Evolution of reciprocity with limited payoff memory

Nikoleta E. Glynatsi[1], Alex McAvoy[2,3,†], Christian Hilbe[1,†]

[1]Max Planck Research Group on the Dynamics of Social Behavior,
Max Planck Institute for Evolutionary Biology, Plön, Germany
[2]School of Data Science and Society, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599
[3]Department of Mathematics, University of North Carolina at Chapel Hill,
Chapel Hill, NC 27599
† A.M. and C.H. contributed equally to this work.

Abstract

Direct reciprocity is a mechanism for the evolution of cooperation in repeated social interactions. According to this literature, individuals naturally learn to adopt conditionally cooperative strategies if they have multiple encounters with their partner. Corresponding models have greatly facilitated our understanding of cooperation, yet they often make strong assumptions on how individuals remember and process payoff information. For example, when strategies are updated through social learning, it is commonly assumed that individuals compare their average payoffs. This would require them to compute (or remember) their payoffs against everyone else in the population. To understand how more realistic constraints influence direct reciprocity, we consider the evolution of conditional behaviors when individuals learn based on more recent experiences. Even in the most extreme case that they only take into account their very last interaction, we find that cooperation can still evolve. However, such individuals adopt less generous strategies, and they cooperate less often than in the classical setup with average payoffs. Interestingly, once individuals remember the payoffs of two or three recent interactions, cooperation rates quickly approach the classical limit. These findings contribute to a literature that explores which kind of cognitive capabilities are required for reciprocal cooperation. While our results suggest that some rudimentary form of payoff memory is necessary, it suffices to remember a few interactions.

Keywords: Evolution of cooperation; evolutionary dynamics; direct reciprocity; repeated prisoner's dilemma; social learning

1 Introduction

Evolutionary game theory describes the dynamics of populations when an individual's fitness depends on the traits or strategies of other population members (1–4). This theory can be used to describe the dynamics of animal conflict (5), cancer cells (6), and of cooperation (7). Respective models translate strategic interactions into games (8). These games specify how individuals (players) interact, which strategies individuals can choose, and what fitness consequences (or payoffs) the different strategies have. In addition, these models also specify the mode by which successful strategies spread over time. In models of biological evolution, individuals with a high fitness produce more offspring; in models of cultural evolution, such individuals are imitated more often. Although biological and cultural evolution are sometimes treated as equivalent, there can be important differences (9–11). For example, models of biological evolution do not require individuals to have any particular cognitive abilities. Here, it is the evolutionary process itself that biases the population towards strategies with higher fitness. In contrast, in models of cultural evolution, individuals need to be aware of the different strategies present in the population, and they need to identify those strategies with a higher payoff. As a consequence, evolutionary outcomes may depend on how easily different behaviors can be learned (12), as well as on how readily payoffs can be compared.

These difficulties to learn strategies by social imitation are particularly pronounced in models of direct reciprocity. This literature follows Trivers' insight that individuals have more of an incentive to cooperate in social dilemmas when they interact repeatedly (13). In repeated interactions, individuals can condition their behavior on their past experiences with their interaction partner. They may use strategies such as Tit-for-Tat (14, 15) or Generous Tit-for-Tat (16, 17) to preferentially cooperate with other cooperators. Such conditional strategies approximate human behavior fairly well (18–22) and they have also been documented in several other species (23–25)–although direct reciprocity is generally more difficult to demonstrate in animals (26–28). However, at the outset, it is not clear how easy it is to *learn* reciprocal strategies by social imitation. As one obstacle, even if others' strategies are perfectly observable, individuals might find it difficult to identify which ones have the highest payoff. After all, the payoff of a strategy of direct reciprocity is not determined by the outcome of any single round. Rather, it is determined by how well this strategy fares over an entire sequence of rounds, against many different population members. In practice, such information might be difficult both to obtain and to process.

Most models of direct reciprocity do not address these difficulties (29–46) and, instead, assume individuals can easily copy the strategies of others. Similarly, they assume that updating decisions are based on the strategies' average (or expected) payoffs, which are based on all rounds and all interactions. These assumptions create a curious inconsistency in how models represent an individual's cognitive abilities. On the one hand, when playing the game, individuals are often assumed to have restricted memory. Respective studies typically assume that individuals make their decisions each round based on the outcome of the last

round only (with only a few exceptions, see Refs. 47–52). Yet when learning new strategies, individuals are assumed to remember (or compute) each others' precise average payoff across many rounds and many interaction partners. Herein, we wish to explore whether this latter assumption is actually necessary for the evolution of reciprocity through social imitation. We ask whether individuals can learn to adopt reciprocal strategies even when learning is based on payoff information from a limited number of rounds.

To explore that question, we theoretically study imitation dynamics in the repeated prisoner's dilemma, using two extreme scenarios. The first scenario is the classical modeling approach. Here, individuals update their strategies based on their expected payoffs. We contrast this model with an alternative scenario where individuals update their strategies based on the very last (one-shot) payoff they obtained. We find that individuals with limited payoff memory tend to adopt less generous strategies. Yet moderate levels of cooperation can still evolve. Moreover, as we increase the individuals' payoff memory to include the last two or three one-shot payoffs, cooperation rates quickly approach the rates observed in the classical baseline.

2 Model and Methods

To explore the impact of limited payoff memory, we adapt existing models of the evolution of direct reciprocity. These models involve two different time scales. The short time scale describes the game dynamics. Here, individuals with fixed strategies are randomly matched to interact with each other in repeated social dilemmas. The long time scale describes the evolutionary dynamics. Here, individuals can update their repeated-game strategies based on the payoffs they yield. In the following, we introduce the basic setup of our model; all details and derivations are described in the electronic supplementary material.

Description of the game dynamics. We consider a well-mixed population consisting of N players. Players are randomly matched in pairs to participate in a repeated donation game (53) with their respective co-player. Each round, they can either cooperate (C) or defect (D). A cooperating player provides a benefit b to the other player at their own cost c, with $0 < c < b$. A defecting player provides no benefit and pays no cost. Thus, the players' payoffs in a single round are given by the matrix

$$\begin{array}{c} \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cc} C & D \end{array} \\ \begin{pmatrix} b-c & -c \\ b & 0 \end{pmatrix} \end{array}. \tag{1}$$

In particular, payoffs take the form of a prisoner's dilemma: Mutual cooperation yields a better payoff than mutual defection ($b-c > 0$), but each player individually prefers to defect, independent of the co-player's action ($b > b-c$ and $0 > -c$). To incorporate repetition, we assume that after each round, there is a constant continuation probability δ of interacting for another round. For $\delta = 0$, we recover the case of a

conventional (one-shot) prisoner's dilemma. Here, mutual defection is the only equilibrium. As δ increases, the game turns into a repeated game. Here, additional equilibria emerge, with some of them allowing for full cooperation (54–57).

In a one-shot donation game, players can only choose among two pure strategies (they can either cooperate or defect). In the repeated game, strategies become arbitrarily complex. Here, strategies are contingent rules, telling players what to do depending on the outcome of all previous rounds. For simplicity, in the following we assume individuals use *reactive strategies* (17). A reactive strategy only depends on the other player's action in the last round. Such strategies can be written as a three-dimensional tuple $s = (y, p, q)$. The first entry y is the probability that the player opens with cooperation in the first round. The two other entries are the probabilities that the player cooperates in all subsequent rounds, depending on whether the co-player cooperated (p) or defected (q) in the previous round. The set of reactive strategies is simple enough to facilitate an explicit mathematical analysis (1). Yet it is rich enough to capture several important strategies of repeated games. For example, it contains $\texttt{ALLD} = (0, 0, 0)$, the strategy that always defects. Similarly, it contains Tit-for-Tat, $\texttt{TFT} = (1, 1, 0)$, the strategy that copies the co-player's previous action (and that cooperates in the first round). Finally, it contains Generous Tit-for-Tat, $\texttt{GTFT} = (1, 1, q)$, where $q > 0$ reflects a player's generosity in response to a co-player's defection (16, 17).

In the short run, the players' strategies are taken to be fixed. Players use their strategies to decide whether to cooperate in a series of repeated games against all other population members. In the long run, however, the players' strategies may change depending on the payoffs they yield, as we describe in the following.

Description of the evolutionary dynamics. Herein, we assume population members update their strategies based on social learning. To model these strategy updates, we consider a pairwise comparison process (58). This process assumes that at regular time intervals, one population member is randomly selected, and given the chance to revise its strategy. We refer to this player as the "learner". With probability μ (reflecting a mutation rate), the learner simply adopts a random strategy (all reactive strategies have the same probability to be chosen). With the converse probability $1-\mu$, the learner randomly picks a "role model" from the population. The learner then compares its own payoff π_{L} from the repeated game to the role model's payoff π_{RM}. The learner adopts the role model's strategy with a probability φ described by a Fermi function (59, 60),

$$\varphi\left(\pi_{\mathrm{L}}, \pi_{\mathrm{RM}}\right) = \frac{1}{1+e^{-\beta\left(\pi_{\mathrm{RM}}-\pi_{\mathrm{L}}\right)}}. \tag{2}$$

The selection strength parameter $\beta \geq 0$ indicates how sensitive players are to payoff differences. For $\beta = 0$, payoff differences are irrelevant, and the learner simply adopts the role model's strategy with probability $1/2$. As the selection strength β increases, players are increasingly biased to imitate the role model only if it has the higher payoff.

We deviate from previous models in how we interpret the payoffs π_{L} and π_{RM}, which form the basis of the pairwise comparisons in Eq. (2). In previous work, these payoffs are taken to be the respective players' expected payoffs. We interpret that setup as a model with perfect payoff memory. There, the payoffs π_{L} and π_{RM} represent an average over all possible repeated games the two individuals have played with all population members (Figure 1, upper left panel). The use of expected payoffs is mathematically convenient, because explicit formulas for these payoffs are available (1). Herein, we compare this model of perfect payoff memory to a model with limited payoff memory. In the latter model, the players' payoffs π_{L} and π_{RM} are taken to be the payoffs that each player received in their very last round prior to making social comparisons. That is, players only consider the very last repeated game they participated in, and there they only take into account the outcome of the very last round (Figure 1, lower left panel). This assumption could reflect, for example, a strong recency bias in how individuals evaluate payoffs. In addition to this extreme case of limited payoff memory, later on we also explore cases in which players take into account the outcome of two, three, four, or more recent rounds.

Both in the case of perfect and limited memory, we iterate the elementary strategy update step described above many times. This gives rise to a stochastic process that describes which strategies players adopt over time. Provided the selection strength is finite, this process satisfies the mathematical property of being ergodic. This implies that in the long run, the time average of the players' cooperation rates converges, and that this limit is independent of the initial population. We explore the dynamics of this process mathematically and with numerical simulations. These simulations are run sufficiently long for convergence to occur (we numerically checked that independent runs are within a 1% error tolerance). For the results presented in the following, we assume that mutations are rare ($\mu \to 0$). This assumption is fairly common in evolutionary game theory, both because it makes some computations more efficient (61–63) and because the results can be interpreted more easily. However, in Section 3 of the electronic supplementary material we show that our main results continue to hold for strictly positive mutation rates.

3 Results

Stability of cooperative populations. To get some intuition for the differences between perfect and limited payoff memory, we first analyze when full cooperation is stable in either scenario. For a population to be fully cooperative, players need to cooperate in the first round (that is, $y = 1$) and after any round with mutual cooperation (that is, $p = 1$). We refer to such a strategy as Generous Tit-for-Tat, $\texttt{GTFT} = (1, 1, q)$. To explore whether such a population consisting of GTFT players is stable, we introduce a single mutant who adopts ALLD. We say *(full) cooperation is stochastically stable* if the single mutant is more likely to imitate the residents than vice versa. For simplicity, we consider a large population ($N \to \infty$) and strong selection ($\beta \to \infty$). More general results are derived in the electronic supplementary material.

In the case of perfect payoff memory, it is straightforward to characterize when cooperation is stochasti-

cally stable. Here, we simply need to compute the players' expected payoffs. Because the population mostly consists of residents, and because residents mutually cooperate with each other, their expected payoff is $\pi_{\texttt{GTFT}} = b-c$. On the other hand, the defecting mutant only interacts with residents. Given the residents' strategy, the mutant receives a benefit b in the first round, and in every subsequent round with probability q. As a result, the mutant's expected payoff is $\pi_{\texttt{ALLD}} = (1-\delta+\delta q)b$. For perfect payoff memory, the requirement for cooperation to be stochastically stable reduces to the condition $\pi_{\texttt{GTFT}} > \pi_{\texttt{ALLD}}$, which yields

$$q < 1-\frac{c}{\delta b}. \tag{3}$$

In particular, we recover the previous observation that $q = 1 - c/(\delta b)$ is the maximum generosity that cooperators should have (16, 17, 64). Moreover, because q needs to be non-negative, we also conclude that cooperation can only be stable if $\delta > c/b$. Again, this condition for the feasibility of direct reciprocity already exists in the literature (7).

The logic of the case with limited payoff memory is somewhat different. Here we need to compute how likely each player obtains one of the four possible payoffs $\{b-c, -c, b, 0\}$ in the very last round of a game, before they make social comparisons. Because residents almost always interact with other residents, their last one-shot payoff is $\pi_{\texttt{GTFT}} = b-c$ almost surely. For the defecting mutant, there are two possibilities. If the mutant's co-player happens to cooperate in the last round, the mutant receives $\pi_{\texttt{ALLD}} = b$. This case occurs with probability $1-\delta+\delta q$. On the other hand, if the co-player defects in the last round, the mutant receives $\pi_{\texttt{ALLD}} = 0$. This occurs with the converse probability $\delta(1-q)$. Because $b-c<b$, residents tend to imitate the mutant in the first case. Because $b-c>0$, mutants tend to imitate the resident in the second case. Cooperation is stochastically stable if the first case is less likely than the second, which yields the condition

$$q < 1-\frac{1}{2\delta}. \tag{4}$$

Interestingly, this condition no longer depends on the exact payoff values, b and c. This independence arises because of our assumption of strong selection, in which case only the payoff ordering $b > c > 0$ matters. Because q is non-negative, condition (4) can be satisfied only if $\delta > 1/2$. That is, players need to interact in more than two rounds, on average.

By comparing the two cases, we find that payoff memory affects whether a conditionally cooperative strategy $(1, 1, q)$ is viable. With perfect memory, the maximum generosity q must satisfy Eq. (3). In particular, this generosity can become arbitrarily large, provided the game's benefit-to-cost ratio b/c and the continuation probability δ are sufficiently large. In contrast, with limited payoff memory, the maximum generosity is bounded by one half, and it is independent of the benefit-to-cost ratio.

Evolutionary dynamics of reciprocity. The previous static observations describe whether full cooperation,

once established, can be sustained. In a next step, we explore in which cases cooperation actually evolves. To this end, we turn to simulations. We have run separate simulations for perfect and limited payoff memory. In each case, we consider both a low and a high benefit of cooperation ($b/c=3$ and $b/c=10$, respectively). For each simulation, we record which strategies (y,p,q) the players adopt over time. Figure 1 depicts the conditional cooperation probabilities p and q (we omit the opening move y because we use a discount factor δ close to one, such that first-round behavior is largely irrelevant). In all simulations, we find that the players' strategies cluster in two regions of the strategy space. The first region corresponds to a neighborhood of ALLD with $(p,q)\approx(0,0)$. The second region corresponds to a thin strip of cooperative strategies with $(p,q)\approx(1,q)$. Within this strip, we observe that most strategies satisfy the constraints on q suggested by the inequalities (3) and (4). That is, with perfect memory, most evolving strategies have $q<1-c/b$, whereas with limited payoff memory, most strategies have $q<1/2$. In particular, for limited payoff memory, changes in the benefit parameter have no effect on the qualitative distribution of strategies.

In each case, the evolutionary dynamics follow a similar cyclic pattern (as described in Refs. 17, 33): Resident populations of defectors are most likely invaded by strategies close to TFT. Once the population adopts conditionally cooperative strategies $(1,1,q)$, neutral drift may introduce larger values of generosity q. If the resident's generosity q violates the conditions (3) and (4), defectors can re-invade, and the cycle starts again. The relative time spent near ALLD and near the strip of conditionally cooperative strategies depends on the memory (Supplementary Figure 3, depicting the case of high benefits). For perfect memory, we find that ALLD is replaced relatively quickly by more cooperative strategies. Here, it takes on average 159 invading mutants until ALLD is successfully replaced. In contrast, for limited memory, ALLD is more robust, resisting on average 798 mutant strategies. This picture reverses for an initial population that adopts GTFT. Such populations are much more robust under perfect memory than they are under limited memory. Overall, we find that the impact of memory on the population's average cooperation rate is substantial. For perfect memory, this rate is 52% for low benefits, and 98% for high benefits. For limited payoff memory, the evolving cooperation rates are smaller but still strictly positive, with 37% cooperation for low benefits and 51% cooperation for high benefits (Figure 1).

To further investigate the influence of different parameters, we have systematically varied the benefit b and the selection strength β in Figure 2. According to Figure 2*a*, perfect memory consistently results in a higher cooperation rate, and this relative advantage further increases with an increasing benefit b. Interestingly, for limited payoff memory, the cooperation rate remains stable at approximately 50% once $b\geq 5$. This finding is reminiscent of our earlier static results, which also suggested that changes in the game's benefits may have a negligible role on the stochastic stability of full cooperation. With respect to the effect of different selection strengths, Figure 2*b* suggests that both perfect and limited payoff memory yield similar cooperation rates for weak selection ($\beta\ll 1$). This is not a coincidence: it is known that, under weak selection, stochastic payoffs can be replaced by their (deterministic) expectations without altering the evo-

lutionary dynamics (65)–and perfect payoff memory corresponds to the expected value of the payoffs in the limited payoff memory model, due to the law of large numbers. Beyond weak selection, increasing selection has a positive effect under perfect payoff memory, but a negative effect under limited payoff memory.

The effect of increasing individual payoff memory. So far, we have taken a rather extreme interpretation of limited payoff memory. We assumed that individuals update their strategies based on their experience in a single round of the prisoner's dilemma, against a single co-player. The limited payoff memory framework can be expanded in various ways. In particular, individuals may recall a larger number of rounds, they may recall their interactions with several co-players, or both. To gain further insights on the impact of payoff memory, we explore four additional scenarios. In the first scenario, players recall the payoffs they obtained in the last two rounds against a single co-player. In the second scenario, players recall their last-round payoffs against two co-players. In the third scenario, they recall the two last rounds against two co-players. Finally, in the last scenario, players update based on the average payoff they receive over all rounds with a single co-player. (Further extensions are possible, but we do not explore them here.)

For most scenarios, we can again derive an analytical condition for when cooperation is stochastically stable. As before, we assume populations are large and that selection is strong. For simplicity, we also assume that the game continues almost certainly after each round (i.e., δ approaches one). The details of this analysis can be found in the electronic supplementary material. In the first two scenarios, we interestingly find that for $b > 2c$, cooperation is stochastically stable when $q < \frac{\sqrt{2}}{2} \approx 0.707$. Comparing this condition with the more stringent condition in Eq. (4) suggests that there are now more conditionally cooperative strategies that can sustain cooperation. Hence, cooperation should evolve more easily. In the last scenario, we find that cooperation is stochastically stable when $q < 1 - \frac{c}{b}$, which is the same condition as in Eq. (3), even though only a single co-player is considered instead of the whole population.

We complement these analytical results with additional simulations; see Figure 3. We observe that a minimal increase in the players' payoff memory (compared to the baseline case with a single round recalled) can promote cooperation considerably. Specifically, in all four scenarios with extended memory, we see similar cooperation rates that approach the rates observed under perfect memory. These results suggest that while it takes *some* payoff memory to sustain substantial cooperation rates, the memory requirements are rather modest. Already remembering a few interactions, either with the same co-player or across different co-players, may provide players with enough information to adopt reciprocal strategies.

Beyond reactive strategies and the donation game. While the results presented in the main text focus on reactive strategies and the donation game, the observed patterns hold more generally. To illustrate this point in more detail, in the electronic supplementary material we first consider the dynamics of the donation game among memory-1 strategies. Here, players take into account both their co-player's and their own last move,

see Refs. (66, 67). Also in that case, we also observe that perfect memory leads to systematically higher cooperation rates (Supplementary Figures 5,6). Again, this advantage of perfect memory is particularly pronounced for strong selection, or when there is a high benefit of cooperation (Supplementary Figure 7).

In a second step, we also analyze the evolutionary dynamics of a different social dilemma, the snowdrift game (68). In this game, individuals can again cooperate or defect. However, now they get the benefit b if at least one of them cooperates. Cooperators in turn pay a cost of c or $c/2$, depending on whether they are the only one to cooperate or not. Compared to the donation game, the snowdrift game represents a weaker form of a social dilemma, because defection is no longer the dominant action in any single round (69). Again, we derive analytical results in the case of large populations and strong selection. For perfect memory, cooperation is stochastically stable when $q < 1-c/(2\delta b)$. Interestingly, for limited payoff memory, we obtain the same condition as for the donation game, $q < 1-1/(2\delta)$. We explain this invariance in detail in the electronic supplementary material: compared to the donation game, the ordering of one-shot payoffs in the snowdrift game only differs in one instance (the *sucker's payoff* now exceeds the *punishment payoff*); however, this instance is irrelevant for the stochastic stability of cooperation. Further evolutionary simulations suggest that again, perfect memory leads to higher cooperation rates than limited memory; but now even limited memory can result in substantial cooperation (Supplementary Figures 8, 9).

4 Discussion

In economics, if payoff is measured in terms of money and a decision is to be made at time t in the future, then currency accumulated early on is weighted more in that decision because it has more time to accumulate interest. Such a model discounts the future relative to the past. It also does not necessarily require "memory" because rewards are accumulated into a factor used in decision-making; the specific time stamps of rewards do not themselves provide better information beyond their effects on total payoff. In a similar fashion, the probabilistic interpretation of discounting as a continuation probability (15) also effectively discounts the future relative to the past. A foraging animal deciding between two behaviors might tend to choose the one that yields a moderate reward sooner relative to a larger reward later, since earlier rewards (e.g., food) contribute to immediate survival, and there is no guarantee that later rewards will happen at all (70).

Within the context of a single repeated game, the model we consider here is, in some ways, dual to the classical model of temporally-discounted rewards in repeated games. Instead of making decisions based on expected rewards in the future, we consider individuals who make decisions based on actual rewards in the past. Thus, the ability to estimate the future payoff of a strategy is replaced by the memory of how this strategy previously fared against others. This involves two time scales: interaction partners and rounds within those interactions. As a result, we are dealing with a model that discounts the past rather than the future. Intriguingly, treating payoffs in this manner is reminiscent of the reward-smoothing technique of "eligibility traces" in reinforcement learning (71), which uses past rewards (discounted appropriately) to shape present

perceived payoff. There is a sound basis for this method in neuroscience, where rewards and (temporal-difference) learning are associated to dopaminergic neurons (72) and spike-timing-dependent plasticity (73). This suggests a more biologically-encoded interpretation of memory, which is equally applicable to models of direct reciprocity where rewards have a neurological basis.

Of course, the precise nature of "memory" also depends on what payoffs in a game represent, which should be taken into account when applying game-theoretic models. For example, a payoff stream of monetary currency might truly accumulate and not require memory on the parts of agents. Even in the context of money, however, not all of what was obtained in the past is necessarily available at the time a decision is made, which brings memory into play. The serial position effect in human psychology shows that in an ordered list of items (e.g., words), humans tend to have difficulty remembering the entirety of sequences, demonstrating moderate recall for those items coming earlier (primacy effect), substantial recall for those coming later (recency effect), and lower recall for those in between (74). It is therefore reasonable that when presented with a stream of payoffs, whether on the timescale of pairing for repeated interactions or in a stream of one-shot games, players might be able to effectively incorporate only the most recent payoffs.

In fact, even beyond specific psychological considerations, a curious interpretation of payoffs arises from the formula commonly used for expected payoffs in repeated games. If $\delta \in [0,1)$ is the probability of continuing to another round in the game, then the expected payoff to an agent is $(1-\delta)\sum_{t=0}^{\infty}\delta^t u_t$, where u_t is the reward the agent receives in the stage game at time t. Here, additional stochasticity arises due to uncertainty in the game length, and an agent might not be able to compute his or her expected payoff for use in decision-making. The probability that the game terminates after the interaction at time T is $\delta^T(1-\delta)$, in which case $(1-\delta)\sum_{t=0}^{\infty}\delta^t u_t$ is exactly the expected payoff the agent receives at time T, i.e. *in the last round*. As an unbiased estimator of this expectation, the agent might thus use u_T as a proxy for "success" when evaluating his or her behavior. This gives a purely model-driven justification for why considering payoff in the last round of the game can result in more realistic extensions of traditional models.

We note that the expected *total* payoff in the game, $u_0+u_1+\cdots+u_T$, is given by $\sum_{t=0}^{\infty}\delta^t u_t$. This version of "expected payoff" appears less common in the literature on direct reciprocity than its normalization, $(1-\delta)\sum_{t=0}^{\infty}\delta^t u_t$, likely owing to the fact that payoffs can grow arbitrarily large with sufficiently long time horizons ($\delta \to 1^-$). Non-normalized payoffs interfere with selection intensity (β) in models of social imitation, which is (presumably) why they appear less frequently in the literature. On that point, we note that differences between realized and expected payoffs disappear in the limit of weak selection, which is known in a general setting (65). Non-weak selection can introduce substantial differences between models with realized and expected payoffs (75), which is especially important to understand in models of social systems with cultural transmission (76).

Our main contribution is an application of these ideas to direct reciprocity, which is one of the key mechanisms to explain why unrelated individuals might cooperate (7). According to this mechanism, cooperation

pays if it makes the interaction partner more cooperative in future. To describe which strategies are most effective, the previous theoretical literature assumes that the evolutionary dynamics are driven by the players' expected payoffs (29–44). To the extent that strategies are learned (not inherited), this assumption seems to impose rather stringent requirements on the individuals' cognitive abilities. In the most extreme case, individuals would have to remember (or compute) their payoffs against all population members, for all possible ways in which their repeated games may unfold. This assumption introduces a curious inconsistency in how these models represent an individual's cognitive abilities. For playing their games, individuals are often assumed to only recall the outcome of the very last round. Yet to update their strategies, individuals are implicitly assumed to have a record of the outcome of all rounds, across all interaction partners.

It is natural to ask, then, to what extent perfect payoff memory is in fact required for the evolution of reciprocity. To this end, we consider a model in which individuals only remember the payoff of their very last interaction, or the payoffs of the last few interactions. By only considering an individual's most recent experiences, the evolutionary process is subject to additional stochasticity. Strategies that perform well on average (across an entire repeated game and across many interaction partners) may still get replaced if the respective player happened to yield an inferior payoff in the very last round. A similar element of stochasticity has been previously explored in the context of one-shot (non-repeated) games (77–81). This literature studies which strategies are selected for when individuals only interact with a finite sample of population members. In the respective models, individuals can only choose among two strategies. They can either cooperate or defect, and stochastic sampling affects which of these two strategies is favored. Instead, in repeated games, players have access to a large set of strategies (in our case, all reactive strategies; 82). Here, stochastic sampling does not only affect whether cooperative or non-cooperative strategies are favored; it also affects *which* conditionally cooperative strategies are favored.

To explore the effect of payoff memory, we only considered the simplest model of reciprocity. This leaves several possible directions for future work. For example, for both our analytical analysis and our simulations, we assumed that populations are *well-mixed*. As a result, any population member is equally likely to be a given player's last co-player. A similar analysis for structured populations would be desirable, as population structure can often promote cooperation without an explicit need for reciprocity (48, 83). Similarly, we assumed that after identifying a suitable role model, players can easily imitate that role model's strategy. In the context of our study, this might be a particular simplification, because players with limited payoff memory may also have less accurate information to infer a co-player's strategy. Finally, we have taken a player's payoff memory to be a given parameter of the model. Instead, future models could explore how that payoff memory might co-evolve along with the players' strategies (similar to previous work on the evolution of the players' feasible strategy sets, e.g. 43, 49, 52). Here, individuals might experience an interesting trade-off. Better payoff-memory would help them to make better decisions, yet it might also come with higher cognitive costs. For such a setup, our results suggest that in many instances, a modest amount of memory

may be sufficient to achieve good outcomes.

5 Conclusion

Herein, we combine analytical methods and computer simulations to explore the impact of payoff memory on the evolution of reciprocal altruism. In the most extreme case, we consider individuals who update their strategies based on only one piece of information: the last round of a single repeated game. Compared to exisiting models where decisions are made based on expected payoffs (perfect payoff memory), we find that individuals are less generous, and they tend to be less cooperative overall (Figure 1). However, once individuals update their strategies based on two or more recent experiences, overall cooperation rates quickly approach the levels observed under perfect payoff memory (Figure 3). These findings suggest that models based on expected payoffs can serve as a useful approximation to more realistic models with limited payoff memory. Our findings also contribute to a wider literature that explores which kinds of cognitive capacities are required for reciprocal altruism to be feasible (e.g., 84, 85). While more payoff memory is always favorable, reciprocal cooperation can already be sustained if individuals have a record of two or three past outcomes. We believe that this kind of result, derived entirely within a theoretical model, is crucial for making model-informed deductions about reciprocity in natural systems.

Ethics. This work is purely theoretical and did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. All data and code used in this manuscript are openly available at `https://zenodo.org/records/10066227`.

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. N.G.: conceptualization, formal analysis, investigation, methodology, visualization, writing - original draft, writing – review and editing; A.M.: conceptualization, formal analysis, methodology, writing – review and editing; C.H.: conceptualization, formal analysis, methodology, writing – review and editing.

Conflict of interest declaration. We declare we have no competing interests.

Funding. N.G. and C.H. acknowledge generous support by the European Research Council Starting Grant 850529: E-DIRECT, and by the Max Planck Society.

References

[1] Hofbauer, J., Sigmund, K. *et al. Evolutionary games and population dynamics* (Cambridge university press, 1998).

[2] Nowak, M. A., Sasaki, A., Taylor, C. & Fudenberg, D. Emergence of cooperation and evolutionary stability in finite populations. *Nature* **428**, 646–650 (2004).

[3] Hauert, C. & Szabó, G. Game theory and physics. *American Journal of Physics* **73**, 405–414 (2005).

[4] Traulsen, A. & Glynatsi, N. E. The future of theoretical evolutionary game theory. *Philosophical Transactions B* (2022).

[5] Maynard Smith, J. & Price, G. R. The logic of animal conflict. *Nature* **246**, 15–18 (1973).

[6] Stein, A. *et al.* Stackelberg evolutionary game theory: how to manage evolving systems. *Philosophical Transactions of the Royal Society B* **378**, 20210495 (2023).

[7] Nowak, M. A. Five rules for the evolution of cooperation. *Science* **314**, 1560–1563 (2006).

[8] Smith, J. M. *Evolution and the Theory of Games* (Cambridge university press, 1982).

[9] Wu, B., Bauer, B., Galla, T. & Traulsen, A. Fitness-based models and pairwise comparison models of evolutionary games are typically different—even in unstructured populations. *New Journal of Physics* **17**, 023043 (2015).

[10] Smolla, M. *et al.* Underappreciated features of cultural evolution. *Philosophical Transactions of the Royal Society B* **376**, 20200259 (2021).

[11] Denton, K. K., Ram, Y. & Feldman, M. W. Conformity and content-biased cultural transmission in the evolution of altruism. *Theoretical Population Biology* **143**, 52–61 (2022).

[12] Chatterjee, K., Zufferey, D. & Nowak, M. A. Evolutionary game dynamics in populations with different learners. *Journal of Theoretical Biology* **301**, 161–173 (2012).

[13] Trivers, R. L. The evolution of reciprocal altruism. *The Quarterly review of biology* **46**, 35–57 (1971).

[14] Rapoport, A. & Chammah, A. M. *Prisoner's Dilemma* (University of Michigan Press, Ann Arbor, 1965).

[15] Axelrod, R. & Hamilton, W. D. The evolution of cooperation. *Science* **211**, 1390–1396 (1981).

[16] Molander, P. The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution* **29**, 611–618 (1985).

[17] Nowak, M. A. & Sigmund, K. Tit for tat in heterogeneous populations. *Nature* **355**, 250–253 (1992).

[18] Fischbacher, U., Gächter, S. & Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Economic Letters* **71**, 397–404 (2001).

[19] Rand, D. G. & Nowak, M. A. Human cooperation. *Trends in Cogn. Sciences* **117**, 413–425 (2012).

[20] Dal Bó, P. & Fréchette, G. R. Strategy choice in the infinitely repeated prisoner's dilemma. *American Economic Review* **109**, 3929–3952 (2019).

[21] Montero-Porras, E., Grujić, J., Fernández Domingos, E. & Lenaerts, T. Inferring strategies from observations in long iterated prisoner's dilemma experiments. *Scientific Reports* **12**, 7589 (2022).

[22] Rossetti, C. & Hilbe, C. Direct reciprocity among humans. *Ethology* https://doi.org/10.1111/eth.13407 (2023).

[23] Carter, G. G. & Wilkinson, G. S. Food sharing in vampire bats, reciprocal help predicts donations more than relatedness or harassment. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20122573 (2013).

[24] Schweinfurth, M. K., Aeschbacher, J., Santi, M. & Taborsky, M. Male norway rats cooperate according to direct but not generalized reciprocity rules. *Animal Behaviour* **152**, 93–101 (2019).

[25] Voelkl, B. *et al.* Matching times of leading and following suggest cooperation through direct reciprocity during V-formation flight in ibis. *Proceedings of the National Academy of Sciences USA* **112**, 2115–2120 (2015).

[26] Clutton-Brock, T. Cooperation between non-kin in animal societies. *Nature* **462**, 51–57 (2009).

[27] Silk, J. B. Reciprocal altruism. *Current Biology* **23**, 827–828 (2013).

[28] Taborsky, M. Social evolution: Reciprocity there is. *Current Biology* **23**, 486–488 (2013).

[29] Brauchli, K., Killingback, T. & Doebeli, M. Evolution of cooperation in spatially structured populations. *Journal of Theoretical Biology* **200**, 405–417 (1999).

[30] Brandt, H. & Sigmund, K. The good, the bad and the discriminator - errors in direct and indirect reciprocity. *Journal of Theoretical Biology* **239**, 183–194 (2006).

[31] Ohtsuki, H. & Nowak, M. A. Direct reciprocity on graphs. *Journal of Theoretical Biology* **247**, 462–470 (2007).

[32] Szolnoki, A., Perc, M. & Szabó, G. Phase diagrams for three-strategy evolutionary prisoner's dilemma games on regular graphs. *Physical Review E* **80**, 056104 (2009).

[33] Imhof, L. A. & Nowak, M. A. Stochastic evolutionary dynamics of direct reciprocity. *Proceedings of the Royal Society B: Biological Sciences* **277**, 463–468 (2010).

[34] van Segbroeck, S., Pacheco, J. M., Lenaerts, T. & Santos, F. C. Emergence of fairness in repeated group interactions. *Physical Review Letters* **108**, 158104 (2012).

[35] Grujic, J., Cuesta, J. A. & Sanchez, A. On the coexistence of cooperators, defectors and conditional cooperators in the multiplayer iterated prisoner's dilemma. *Journal of Theoretical Biology* **300**, 299–308 (2012).

[36] Martinez-Vaquero, L. A., Cuesta, J. A. & Sanchez, A. Generosity pays in the presence of direct reciprocity: A comprehensive study of $2\times$ 2 repeated games. *PLoS One* **7**, e35135 (2012).

[37] Stewart, A. J. & Plotkin, J. B. From extortion to generosity, evolution in the iterated prisoner's dilemma.

Proceedings of the National Academy of Sciences USA **110**, 15348–15353 (2013).

[38] Pinheiro, F. L., Vasconcelos, V. V., Santos, F. C. & Pacheco, J. M. Evolution of all-or-none strategies in repeated public goods dilemmas. *PLoS Comput Biol* **10**, e1003945 (2014).

[39] Stewart, A. J. & Plotkin, J. B. The evolvability of cooperation under local and non-local mutations. *Games* **6**, 231–250 (2015).

[40] Baek, S. K., Jeong, H.-C., Hilbe, C. & Nowak, M. A. Comparing reactive and memory-one strategies of direct reciprocity. *Scientific reports* **6**, 1–13 (2016).

[41] McAvoy, A. & Nowak, M. A. Reactive learning strategies for iterated games. *Proceedings of the Royal Society A* **475**, 20180819 (2019).

[42] Glynatsi, N. E. & Knight, V. A. Using a theory of mind to find best responses to memory-one strategies. *Scientific Reports* **10**, 17287 (2020).

[43] Schmid, L., Hilbe, C., Chatterjee, K. & Nowak, M. A. Direct reciprocity between individuals that use different strategy spaces. *PLoS Computational Biology* **18**, e1010149 (2022).

[44] Murase, Y., Hilbe, C. & Baek, S. K. Evolution of direct reciprocity in group-structured populations. *Scientific Reports* **12**, 18645 (2022).

[45] Cooney, D. B. Assortment and reciprocity mechanisms for promotion of cooperation in a model of multilevel selection. *Bulletin of Mathematical Biology* **84**, 126 (2022).

[46] Chen, X. & Fu, F. Outlearning extortioners: Unbending strategies can foster reciprocal fairness and cooperation. *PNAS Nexus* **2**, pgad176 (2023).

[47] Hauert, C. & Schuster, H. G. Effects of increasing the number of players and memory size in the iterated prisoner's dilemma: a numerical approach. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **264**, 513–519 (1997).

[48] van Veelen, M., García, J., Rand, D. G. & Nowak, M. A. Direct reciprocity in structured populations. *Proceedings of the National Academy of Sciences USA* **109**, 9929–9934 (2012).

[49] Stewart, A. J. & Plotkin, J. B. Small groups and long memories promote cooperation. *Scientific reports* **6**, 1–11 (2016).

[50] Hilbe, C., Martinez-Vaquero, L. A., Chatterjee, K. & Nowak, M. A. Memory-n strategies of direct reciprocity. *Proceedings of the National Academy of Sciences* **114**, 4715–4720 (2017).

[51] Li, J. *et al.* Evolution of cooperation through cumulative reciprocity. *Nature Computational Science* **2**, 677–686 (2022).

[52] Murase, Y. & Baek, S. K. Grouping promotes both partnership and rivalry with long memory in direct reciprocity. *PLoS Computational Biology* **19**, e1011228 (2023).

[53] Sigmund, K. *The calculus of selfishness* (Princeton University Press, 2010).

[54] Friedman, J. A non-cooperative equilibrium for supergames. *Review of Economic Studies* **38**, 1–12 (1971).

[55] Akin, E. The iterated prisoner's dilemma: Good strategies and their dynamics. In Assani, I. (ed.) *Ergodic Theory, Advances in Dynamics*, 77–107 (de Gruyter, Berlin, 2016).

[56] Hilbe, C., Traulsen, A. & Sigmund, K. Partners or rivals? Strategies for the iterated prisoner's dilemma. *Games and Economic Behavior* **92**, 41–52 (2015).

[57] Stewart, A. J. & Plotkin, J. B. Collapse of cooperation in evolving games. *Proceedings of the National Academy of Sciences USA* **111**, 17558 – 17563 (2014).

[58] 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).

[59] Blume, L. E. The statistical mechanics of best-response strategy revision. *Games and Economic Behavior* **11**, 111–145 (1995).

[60] Szabó, G. & Tőke, C. Evolutionary Prisoner's Dilemma game on a square lattice. *Physical Review E* **58**, 69–73 (1998).

[61] Fudenberg, D. & Imhof, L. A. Imitation processes with small mutations. *Journal of Economic Theory* **131**, 251–262 (2006).

[62] Wu, B., Gokhale, C. S., Wang, L. & Traulsen, A. How small are small mutation rates? *Journal of Mathematical Biology* **64**, 803–827 (2012).

[63] McAvoy, A. Comment on "Imitation processes with small mutations". *J. Econ. Theory* **159**, 66–69 (2015).

[64] Schmid, L., Chatterjee, K., Hilbe, C. & Nowak, M. A unified framework of direct and indirect reciprocity. *Nature Human Behaviour* **5**, 1292–1302 (2021).

[65] McAvoy, A., Allen, B. & Nowak, M. A. Social goods dilemmas in heterogeneous societies. *Nature Human Behaviour* **4**, 819–831 (2020).

[66] Nowak, M. A. & Sigmund, K. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* **364**, 56–58 (1993).

[67] Imhof, L. A., Fudenberg, D. & Nowak, M. A. Tit-for-tat or win-stay, lose-shift? *Journal of Theoretical Biology* **247**, 574–580 (2007).

[68] Doebeli, M. & Hauert, C. Models of cooperation based on the prisoner's dilemma and the snowdrift game. *Ecology letters* **8**, 748–766 (2005).

[69] Nowak, M. A. Evolving cooperation. *Journal of Theoretical Biology* **299**, 1–8 (2012).

[70] Stephens, D. W. & Krebs, J. R. *Foraging Theory*. Monographs in Behavior and Ecology (Princeton University Press, 1986).

[71] Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (The MIT Press, 2018), second edn.

[72] Schultz, W. Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology* **80**, 1–27 (1998).

[73] Dan, Y. & Poo, M.-M. Spike Timing-Dependent Plasticity of Neural Circuits. *Neuron* **44**, 23–30 (2004).

[74] Murdock, B. B. The serial position effect of free recall. *Journal of Experimental Psychology* **64**, 482–488 (1962).

[75] McAvoy, A., Rao, A. & Hauert, C. Intriguing effects of selection intensity on the evolution of prosocial behaviors. *PLOS Computational Biology* **17**, e1009611 (2021).

[76] Cavalli-Sforza, L. L. & Feldman, M. W. *Cultural Transmission and Evolution: A Quantitative Approach*. Cultural Transmission and Evolution: A Quantitative Approach (Princeton University Press, 1981).

[77] Sánchez, A. & Cuesta, J. A. Altruism may arise from individual selection. *Journal of theoretical biology* **235**, 233–240 (2005).

[78] Roca, C. P., Cuesta, J. A. & Sánchez, A. Time scales in evolutionary dynamics. *Physical review letters* **97**, 158701 (2006).

[79] Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic payoff evaluation increases the temperature of selection. *Journal of theoretical biology* **244**, 349–356 (2007).

[80] Woelfing, B. & Traulsen, A. Stochastic sampling of interaction partners versus deterministic payoff assignment. *Journal of Theoretical Biology* **257**, 689–695 (2009).

[81] Hauert, C. & Miekisz, J. Effects of sampling interaction partners and competitors in evolutionary games. *Physical Review E* **98**, 052301 (2018).

[82] Nowak, M. & Sigmund, K. Game-dynamical aspects of the prisoner's dilemma. *Applied Mathematics and Computation* **30**, 191–213 (1989).

[83] Szabó, G. & Fáth, G. Evolutionary games on graphs. *Physics Reports* **446**, 97–216 (2007).

[84] Stevens, J. R., Volstorf, J., Schooler, L. J. & Rieskamp, J. Forgetting constrains the emergence of cooperative decision strategies. *Frontiers in Psychology* **1**, 235 (2011).

[85] Volstorf, J., Rieskamp, J. & Stevens, J. R. The good, the bad, and the rare: Memory for partners in social interactions. *PloS one* **6**, e18945 (2011).

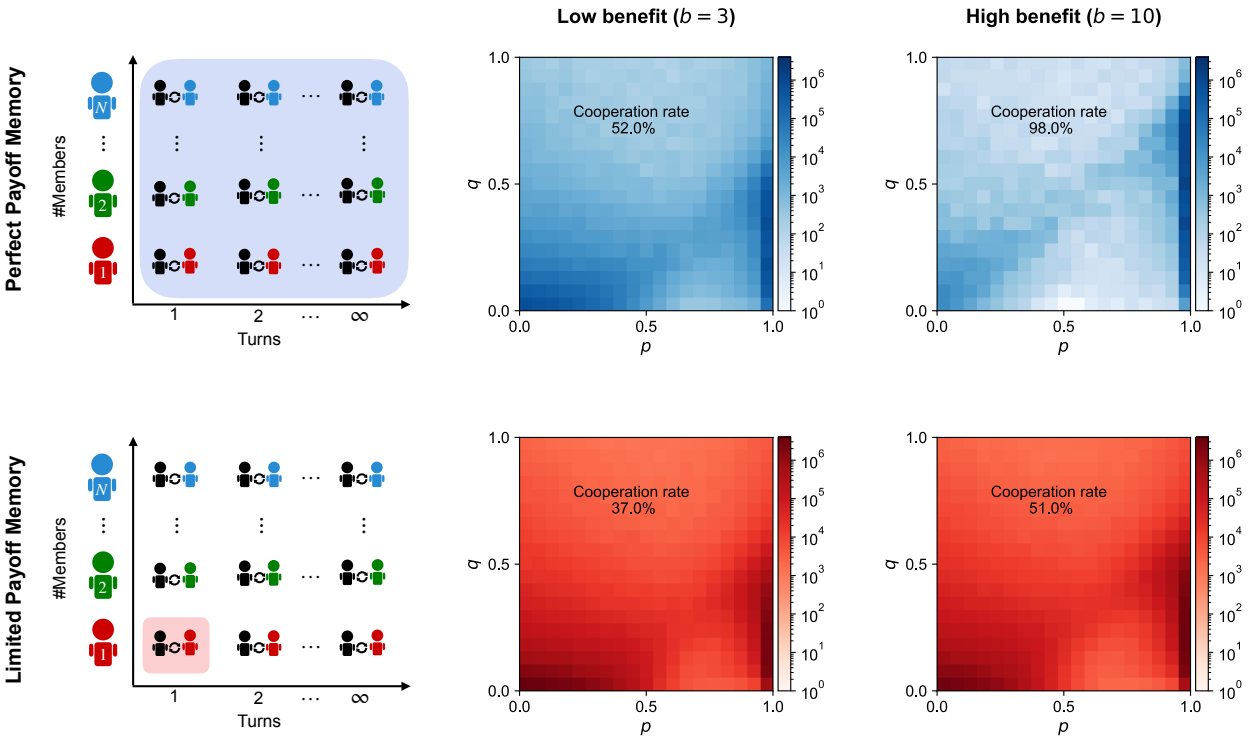

Figure 1: Evolutionary dynamics under perfect and limited payoff memory. The leftmost panels give a schematic overview of the two main scenarios we compare. The two scenarios differ in how many past interactions individuals take into account when updating their strategy. In the scenario with perfect payoff memory, individuals consider all their past interactions (against all population members, and taking every turn of each repeated game into account). In the scenario with limited payoff memory, individuals only consider their very last interaction (against one specific population member, taking into account only one round of the repeated game). The four panels on the right side depict the outcome of evolutionary simulations for repeated games with either a low or a high benefit of cooperation. Colors represent how often the respective region of the strategy space is visited over time. In all four panels, two regions are visited particularly often. One region corresponds to a neighborhood of ALLD with $p \approx q \approx 0$ (lower left corner). The other region corresponds to a strip of conditionally cooperative strategies with $p \approx 1$ and q satisfying the constraints (3) and (4), respectively (lower right corner). The resulting average cooperation rate depends on which of these two neighborhoods is visited more often. Simulations are run for $T = 10^7$ time steps, using a cost $c = 1$, a continuation probability of $\delta = 0.999$ and a selection strength of $\beta = 1$, in a population of size $N = 100$.

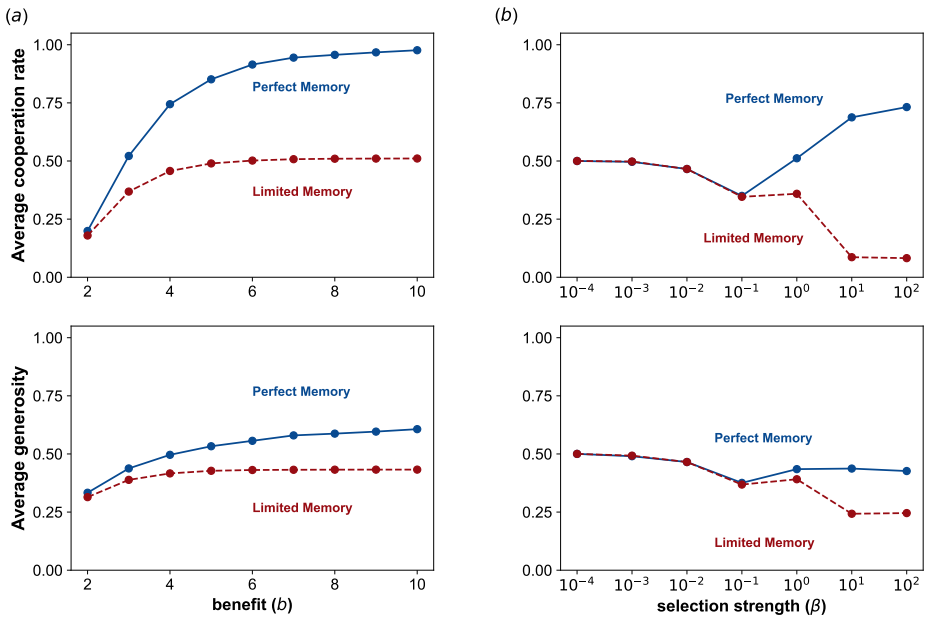

Figure 2: Evolution of direct reciprocity for different parameter values. To explore the robustness of our results, we have run simulations for different benefit values (left panels, a), and for different selection strengths (right panels, b). In each case, we record the resulting average cooperation rate over the entire simulation (upper panels). In addition we record the individuals' average generosity. Here, we only take into account those residents with $p \approx 1$ and we compute the average of their cooperation probability q. These simulations suggest that perfect payoff memory consistently leads to more cooperation and more generosity. Unless explicitly varied, the parameters of the simulation are $N = 100$, $b = 3$, $c = 1$, $\beta = 1$, $\delta = 0.99$. Simulations are run for $T = 5 \times 10^7$ time steps, and each point represents a single simulation run.

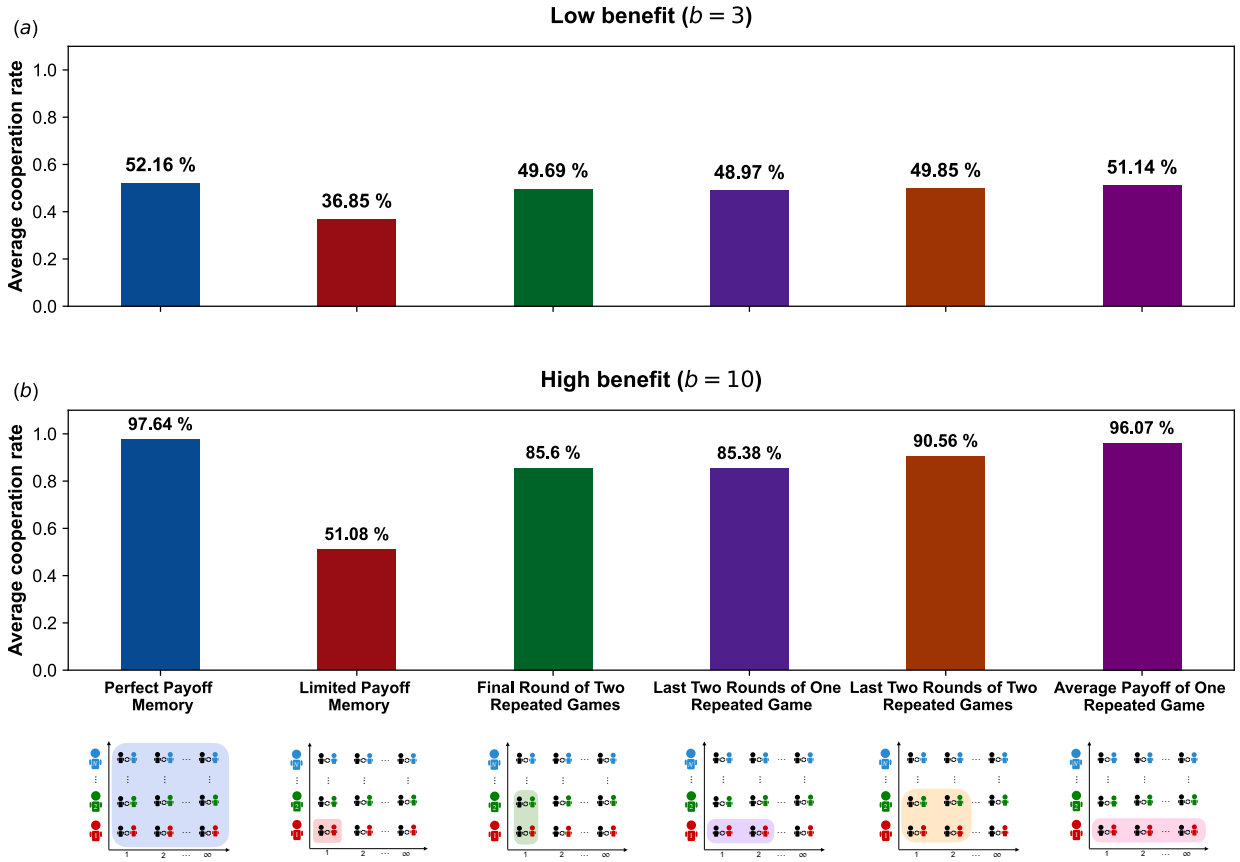

Figure 3: Average cooperation rates for different payoff memories. We vary how much information individuals take into account when updating their strategies. From left to right, we consider the following cases. (*i*) Updating occurs based on expected payoffs (perfect memory), (*ii*) it occurs based on 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 average payoff of one interaction. Again, simulations are run either for a comparably low benefit of cooperation ($b/c = 3$), or for a high benefit ($b/c = 10$). We observe that perfect memory always yields the highest cooperation rate. However, when individuals take into account at least two past interactions – cases (*iii*) to (*vi*) – evolving cooperation rates are close to this optimum. Simulations are run for $T = 10^7$ time steps, based on the parameters $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.