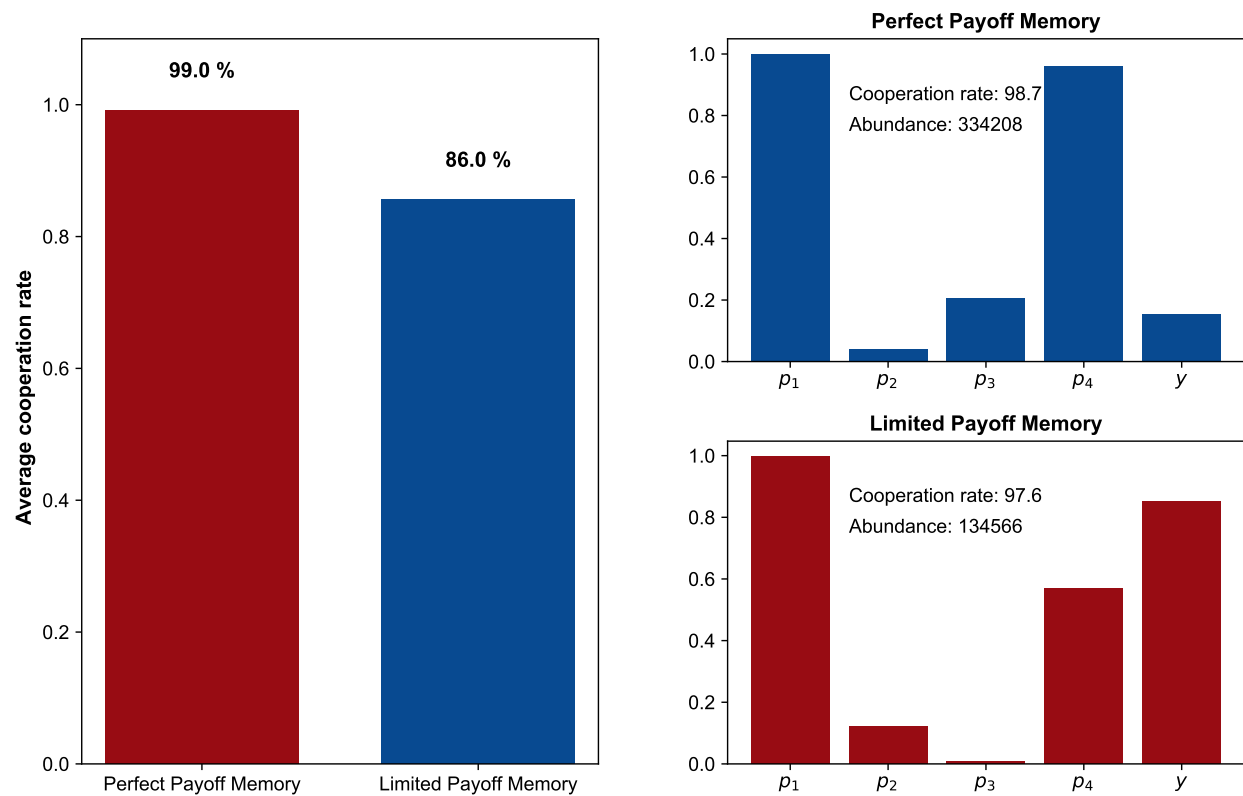
by using the memory-one representation (y, p, q, p, q) . Conversely, there are memory-one strategies that cannot be represented by a reactive strategy, such as the strategy $\text{WSLS} = (1, 1, 0, 0, 1)$. Therefore, the set of memory-one strategies is a strict superset of the reactive strategies.

To explore the dynamics among memory-one strategies, we perform four separate simulations, for both perfect and limited memory, and for two different benefit values. Fig. 5 shows the results for a comparably low benefit of cooperation; Fig. 6 shows the corresponding results for a large benefit. These simulations suggest that even when individuals are allowed to use memory-one strategies, there is more cooperation when players update their strategies based on perfect payoff memory.

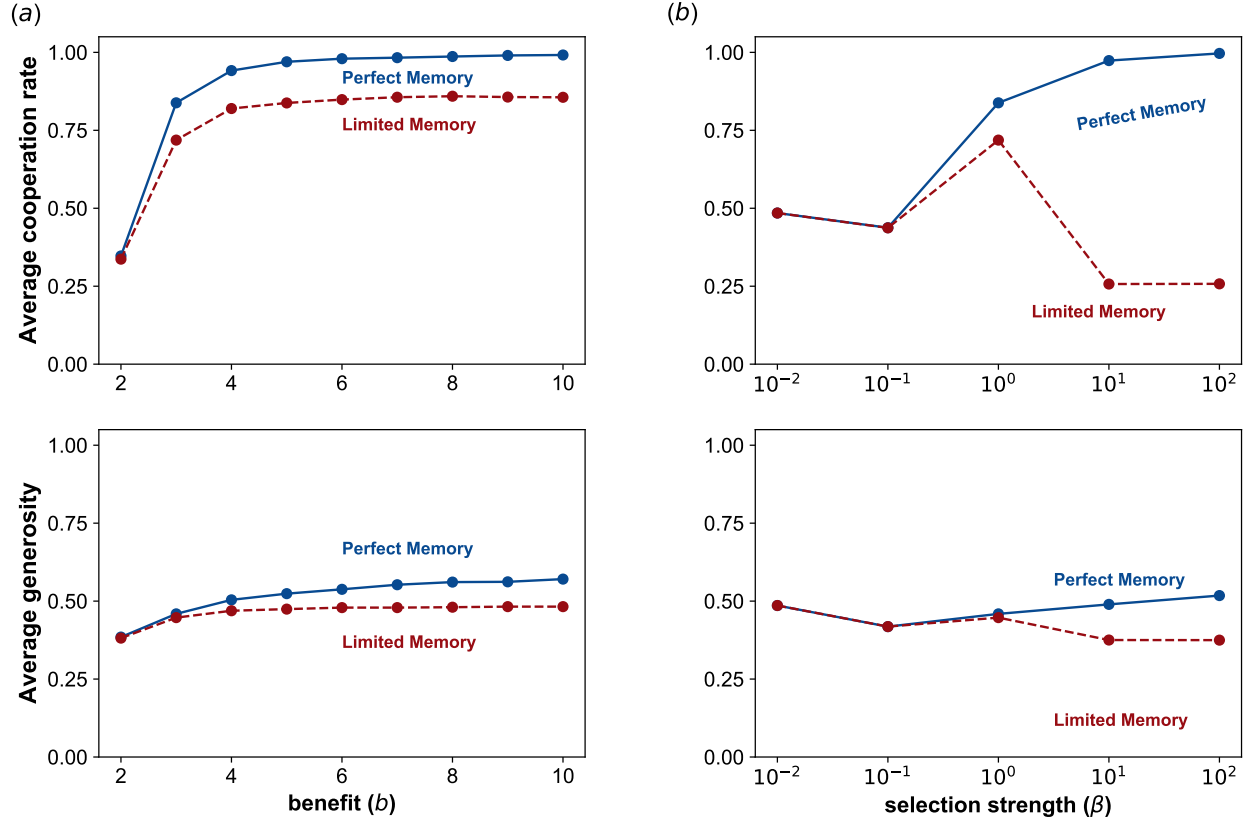
To further investigate the influence of different parameters, we have systematically varied the benefit b and the selection strength β , as shown in Fig. 7. When compared to the case of reactive strategies, what we observe here is that in the case of limited payoff memory, cooperation rates increase with an increasing benefit b and do not remain stable (Fig. 7(a)). However, even in this scenario, we still observe that perfect memory consistently results in a higher cooperation rate. Regarding the effect of different selection strengths, as depicted in Fig. 7(b), the results once again suggest that both perfect and limited payoff memory yield similar cooperation rates for weak selection ($\beta < 1$). Beyond weak selection, increasing selection has a positive effect under perfect payoff memory but a negative effect under limited payoff memory.



Supplementary Figure 5: Evolution of memory-one strategies when the benefit of cooperation is comparably small. We perform two independent simulations. In one simulation, individuals use expected payoffs. In the other, they update their strategy based on the payoff they received in the last round of their last interaction. We run each simulation for $T = 10^8$ time steps. For each time step, we record the current resident population, which is now of the form (y, p_1, p_2, p_3, p_4) . In the left panel, we report the cooperation rates for each simulation. For memory-one strategies, like for reactive strategies, the use of expected payoffs results in more cooperation. The right panel reports the most abundant strategy of each simulation. Abundance is the number of mutants a strategy can repel before being invaded. The most abundant strategies have some similarities, namely, $p_1 \approx 1$, $p_3 \approx 0$ and $p_4 > \frac{1}{2}$. Parameters: $N = 100$, $c = 1$, $b = 3$, $\beta = 1$, $\delta = 0.999$.



Supplementary Figure 6: Evolution of memory-one strategies when the benefit is high. We consider a similar setup as in Fig. 5, but now using a larger benefit value $b=10$.



Supplementary Figure 7: Evolution of direct reciprocity for different parameter values for memory-one strategies. To further explore the robustness of our results in the case of memory-one strategies, we run simulations for different benefit values (left panels, A) and for different selection strengths (right panels, B). These simulations suggest that perfect payoff memory consistently leads to more cooperation and more generosity. Unless explicitly varied, the parameters of the simulation are $N = 100$, $b = 3$, $c = 1$, $\beta = 1$, $\delta = 0.999$. Simulations are run for $T = 5 \times 10^7$ time steps for each parameter combination.

References

- [1] Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).
- [2] Traulsen, A., Pacheco, J. M. & Nowak, M. A. Pairwise comparison and selection temperature in evolutionary game dynamics. *Journal of theoretical biology* **246**, 522–529 (2007).
- [3] Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *Journal of Economic Theory* **131**, 251–262 (2006).
- [4] Wu, B., Gokhale, C. S., Wang, L. & Traulsen, A. How small are small mutation rates? *Journal of Mathematical Biology* **64**, 803–827 (2012).
- [5] McAvoy, A. Comment on “Imitation processes with small mutations”. *J. Econ. Theory* **159**, 66–69 (2015).
- [6] Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).
- [7] Sigmund, K. *The calculus of selfishness* (Princeton University Press, 2010).
- [8] Molander, P. The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution* **29**, 611–618 (1985).
- [9] Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).
- [10] Schmid, L., Chatterjee, K., Hilbe, C. & Nowak, M. A unified framework of direct and indirect reciprocity. *Nature Human Behaviour* **5**, 1292–1302 (2021).
- [11] Tkadlec, J., Hilbe, C. & Nowak, M. A. Mutation enhances cooperation in direct reciprocity. *Proceedings of the National Academy of Sciences USA* **120**, e2221080120 (2023).

CH: Adjust symbols in main text – for role model, strategy, etc