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

Abstract

Direct reciprocity can explain how cooperation emerges in repeated social interactions. According to this literature, individuals should naturally learn to adopt conditionally cooperative strategies such as Tit-for-Tat. Corresponding models have greatly facilitated our understanding of cooperation. Yet they often make strong assumptions on how individuals remember and process information. For example, when strategies are updated through social learning, it is commonly assumed that individuals update their strategies based on their average payoffs. This would require them to compute (or remember) their payoffs against all other population members. Instead, herein, we introduce a theoretical framework to study the evolution of reciprocity when individuals learn based on their most 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 tend to cooperate less often than in the classical framework with expected payoffs. Interestingly, once individuals remember the payoffs of two or three recent interactions, evolving 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 memory facilitates the evolution of reciprocity, it only takes a rather modest amount of payoff memory for cooperation to emerge.

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

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 (1–3). This theory can be used to describe the dynamics of animal conflict (4), cancer cells (5), and of cooperation (6). Respective models translate strategic interactions into games (7). These games specify how individuals (players) interact, which strategies they can choose, and what fitness consequences (or payoffs) their strategies have. In addition, these models also specify the mode by which successful strategies spread over time. Models of biological evolution posit that individuals with a high fitness tend to produce more offspring; models of cultural evolution instead assume that such individuals are imitated more often. While biological and cultural evolution are sometimes treated as equivalent, there can be important differences (8; 9). 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 that are available, and they need to identify those strategies with a higher payoff. As a consequence, evolutionary outcomes may often depend on how easily different behaviors can be learned (10), and on how easy payoff comparisons are.

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 (11). In repeated interactions, individuals can condition their future behavior on their past experiences with their interaction partner. They may use strategies such as Tit-for-Tat (12) or Generous Tit-for-Tat (13; 14) to preferentially cooperate with those partners who cooperated in the past. Such conditional strategies approximate human behavior fairly well (15–18) and they have also been documented in several other species (19–21) – although direct reciprocity is generally more difficult to demonstrate in animals (22–24). However, at the outset, it is not clear how easy it is to *learn* such strategies by social imitation. As one obstacle, individuals usually only observe other population members' actions (i.e., whether they cooperate or defect). These observations alone may not suffice to infer the underlying strategies (i.e., the contingent rules that determine whether to cooperate or defect in a given situation). As another obstacle, even if others' strategies are 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 both difficult to obtain and difficult to process.

Most models of direct reciprocity abstract from these difficulties (25–40). Instead they assume individuals can easily copy the strategies of others. Similarly, they assume that updating decisions are based on the strategies' average (or expected) payoffs, taking into account all rounds and 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. 41–45). 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 only based on payoff information from a limited number of rounds.

To explore that question, we first consider two extreme scenarios. The first scenario corresponds to the usual modeling approach. Here, individuals update their strategies for a repeated prisoner's dilemma based on pairwise comparisons (46), based on their strategies' expected payoff. We contrast this model with an alternative scenario where individuals update their strategies based on the very last (one-shot) payoff they obtained. We observe 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 case. Overall, these findings suggest that while memory is important, already minimal payoff information may suffice for the evolution of direct reciprocity based on social learning. They also suggest that the classical model of reciprocity (based on expected payoffs) can be interpreted as a useful approximation to more realistic models that provide a more realistic depiction of cognitive constraints.

2 Model Setup

A pairwise comparison process (46) starts with assigning all individuals of the population the same strategy. Thenceforth each elementary time step of the process consists of the mutation phase, the game phase and the update phase. In the game phase individuals are matched in pairs and that they participate in a repeated 2 person donation game; a special case of the prisoner's dilemma. In the donation game there are two actions: cooperation (C) and defection (D). By cooperating a player provides a benefit b to the other player at their cost c, with $0 < c < b$. Thus the payoffs for a player in each turn are given by,

$$\begin{array}{rc} & \begin{array}{cc}\text{cooperate} & \text{defect}\end{array} \\ \begin{array}{r}\text{cooperate} \\ \text{defect}\end{array} & \left(\begin{array}{cc} b-c & -c \\ b & 0 \end{array}\right). \end{array} \tag{1}$$

Let $\mathbf{u} = (b-c, -c, b, 0)$ be the round payoffs in a vector format, and let $\mathcal{U} = \{r, s, t, p\}$ denote the set of feasible payoffs, where r denotes the payoff of mutual cooperation, s the sucker's payoff, t the temptation to defect payoff, and p the punishment payoff.

In repeated games there are infinite many strategies, however, similar to the literature we will assume that individuals use reactive strategies. A reactive strategy considers only the previous action of the other player, and thus, a reactive strategy s can be written as a three-dimensional vector $s = (y, p, q)$. The parameter y is the probability that the strategy opens with a cooperation and p, q are the probabilities that the strategy cooperates given that the opponent cooperated and defected equivalently. The play between a pair of reactive strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$ can be model as a Markov process with the transition matrix M (47),

$$M = \begin{bmatrix} p_1 p_2 & p_1 (1 - p_2) & p_2 (1 - p_1) & (1 - p_1)(1 - p_2) \\ p_2 q_1 & q_1 (1 - p_2) & p_2 (1 - q_1) & (1 - p_2)(1 - q_1) \\ p_1 q_2 & p_1 (1 - q_2) & q_2 (1 - p_1) & (1 - p_1)(1 - q_2) \\ q_1 q_2 & q_1 (1 - q_2) & q_2 (1 - q_1) & (1 - q_1)(1 - q_2) \end{bmatrix} \tag{2}$$

and the stationary vector $\mathbf{v}(s_1, s_2)$ which is the solution to $\mathbf{v}(s_1, s_2) \times M = \mathbf{v}(s_1, s_2)$.

In the update stage two individuals are randomly selected. From the two individuals, one serves as the ‘learner’ and the other as the ‘role model’. The learner adopts the role model’s strategy with a probability ρ given by,

$$\rho(\pi_L, \pi_{RM}) = \frac{1}{1+e^{-\beta(\pi_{RM}-\pi_L)}}. \tag{3}$$

π_L and π_{RM} are the updating payoffs/fitness of the learner and the role model respectively. The updating payoffs are a measure of how successful individuals are in the current standing of the population. The parameter β is known as the selection strength, namely, it shows how important the payoff difference is when the learner is considering adopting the strategy of the role model.

For the results presented here we assume that mutations are rare ($\mu \rightarrow 0$). In fact, so rare that only two different strategies can be present in the population at any given time. However, in the Supplementary Information Section 9 we show that the main result holds for $\mu \neq 0$. The case of low mutation is vastly adopted because it allows us to explicitly calculate the fixation probability of a newly introduced mutant. More specifically, at each step one individual adopts a mutant strategy randomly selected from the set of feasible strategies. The fixation probability ϕ_M of the mutant strategy can be calculated explicitly as,

$$\varphi_M = \frac{1}{1 + \sum\limits_{i=1}^{N-1} \prod\limits_{k}^{i} \frac{\lambda_k^-}{\lambda_k^+}}, \tag{4}$$

where k is the number of mutants and λ_k^-, λ_k^+ are the probabilities that the number of mutants decreases and increases respectively. Depending on the fixation probability ϕ_M the mutant either fixes (becomes the new resident) or goes extinct. Regardless, in the next elementary time step another mutant strategy is introduced to the population. We iterate this elementary population updating process for a large number of mutant strategies and we record the resident strategies at each time step. The probabilities λ_k^- and λ_k^+ depend on the fitness of the mutant and the resident strategies. In the next section we present how fitness is calculated in the cases of perfect and limited payoff memory.

2.1 Fitness based on Perfect and Limited Payoff Memory

In the perfect payoff memory case an individual updates based on the average payoff against each other member of the population, otherwise known as expected payoffs. The payoff of a reactive strategy s_1 against the reactive strategy s_2 in an infinitely repeated game ($\delta \to 1$) is explicitly calculated as,

$$\langle \mathbf{v}(s_1, s_2), \mathbf{u} \rangle.$$

In a population of size N there are k mutants and $N - k$ residents. Let $s_M = (y_M, p_M, q_M)$ and $s_R = (y_R, p_R, q_R)$ denote the strategies of a mutant and a resident, the expected payoffs π_R and π_M are given by,

$$\begin{aligned} \pi_R &= \frac{N-k-1}{N-1} \cdot \langle \mathbf{v}(s_R, s_R), \mathbf{u} \rangle + \frac{k}{N-1} \cdot \langle \mathbf{v}(s_R, s_M), \mathbf{u} \rangle, \\ \pi_M &= \frac{N-k}{N-1} \cdot \langle \mathbf{v}(s_M, s_R), \mathbf{u} \rangle + \frac{k-1}{N-1} \cdot \langle \mathbf{v}(s_M, s_M), \mathbf{u} \rangle. \end{aligned} \tag{5}$$

The probabilities that the number of mutants decreases and increases, λ_k^- and λ_k^+, in the perfect payoff memory case are defined as,

$$\lambda_k^- = \rho(\pi_M, \pi_R) \quad \text{and} \quad \lambda_k^+ = \rho(\pi_R, \pi_M). \tag{6}$$

In the case of limited payoff memory we initially define the probability that a reactive strategy receives the payoff $u \in \mathcal{U}$ in the very last round of the game. This is given by Proposition 1 (see Supplementary Information Section 2.2.1 for proof).

Proposition 1. *Consider a repeated game, with continuation probability δ, between players with reactive strategies $s_1 = (y_1, p_1, q_1)$ and $s_2 = (y_2, p_2, q_2)$ respectively. Then the probability that the s_1 player receives the payoff $u \in \mathcal{U}$ in the very last round of the game is given by $v_u(s_1, s_2)$, as given by Equation (7).*

$$
\begin{aligned}
v_r(s_1,s_2) &= (1-\delta)\frac{y_1 y_2}{1-\delta^2 l_1 l_2} + \delta\frac{\Big(q_1 + l_1\big((1-\delta)y_2 + \delta q_2\big)\Big)\Big(q_2 + l_2\big((1-\delta)y_1 + \delta q_1\big)\Big)}{(1-\delta l_1 l_2)(1-\delta^2 l_1 l_2)},\\
v_s(s_1,s_2) &= (1-\delta)\frac{y_1 \bar{y}_2}{1-\delta^2 l_1 l_2} + \delta\frac{\Big(q_1 + l_1\big((1-\delta)y_2 + \delta q_2\big)\Big)\Big(\bar{q}_2 + \bar{r}_2\big((1-\delta)y_1 + \delta p_1\big)\Big)}{(1-\delta l_1 l_2)(1-\delta^2 l_1 l_2)},\\
v_t(s_1,s_2) &= (1-\delta)\frac{\bar{y}_1 y_2}{1-\delta^2 l_1 l_2} + \delta\frac{\Big(\bar{q}_1 + \bar{r}_1\big((1-\delta)y_2 + \delta p_2\big)\Big)\Big(q_2 + l_2\big((1-\delta)y_1 + \delta q_1\big)\Big)}{(1-\delta l_1 l_2)(1-\delta^2 l_1 l_2)},\\
v_p(s_1,s_2) &= (1-\delta)\frac{\bar{y}_1 \bar{y}_2}{1-\delta^2 l_1 l_2} + \delta\frac{\Big(\bar{q}_1 + \bar{r}_1\big((1-\delta)y_2 + \delta p_2\big)\Big)\Big(\bar{q}_2 + \bar{r}_2\big((1-\delta)y_1 + \delta p_1\big)\Big)}{(1-\delta l_1 l_2)(1-\delta^2 l_1 l_2)}.
\end{aligned}
\tag{7}
$$

In these expressions, we have used the notation $l_i := p_i - q_i$, $\bar{y}_i = 1 - y_i$, $\bar{q}_i := 1 - q_i$, *and* $\bar{l}_i := \bar{p}_i - \bar{q}_i = -l_i$ *for* $i \in \{1,2\}$.

In the case of limited payoffs memory both the role model and the learner estimate their fitness after interacting with a single member of the population. At each time step there are five possible pairings. They interact with each other with a probability $\frac{1}{N-1}$, and they do not interact with other with a probability $1 - \frac{1}{N-1}$. In the latter case, each of them can interact with either a mutant or a resident. Both of them interact with a mutant with a probability $\frac{(k-1)(k-2)}{(N-2)(N-3)}$ and both interact with a resident with a probability $\frac{(N-k-1)(N-k-2)}{(N-2)(N-3)}$. The last two possible pairings are that either of them interacts with a resident whilst the other interacts with a mutant, and this happens with a probability $\frac{(N-k-1)(k-1)}{(N-2)(N-3)}$. We define the probability that the randomly chosen resident obtained a payoff of u_R in the last round of his respective game, and that the mutant obtained a payoff of u_M as $x(u_R, u_M)$.

$$
\begin{aligned}
x(u_R,u_M) = \quad & \frac{1}{N-1}\cdot v_{u_R}(s_R,s_M)\cdot 1_{(u_R,u_M)\in\mathcal{U}_F^2}\\
&+\left(1-\frac{1}{N-1}\right)\Big[\tfrac{k-1}{N-2}\tfrac{k-2}{N-3}v_{u_R}(s_R,s_M)v_{u_M}(s_M,s_M) + \tfrac{k-1}{N-2}\tfrac{N-k-1}{N-3}v_{u_R}(s_R,s_M)v_{u_M}(s_M,s_R)\\
&\qquad + \tfrac{N-k-1}{N-2}\tfrac{k-1}{N-3}v_{u_R}(s_R,s_R)v_{u_M}(s_M,s_M) + \tfrac{N-k-1}{N-2}\tfrac{N-k-2}{N-3}v_{u_R}(s_R,s_R)v_{u_M}(s_M,s_R)\Big].
\end{aligned}
\tag{8}
$$

The probability that the number of mutants increases and decreases by one in the case of limited payoff memory are now given by,

$$\lambda_k^+ = \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_R, u_M \in \mathcal{U}} x(u_R, u_M)\rho(u_R, u_M) \quad \text{and} \quad \lambda_k^- = \frac{N-k}{N} \cdot \frac{k}{N} \cdot \sum_{u_R, u_M \in \mathcal{U}} x(u_R, u_M)\rho(u_M, u_R). \tag{9}$$

In this expression, $\frac{(N-k)}{N}$ is the probability that the randomly chosen learner is a resident, and $\frac{k}{N}$ is the probability that the role model is a mutant. The sum corresponds to the total probability that the learner adopts the role model's strategy over all possible payoffs u_R and u_M that the two player may have received in their respective last rounds.

3 Simulation Results

To assess the impact of updating payoffs, we simulate the evolutionary process, recording the strategies adopted by players over time based on perfect and limited payoff memory. We performed two separate runs for each approach varying the value of benefit b. Figure 1 depicts the evolving conditional cooperation probabilities p and q (note that we omit the opening move y, as the discount factor δ is relatively high). The figure suggests that when updating is based on perfect payoff memory players tend to be more generous and more cooperative.

Specifically, we observe that the resident population comprises either defectors or conditional cooperators $(1, q)$. The generosity level q adopted by the resident population depends on whether a defecting strategy can invade. Supplementary Information Section 2 and 3 show that in the perfect payoff memory framework, conditional cooperators of the form $(1, q < \frac{c}{b})$ can arise, however, in the case of limited payoff memory, only conditional cooperators of $(1, q < \frac{1}{2})$ can avoid invasion. The q-values of the resident strategies are generally higher in classical case, indicating that players are more likely to forgive a defection if their fitness depends on interacting with every member of the population. This effect becomes more pronounced as the benefit value increases, as the perfect memory condition on the left-hand side increases, while in the limited memory case, it remains at $q \approx \frac{1}{2}$.

Higher q values lead to a more cooperative population. We compute the average cooperation rate for each simulation, which is the average cooperation rate within the resident population. In the case of perfect payoff memory, the average cooperation rate is consistently higher than that of the last round payoffs.

We further investigate the impact of benefit and selection strength on generosity q and the cooperation rate, as shown in Figure 2. According to Figure 2**A**, perfect memory consistently results in a higher cooperation rate, which increases with increasing benefit. On the other hand, the cooperation rate remains approximately 50% for limited payoff memory once $b = 5$. From Figure 2**B** we observe that for weak selection, $\beta < 1$, the two methods yield similar results, however, as β increases there is variation in the evolving populations. In the case of expected payoffs the resident populations become more cooperative, whereas in the case of limited payoff memory, the resident populations become more defective.

The limited payoff memory framework can be expanded by enabling individuals to observe a greater number of rounds, interact with a larger number of members, or both (refer to Supplementary Information Sections 4-6). To gain further insight into the impact of limited payoff memory, we explore the scenarios of updating based on the last round with two members of the population, the last two rounds with another member of the population, and the last two rounds with two members of the population. To analyze the effects of this framework, we conduct numerical simulations using various fitness methods. The cooperation rates for low and high benefits are presented in Figure 3.

We observe a significant rise in the cooperation rate with the introduction of slightly more information. Specifically, in scenarios involving two rounds or two interactions, the cooperation rates are almost identical. For a large population and a high continuation probability (δ), the conditional cooperators that are adopted by the resident population in these scenarios take on the same form $(1, q < \frac{\sqrt{2}}{2})$. When considering both sets of information (i.e., two rounds and two interactions), cooperation rates experience a greater increase, yet still remain lower compared to the classical scenario.

4 Conclusions

Cooperation can be seen as odd, why is it that we choose to help others at a personal cost? In spite of all the selfish genes', animal and human communities show signs of altruism and cooperation (48–50). Evolutionary game theoretical models have helped us shape our understanding of the evolution of cooperation. In fact, the evolution of cooperation constitutes such a major focus of the field that evolutionary game theory seems to be reduced to the evolution of cooperation (51).

Evolutionary models in the past often feature a curious inconsistency. While these models depict how individuals make decisions in each round by assuming that they only retain memory of the previous round, they also assume that individuals possess a perfect memory when it comes to updating their strategies over time. To be precise, individuals are assumed to remember all of their past interactions and each interaction's outcome when updating strategies.

Here, we investigate the robustness of cooperation as models deviate from the assumption of perfect memory. While prior research has investigated the impact of constraining individuals' interactions, we take into account the limitation of not only interactions but also the information available for each outcome. Additionally, prior studies have only allowed for the adoption of simple strategies such as always cooperating or always defecting. In contrast, we enable the use of more intricate strategies where players can utilize the previous play of their co-player to make decisions.

In our framework, players update their strategies based on a combination of interactions and outcomes. The initial scenario we examined involved using one piece of information: the last round of one interaction.

The outcomes suggest that cooperation faces difficulties in developing when the updating stage utilizes minimal social information. This effect is compounded as the benefit and strength of selection are independently increased. The findings indicate that cooperative players benefit from the ability to engage with all members of the population.

Furthermore, we investigated scenarios where the final two rounds, or the last two instances of interaction, were taken into account. We observed a statistically significant rise in the frequency of cooperative behavior. For a sizable population and a high likelihood of continued interactions, the two cases yield the same result as the overall payoff is influenced by two possible outcomes. Notably, the scenario involving two rounds and two interactions yielded the highest cooperation rate among all the novel methodologies that we tested.

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

[5] Staňková, K., Brown, J. S., Dalton, W. D. & Gatenby, R. A. Optimizing Cancer Treatment Using Game Theory. *JAMA Oncology* **5**, 96–103 (2019).

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

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

[8] 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).

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

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

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

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

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

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

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

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

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

[18] Rossetti, C. & Hilbe, C. Direct reciprocity among humans. *Ethology* in press (2023).

[19] 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).

[20] 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).

[21] 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).

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

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

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

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

[26] 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).

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

[28] 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).

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

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

[31] 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).

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

[33] 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).

[34] 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).

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

[36] 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).

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

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

[39] Schmid, L., Hilbe, C., Chatterjee, K. & Nowak, M. A. Direct reciprocity between individuals that use different

strategy spaces. *PLoS Computational Biology* **18**, e1010149 (2022).

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

[41] 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).

[42] 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).

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

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

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

[46] 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).

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

[48] Milinski, M. Tit for tat in sticklebacks and the evolution of cooperation. *Nature* **325**, 433–435 (1987).

[49] Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature* **418**, 171–174 (2002).

[50] Carter, G. G. *et al.* Development of new food-sharing relationships in vampire bats. *Current Biology* **30**, 1275–1279 (2020).

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

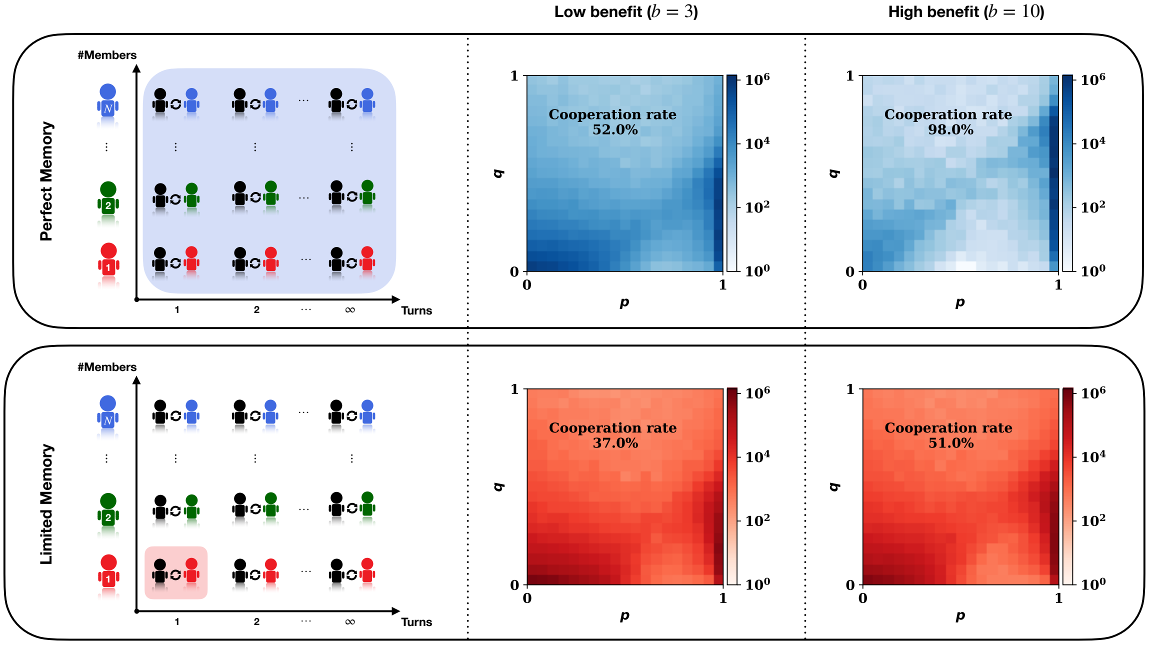

Figure 1: Evolutionary dynamics under perfect and limited payoff memory. (**Schematic illustrations**) On the left panels we show schematic illustrations of the perfect memory and the limited memory cases. The shaded background denotes the game phase information that an individual considers when updating strategies. In the case of perfect payoff memory the entire region is shaded and in the case of limited payoff memory only one turn with a single member of the population. (**Simulations**) We have run four simulations of the pairwise comparison process for $T = 10^7$ time steps. For each time step of the process we record the current resident population (y, p, q). Since simulations are run for a relatively high continuation probability of $\delta = 0.999$, we do not report the players' initial cooperation probability y. The plots show how often the resident population chooses each combination (p, q) of conditional cooperation probabilities in the subsequent rounds. We also report the evolved cooperation rate which is calculated as the average cooperation rate within the resident population. (**Perfect Memory**) In the case of low benefit the resident population either consists of defectors (with $p \approx q \approx 0$) or of conditional cooperators. Conditional cooperators, or otherwise known as generous tit for tat, are a set of strategies that always cooperate following a cooperation ($p \approx 1$) and cooperate with a probability q given that the co-player has defected. q denotes the generosity of a player. The resident population applies a conditional cooperator strategy for which $q \leq 1 - c/b = 0.67$. In the case of high benefit the population mainly consists of conditional cooperators of the form $(p \approx 1, q \leq 1 - 1/10 = 0.9)$. In the Supplementary Information Section 2 we show that a conditional cooperator needs to be of the form $(p \approx 1, q \leq 1 - c/b)$ to not be invaded by defecting strategies. A higher generosity in the population results in a higher average cooperation rate. The average cooperation rate increases from 52% for $b = 3$ to 98% for $b = 10$. (**Limited Memory**) When players update their strategies based on their realized payoffs in the last round, there are two different predominant behaviors regardless of the benefit value. The resident population either consists of defectors (with $p \approx q \approx 0$) or of conditional cooperators. The maximum level of q, consistent with stable cooperation, is somewhat smaller compared to the perfect memory setting. Namely, in the Supplementary Information Section 3 we show that regardless of the value of benefit a conditional cooperator need to be of the form $q < \frac{1}{2}$ to not be invaded by defectors. The evolved cooperation rate only slightly increases from 37% ($b = 3$) to 51% ($b = 10$). Parameters: $N = 100$, $c = 1$, $\beta = 1$, $\delta = 0.999$.

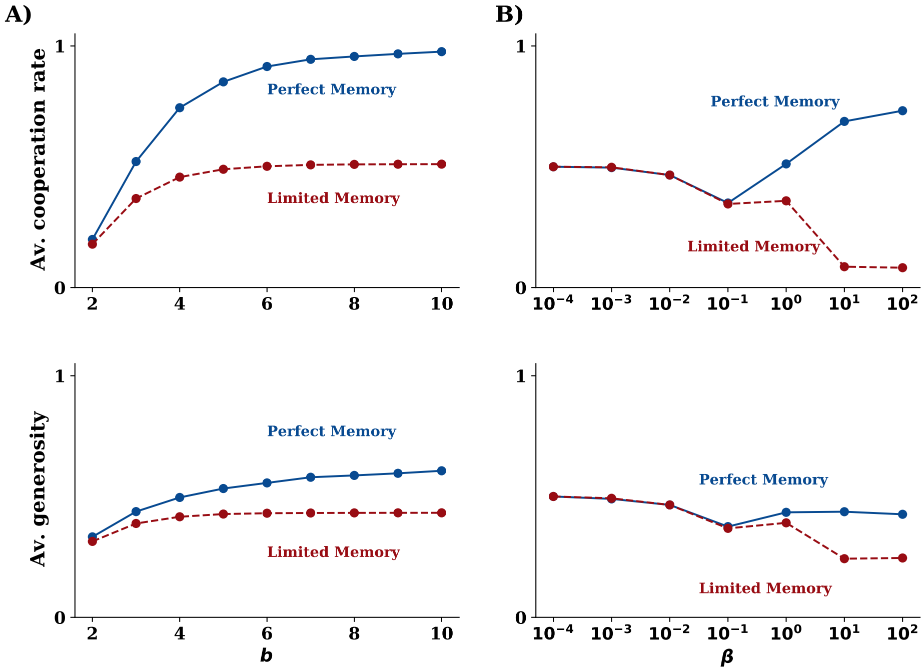

Figure 2: The evolution of cooperation and generosity for different values of benefit (A) and strength of selection (B). We report the average cooperation and the average reciprocity. The average cooperation rate is the average cooperation rate within the resident population. For the average reciprocity we select the residents that have a $p \approx 1$ and we take the average of their cooperation probability q. (**A**) We vary the benefit of cooperation b. In all cases, perfect memory updating payoffs appear to overestimate the average cooperation rate the population achieves. As expected in the case of limited memory the average generosity over the different values of benefit remains the same ($q \approx 0.5$), and as a result so does the average cooperation. (**B**) We vary the selection strength β. For weak selection, $\beta < 1$, the two methods yield similar results. However, as β increases in the case of limited memory payoffs the resident populations become more defective. Note that in the case of perfect payoff memory we see an increase in the cooperation rate even though the generosity remains stable. That is because the generosity does remain the same, however, now cooperative strategies remain fixed as the resident strategy for longer. 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 for each parameter combination.

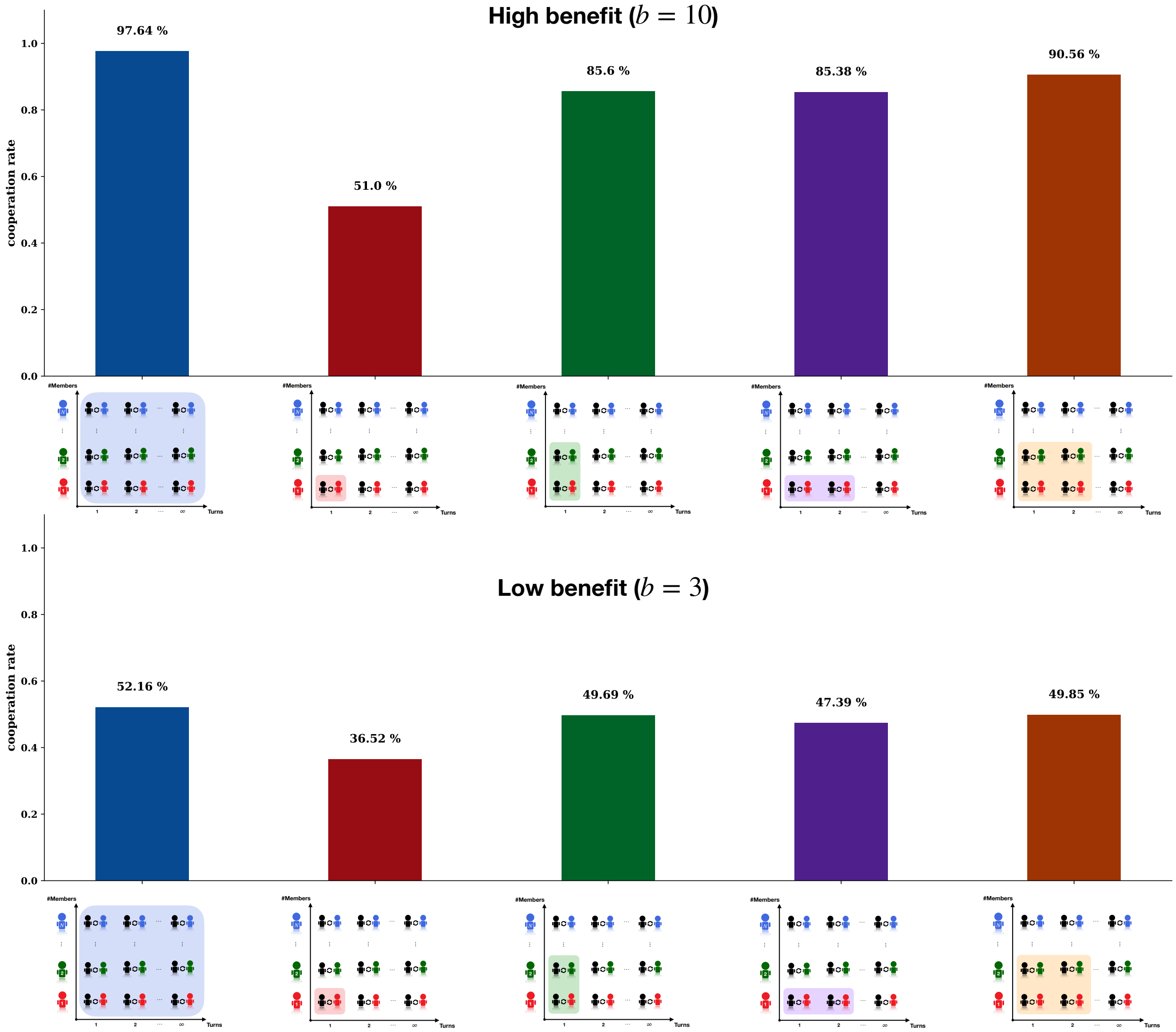

Figure 3: Average cooperation rates for different updating payoffs. From left to right, we present result on the following updating payoffs cases; (a) the expected payoffs (perfect memory), (b) the last round payoff from one interaction (limited memory), (c) the last round payoff from two interactions, (d) the last two rounds payoffs from one interaction, (e) the last two rounds payoffs from two interactions. For the updating payoffs (c) and (d) we have carried out an analysis which has shown that conditional cooperators should adopt a q value smaller than $\frac{\delta-1+\frac{\sqrt{2}}{2}}{\delta}$ and $\frac{\delta+\sqrt{\delta^2+1}-1}{2\delta}$ respectively. Note that as $\delta \to 1$ both right hand sides tend to $\frac{\sqrt{2}}{2}$. Regardless the cooperation rate between for case (c) is slightly higher. In case (d) the cooperation rate is the second highest hinting that as we allow for more information the closer we move to the perfect payoff memory. We performed four pairwise non parametric tests (Mann-Whitney U) to compare the cooperation distributions of the residents in case (b) to cases (a), (c), (d), (e). In all four tests we reject the null hypothesis with $\alpha = 0.05$ and $p \approx 0$. Thus, there is significant difference between the cooperation rates. 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 for each parameter combination.