Evolution of cooperation in stochastic games

Christian Hilbe[1,2], Stepan Simsa[3], Krishnendu Chatterjee[2], Martin A. Nowak[1,4]
[1]Program for Evolutionary Dynamics, Harvard University, Cambridge MA 02138, USA
[2]IST Austria, 3400 Klosterneuburg, Austria
[3]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[4]Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge MA 02138, USA

Social dilemmas occur when individual incentives are misaligned with group interests[1–8]. According to the tragedy of the commons these misalignments can lead to overexploitation and collapse of public resources. The theory of direct reciprocity[9–19] suggests that repeated interactions can alleviate such dilemmas, but previous work has assumed that the public good remains constant in every round. Here we introduce the idea that the public resource itself is changeable and depends on the outcome of previous interactions. A natural setting is that cooperation increases the public resource from one round to the next, while defection decreases it. Thus, cooperation allows the possibility to play a more valuable game, while defection leads to a less valuable game in the next round. The analysis of this idea requires the theory of stochastic games[20–23] and its introduction into evolutionary game theory. Surprisingly, we find that the dependency of the public resource on previous interactions can greatly enhance the propensity for cooperation. For these results, the interaction of reciprocity and payoff feedback proves crucial: neither repeated interactions in a constant environment, nor one-shot interactions in a changing environment can yield similar cooperation rates. Our framework shows which feedbacks between exploitation and environment - either naturally occurring or designed - help to overcome social dilemmas.

The tragedy of the commons leads to the question of how to manage and conserve public resources[1–8]. Any solution to this problem requires an understanding of which processes drive human cooperation, and how institutions, norms, and other feedback mechanisms can be employed to reinforce positive behaviors[24]. These questions are often explored by analyzing stylized social dilemmas, like the public goods game[25] or the collective-risk dilemma[26], that provide valuable insights into the dynamics of cooperation in controlled settings. When subjects interact in such games over multiple rounds, it is typically assumed that the public good remains constant in time, and that its value in every round is independent of the outcome of previous interactions[9–19]. In particular, it is assumed that the same game, with the same payoff values, is played in each round. Here, we explore the emergence of reciprocity when strategic choices in one round affect the game payoffs in subsequent rounds. We introduce a quantitative framework that allows us to capture that humans affect and in turn are affected by the value of the public resource, and that they are able to anticipate and to adapt to such endogenous changes.

Our approach is based on the theory of stochastic games[20,21]. A group of players can find itself in one of multiple states (see Fig. 1). The different states capture how the present physical or social environment affects the players' feasible actions and their payoffs. Stochastic game theory has a rich literature with a wealth of deep results[20–23]. It has applications in computer science[27,28] and in models of industrial organization, capital accumulation and resource extraction.[21]

We consider stochastic games where in each state, players interact in a social dilemma with different payoff values. The players' decision whether to cooperate or to defect does not only affect the current payoffs the players receive, but also the game that will be played in the next round. **Fig. 1** illustrates a scenario reflecting the tragedy of the commons. Mutual cooperation improves the quality of the public resource, leading the players to interact in game 1 with comparably high payoffs. Partial defection, on the other hand, leads to a deterioration of the resource;

players move to game 2 where feasible payoffs are lower. The stochastic game is played for many rounds, and overall payoffs are defined as the players' discounted payoffs per round. Transitions between the different states (or games) can be stochastic or deterministic, state dependent or state independent. The well-studied framework of repeated games is a special case of stochastic games with only one state.

The impact of changing environments on evolutionary dynamics has previously been explored in one-shot, non-repeated games, not using the theory of stochastic games[29–33] (see **SI** Section 1.1 for a discussion of the respective literature). In some scenarios, the co-evolution of the players' strategies and their environment can lead to complex dynamics including oscillations between cooperators and defectors[31,32]. But if cooperators are at a disadvantage in every possible environment, environmental feedback proves ineffective to prevent cooperators from going extinct (**Fig. S14**). One-shot models assume that when players make strategic choices, they only consider their present payoff. In stochastic games, players can take a long-term perspective instead. To find an optimal strategy, they need to take into account how their actions affect the subsequent response of their opponents, and the future state of the environment. As we show below, this interplay of reciprocity and payoff feedback can be crucial for maintaining cooperation.

Traditionally, theoretical work on stochastic games considers players who are rational and who can employ arbitrarily complex strategies, but it does not focus on the dynamics of how players adapt their strategies. We introduce an evolutionary perspective to stochastic games. Players do not need to act rationally, but instead they can experiment with their available strategies and imitate others depending on their success[34]. We focus on simple strategies, which are easy to implement and to interpret[35,36]. Such an evolutionary setup has proved useful to understand the dynamics of cooperation in repeated games[8–19].

We first study evolutionary dynamics of a stochastic game with two states (**Fig. 2**). We focus on individuals using pure memory-one strategies, such that a player's move depends only on the present state and on the outcome of the previous round (see **Methods** and **SI** for all details). We compare the evolving cooperation rates in the stochastic game with the cooperation rates in the two associated repeated games in which players always play the same game (**Fig. 2**). We consider both, two-player interactions representing prisoner's dilemmas, as well as n-player interactions representing public goods games. In both cases, players can cooperate by paying a cost $c > 0$. In the prisoner's dilemma, this yields a benefit $b_i > c$ to the co-player, where the benefit b_i depends on the state i. In the public goods game, aggregated costs are multiplied by some factor r_i with $1 < r_i < n$ (dependent on the state i), and redistributed among all group members. Game 1 is more profitable than game 2 if $b_1 > b_2$ or $r_1 > r_2$. Players only find themselves in game 1 if everyone has cooperated in the previous round. Our simulations show that this feedback can boost cooperation dramatically. For reasonable parameter combinations, we find that in the stochastic game populations quickly adapt towards full cooperation, although neither of the two associated repeated games yields substantial cooperation levels.

In the stochastic game, cooperation evolves because defectors loose out twice: once, because they risk to receive less cooperation from their reciprocal co-player in future, and second because players collectively move towards a less beneficial game. The stochastic game is most effective in boosting cooperation if the benefit in game 1 is intermediate (**Fig. S2**). If b_1 is too low, the double loss present in the stochastic game does not suffice to incentivize mutual cooperation, whereas if b_1 is comparably high, players cooperate in the first game anyway. For an evolutionary setup, we show that stochastic games can lead to cooperation, even if all the individual repeated games alone fail.

We derive a condition for the stability of cooperation in stochastic games with two states and state-independent transitions. A numerical analysis for the two-player case suggests that full cooperation emerges when Win-Stay Lose-Shift[9] (WSLS) becomes stable (**Fig. S6**). This strategy prescribes to cooperate in the next round if and only if both players used the same action in the previous round. In a conventional repeated prisoner's dilemma, WSLS is a Nash equilibrium if $b \geq 2c$.[8] In the stochastic game WSLS is an equilibrium if

$$\big(2q_2 - q_0\big)b_1 + \big(1 - 2q_2 + q_0\big)b_2 \geq 2c. \tag{1}$$

The parameters q_i refer to the conditional probability that the players will be in game 1 in the next round, given that i of them have cooperated in the present round. If mutual cooperation leads to game 1 and mutual defection to game 2, we have $q_2 = 1$ and $q_0 = 0$. Therefore WSLS is stable if $2b_1 - b_2 \geq 2c$. Because $b_1 > b_2$, this condition is easier to satisfy than the respective conditions for the two associated repeated games.

Condition (1) highlights that the stability of cooperation depends on how the states change with respect to the players' decisions. To systematically explore the impact of this exogenous feedback, we have repeated our simulations for all eight deterministic and state-independent two-state games (**Fig. S4**, **Fig. S6**). In six of those stochastic games, players spend more time in the profitable game 1. But only in two of the six stochastic games, players actually cooperate. In line with condition (1), cooperation only evolves if $q_2 = 1$ and $q_0 = 0$, with q_1 turning out to be irrelevant. Stochastic games are most effective in promoting cooperation if mutual cooperation improves the public good, while mutual defection deteriorates it, which is a very natural scenario. Analogous conclusions hold for multi-player interactions with more than two states (see **SI**).

Probabilistic transitions can further enhance the evolution of cooperation. In **Fig. 3a**, we consider a scenario in which mutual cooperation in game 2 leads back to game 1 with probability q. We find that the optimal value of q is intermediate: players should have some chance to return to the better state, but it should not be too easy. In **Fig. 3b** we explore whether individuals become more cooperative if the length of the game is not exogenously given, but affected by the players' decisions. Individuals start in state 1, in which they play a conventional prisoner's dilemma; if one or both of the players defect, there is some probability q that players move towards state 2 in which no further profitable interactions are possible. This form of environmental feedback promotes cooperation; payoffs become maximal for small but positive q. In **Fig. 3c** we consider a model with timeout. Defection leads to a temporal state in which no profitable interactions are possible. The return probability to the regular game is given by q. We derive adaptive dynamics for simple reactive strategies (x, y), where x denotes the cooperation probability after having been in state 1 previously, and y is the cooperation probability after having been in the timeout state. In such a scenario, the fully cooperative strategy $(1, 1)$ can become stable, although unconditional cooperation is never stable in a conventional repeated prisoner's dilemma.

Next we explore the ideal feedback between game payoff and strategic choice. We consider a stochastic game with four players and five states. Defection by a subgroup of players has an immediate, gradual, or delayed negative impact on the benefits of cooperation, or no effect (**Fig. 4**). We obtain the highest cooperation rates for immediate negative impact. The intuitive explanation is as follows: maximum cooperation arises if the players are most incentivized to cooperate in the most valuable game. In the immediate scenario, any deviation from cooperation in game 1 leads to a game with the lowest payoff. Interestingly, even the scenario with a delayed response promotes higher cooperation rates than the stochastic game in which the public good remains unchanged across all states. Amazingly, the lowest cooperation rates are obtained when the benefits of cooperation are high in all five games. We obtain similar conclusions when we consider a state-dependent game in which it may take several successive rounds with mutual defection to end up in the worst possible state (**Fig. S12**, **Fig. S13**).

Direct reciprocity is a mechanism for the evolution of cooperation based on repeated interactions. The standard assumption has been that the same game, with the same payoffs is played again and again. Here we introduce the concept that the game payoff changes in different rounds. In particular, we explore cases where cooperation leads to a more valuable game next round, and defection to a less valuable one. Surprisingly, we find that this setting dramatically boosts cooperation. In the resulting stochastic game, cooperation can prevail even if cooperation is unsuccessful in all individual repeated games. Our observations suggest how naturally occurring or designed feedback can promote cooperation. A tragedy of the commons can be avoided if the environment deteriorates (rapidly) as a consequence of defection. Likewise, cooperation is boosted if there is the prospect of playing for higher gains should the current cooperation succeed. The evolutionary analysis of stochastic games represents a new tool for understanding and influencing human decision making in social dilemmas.

Methods Summary

Stochastic games. To fully describe a stochastic game, one needs to specify five objects: (*i*) the set of players $\mathcal{N}$, (*ii*) the set of possible states S, (*iii*) the set of actions $A(s_i)$ that are available to each player in a given state s_i, (*iv*) the transition function Q which describes how the current state of the environment and the players' actions in a given round determine the state in the next round, and (*v*) a payoff function u, describing how the payoffs of the players in a given round depend on the players' actions and on the present state. The framework of stochastic games does not specify how much time passes between consecutive rounds, nor does it restrict the payoffs that are available in each round. The respective model parameters need to be chosen with respect to the specific application at hand (see **SI** for a detailed description of the framework and how it applies to specific examples). Herein we have considered scenarios in which players face a strict social dilemma in each state, but the framework can easily be adapted to more general payoff constellations (**Fig. S8**).

Throughout the main text, we have considered simple examples of stochastic games. Players can choose between cooperation and defection, and thus their action set is $\{C, D\}$ for each state.

Transitions are symmetric: the transition function Q does not depend on which of the players has cooperated or defected. The payoffs per round are symmetric and in the 2-player case given by payoff matrices. The payoff of a player in the stochastic game is defined as the player's discounted payoff per round over infinitely many rounds. Initially, players are in state 1.

Memory-one strategies. In general, strategies for stochastic games can be arbitrarily complex. A player's action in a given round may depend on the present state and on the whole previous history. To facilitate an evolutionary analysis, we focus on strategies in which players only take into account the present state and the outcome of the previous round. For n-player games with m states, such 'memory-one strategies' can be written as a $2nm$-dimensional vector $\mathbf{p} = (p^i_{a,j})$. Each entry $p^i_{a,j}$ is the player's probability to cooperate in a given round, given that the present state is s_i, and in the previous round the focal player chose action $a \in \{C, D\}$, while j of the $n-1$ other group members have cooperated. When all players use memory-one strategies, the dynamics of a stochastic game can be described by a Markov chain with $m2^n$ possible states $(s_1, C, \ldots, C), \ldots, (s_m, D, \ldots, D)$. In this notation, the first entry refers to the state of the public good in a given round, and the other n entries refer to the players' actions. Using the theory of Markov chains, we compute the players' expected payoffs (see **SI**).

Evolutionary dynamics. To describe how individuals adopt new strategies over time, we consider a standard imitation process[34]. There is a population of size N. Each member of the population is equipped with a memory-one strategy which prescribes how the individual plays the stochastic game. In each evolutionary time step, every player interacts with every other player to derive a payoff from the stochastic game. Then two individuals are drawn randomly from the population, a learner and a role model. The payoffs of those two individuals are π_L and π_R, respectively. The learner adopts the role-model's strategy with probability $\rho = 1/\big(1+e^{-\beta(\pi_R - \pi_L)}\big)$.

The parameter $\beta \geq 0$ corresponds to intensity of selection. For $\beta = 0$, we have random drift. For $\beta > 0$, imitation events are biased in favor of strategies that yield higher payoffs. In addition to imitation events, we allow for random strategy exploration, which correspond to mutations: with probability μ an individual adopts a randomly chosen memory-one strategy instead of imitating a co-player. We analyze the ergodic mutation-selection process with computer simulations. We obtain exact numerical results when exploration events are rare.

Specific methods used for individual figures. Except for the results in Fig. 3c, the main text considers examples where players use pure memory-one strategies subject to small errors (such that $p^i_{a,j}$ is either ε or $1 - \varepsilon$, with $\varepsilon = 0.001$). This assumption allows us to derive numerically exact results in the limit of rare mutations[37]. Further simulations using stochastic memory-one strategies confirm that the respective results are robust (**Fig. S2**). For the evolutionary trajectories of Fig. 2 we have averaged over 100 simulations with mutation rate $\mu = 0.002$. Our numerical results use population size $N = 100$, intermediate selection ($\beta = 1$) for pairwise games and strong selection ($\beta = 10$) for multiplayer games. Except for the stochastic game in Fig. 3b, we assume that future payoffs are not discounted, $\delta \to 1$. Our qualitative results are robust with respect to parameter changes (**Fig. S3**). **Fig. 3c** shows the phase portrait of adaptive dynamics[8] for the game with time-out; the differential equation is derived in the **SI**.

References

[1] Lloyd, W. F. *Two lectures on the checks to population* (Oxford University Press, Oxford, UK, 1833).

[2] Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).

[3] Trivers, R. L. The evolution of reciprocal altruism. *The Quarterly Review of Biology* **46**,

35–57 (1971).

[4] Axelrod, R. *The evolution of cooperation* (Basic Books, New York, NY, 1984).

[5] Ostrom, E. *Governing the commons: The evolution of institutions for collective action* (Cambridge Univ. Press, 1990).

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

[7] Van Lange, P. A. M., Balliet, D., Parks, C. D. & Van Vugt, M. *Social dilemmas – The psychology of human cooperation* (Oxford University Press, Oxford, UK, 2015).

[8] Sigmund, K. *The Calculus of Selfishness* (Princeton Univ. Press, 2010).

[9] 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).

[10] 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 B* **264**, 513–519 (1997).

[11] Killingback, T. & Doebeli, M. The continuous Prisoner's Dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *The American Naturalist* **160**, 421–438 (2002).

[12] 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).

[13] 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).

[14] 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, 2016).

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

[16] 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).

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

[18] Garcia, J. & van Veelen, M. In and out of equilibrium I: Evolution of strategies in repeated games with discounting. *Journal of Economic Theory* **161**, 161–189 (2016).

[19] Stewart, A. J. & Plotkin, J. B. Small groups and long memories promote cooperation. *Scientific Reports* **6**, 26889 (2016).

[20] Shapley, L. S. Stochastic games. *Proceedings of the National Academy of Sciences* **39**, 1095–1100 (1953).

[21] Neyman, A. & Sorin, S. (eds.) *Stochastic games and applications* (Kluwer Academic Press, Dordrecht, 2003).

[22] Mertens, J. F. & Neyman, A. Stochastic games. *International Journal of Game Theory* **10**, 53–66 (1981).

[23] Mertens, J. F. & Neyman, A. Stochastic games have a value. *Proceedings of the National Academy of Sciences USA* **79**, 2145–2146 (1982).

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

[25] Ledyard, J. O. Public goods: A survey of experimental research. In Kagel, J. H. & Roth, A. E. (eds.) *The Handbook of Experimental Economics* (Princeton Univ. Press, 1995).

[26] Milinski, M., Sommerfeld, R. D., Krambeck, H.-J., Reed, F. A. & Marotzke, J. The collective-risk social dilemma and the prevention of simulated dangerous climate change. *Proceedings of the National Academy of Sciences USA* **105**, 2291–2294 (2008).

[27] Alur, R., Henzinger, T. & Kupferman, O. Alternating-time temporal logic. *Journal of the*

ACM **49**, 672–713 (2002).

[28] Miltersen, P. B. & Sorensen, T. B. A near-optimal strategy for a heads-up no-limit texas hold'em poker tournament. *AAMAS'07* 191–197 (2007).

[29] Ashcroft, P., Altrock, P. M. & Galla, T. Fixation in finite populations evolving in fluctuating environments. *Journal of The Royal Society Interface* **11**, 20140663 (2014).

[30] Gokhale, C. S. & Hauert, C. Eco-evolutionary dynamics of social dilemmas. *Theoretical Population Biology* **111**, 28–42 (2016).

[31] Hauert, C., Holmes, M. & Doebeli, M. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B* **273**, 2565–2570 (2006).

[32] Weitz, J. S., Eksin, C., Paarporn, K., Brown, S. P. & Ratcliff, W. C. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences USA* **113**, E7518–E7525 (2016).

[33] Tavoni, A., Schlüter, M. & Levin, S. A. The survival of the conformist: Social pressure and renewable resource management. *Journal of Theoretical Biology* **299**, 152–161 (2012).

[34] Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Physical Review E* **74**, 011909 (2006).

[35] Nowak, M. A. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Applicandae Mathematicae* **20**, 247–265 (1990).

[36] Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435–444 (2006).

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

Acknowledgments. This work was supported by the European Research Council Start Grant 279307: Graph Games (to K.C.), Austrian Science Fund (FWF) Grant P23499-N23 (to K.C.), FWF NFN Grant S11407-N23 Rigorous Systems Engineering/Systematic Methods in Systems Engineering (to K.C.), Office of Naval Research Grant N00014-16-1- 2914 (to M.A.N.), and the John Templeton Foundation (M.A.N.). C.H. acknowledges generous support from the ISTFELLOW program.

Author contributions. All authors conceived the study, performed the analysis, discussed the results and wrote the manuscript.

Author information. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.H. (christian.hilbe@ist.ac.at).

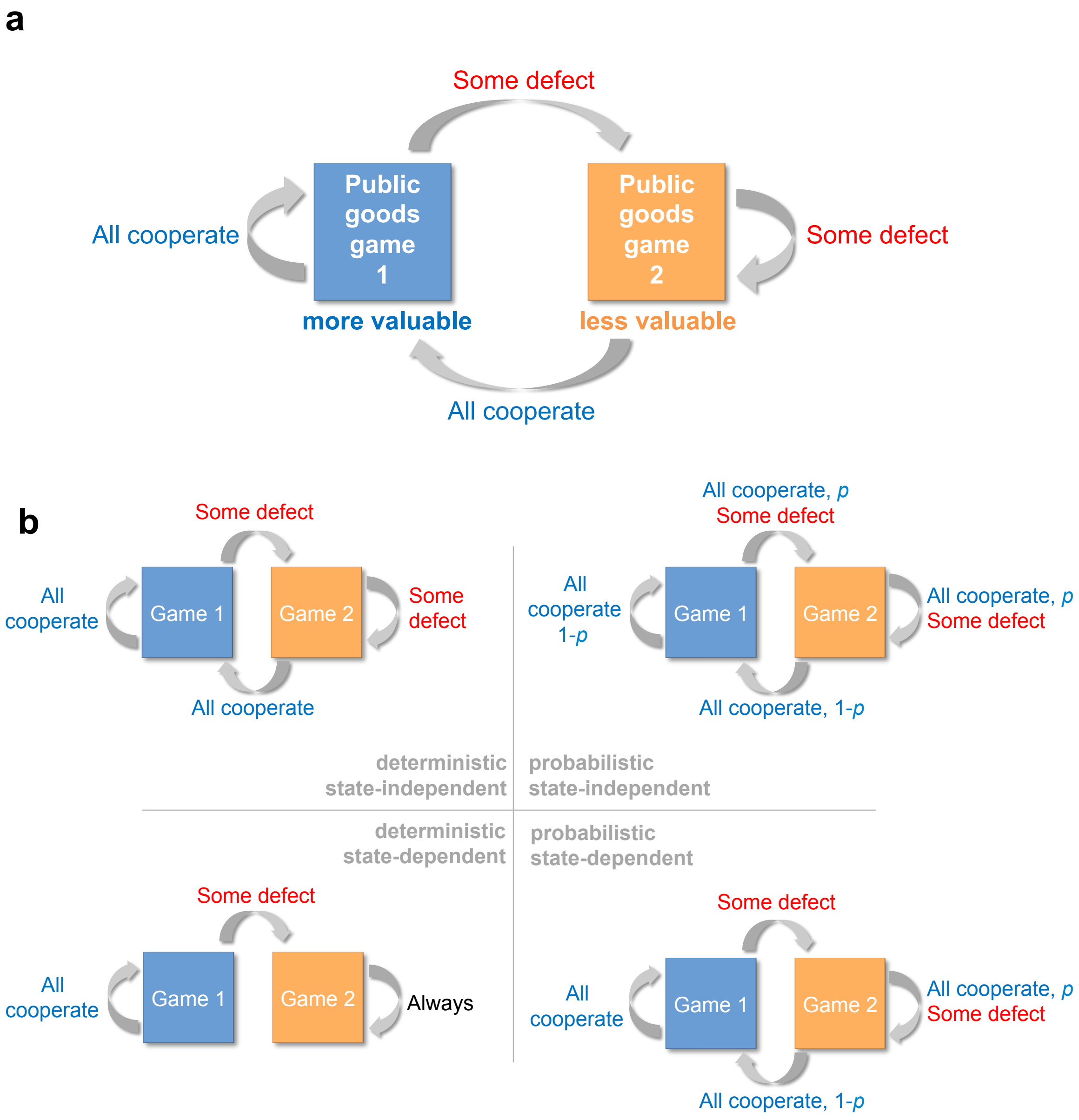

Figure 1: In "stochastic games", the players' decisions in one round determine the game that will be played next round. a, For example, if some players defect in a public goods game, the environment could deteriorate and thereby reduce the value of the public good. If all cooperate, the environment could recover and the original value of the public good might be restored. In this illustration, we show two public goods games with $r_1 > r_2$. **b,** A stochastic game is deterministic if the players' actions and the current game uniquely determine the game that will be played next round, and there are no chance events. It is state-independent, if the game in the next round only depends on the players' actions but not on the current game (state). Thus, we distinguish four different types of stochastic games, depending on whether transitions are deterministic or probabilistic and whether they are state-independent or state-dependent. We note that even a game that only involves deterministic transitions is referred to as a "stochastic game", because it represents a special case of the framework.

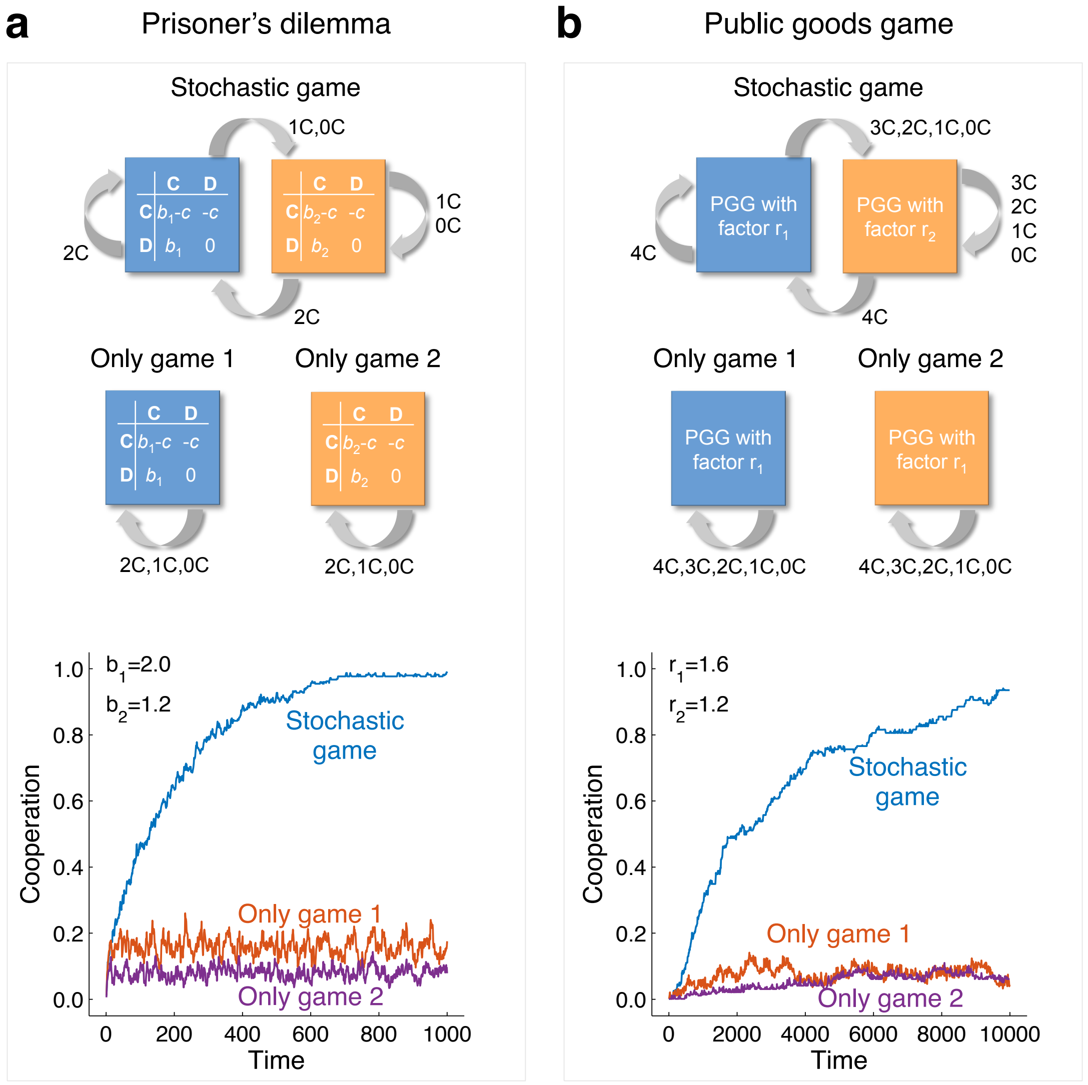

Figure 2: Stochastic games can promote cooperation even if all individual games favor defection. We study **a,** the repeated prisoner's dilemma, which is a two-player game, and **b,** the repeated public goods game (PGG), which here is interpreted as a four-player game. In both cases, the first game has a higher benefit from cooperation than the second game. In the stochastic game, if all players cooperate the next round will be the first game, but if some players defect the next round will be the second game. In the standard, repeated games, the same game is used in every round. An analysis based on evolutionary dynamics (see **Methods**) reveals that each of the standard repeated games fails to support cooperation, while the stochastic game favors cooperation. Parameter values: **a,** $b_1 = 2, b_2 = 1.2, c = 1$. **b,** $r_1 = 1.6, r_2 = 1.2, c = 1$.

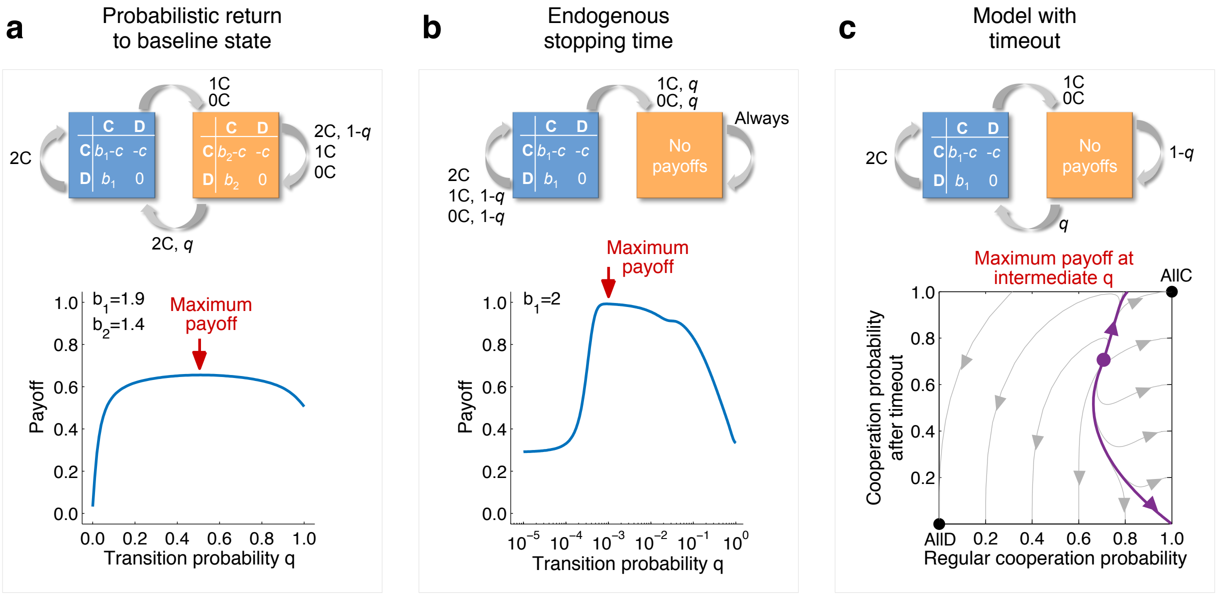

Figure 3: Probabilistic transitions maximize cooperation in three different stochastic games. a, Game 1 is more profitable than game 2, but mutual cooperation in game 2 leads to game 1 only with probability q. The evolving average payoffs are maximized for intermediate q. **b,** Game 1 is left with probability q if at least one player has defected. The optimal value of q is small but positive for games with a finite number of rounds (continuation probability $\delta = 0.999$). **c,** Defection leads to a time-out with an expected duration that depends on a return probability q. We derive the adaptive dynamics for strategies that only take into account if players have been in game 1 in the previous round or in the time out. Depending on the parameters, $AllC$ is a stable endpoint of evolution: a population in which everyone cooperates in game 1 cannot be invaded by any alternative strategy. Again the optimal value of q is intermediate: low values of q increase the area of initial populations that move towards cooperation, but they also make occasional errors more costly (parameters $b_1 = 3$, $c = 1$; $q = 1/2$ is shown).

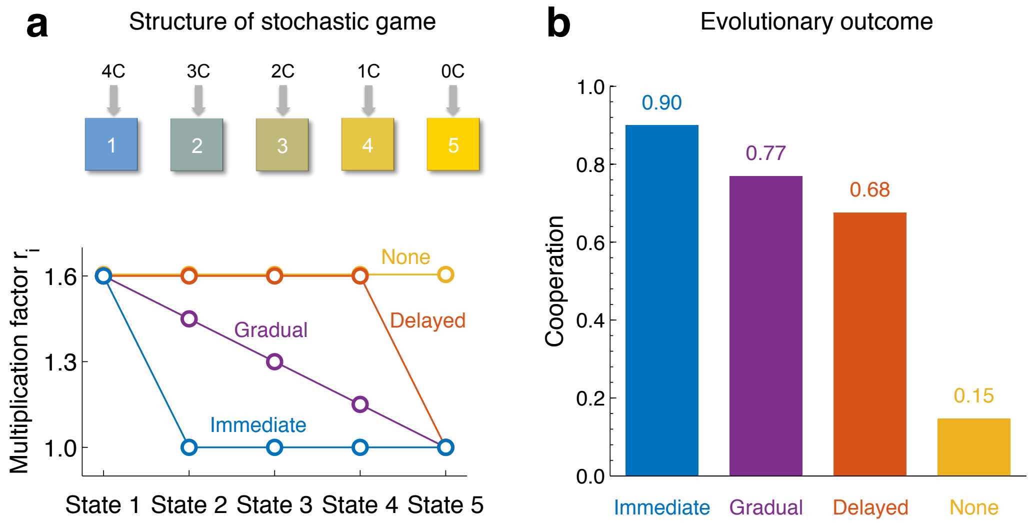

Figure 4: Strong immediate feedback maximizes cooperation. a, A 4-player scenario in which cooperation improves and defection reduces the value of the public good. Transitions are state-independent: the next state only depends on the number of cooperators but not on the previous state. In game 1, contributions to a public good are multiplied by the highest factor $r_1 = 1.6$. In game 5, cooperation does not produce any social benefit, $r_5 = c = 1$. For the payoff in the intermediate games 2, 3 and 4 we distinguish three cases: partial defection has immediate, gradual, or delayed consequences on the multiplication factor of the public good. In addition, we consider a fourth scenario in which the multiplication factor remains high in all states. **b,** An evolutionary analysis confirms that immediately deteriorating public resources are most favorable to cooperation, because they make unilateral exploitation a risky strategy. However, all three stochastic games in which the benefits of cooperation vary lead to substantially more cooperation than the game with no environmental feedback. Results of exact numerical calculations are shown. The ranking is preserved if an alternative regime is considered in which it takes multiple rounds of mutual defection to arrive at the worst state: immediate is better than gradual or delayed (**Fig. S12**).

Supplementary Information:
Evolution of cooperation in stochastic games

Christian Hilbe[1,2], Stepan Simsa[3], Krishnendu Chatterjee[2], Martin A. Nowak[1,4]
[1]Program for Evolutionary Dynamics, Harvard University, Cambridge MA 02138, USA
[2]IST Austria, 3400 Klosterneuburg, Austria
[3]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
[4]Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge MA 02138, USA

Section 1 gives a brief overview of previous work. **Section 2** describes our approach and methods. **Section 3** presents further results for state-independent stochastic games with deterministic transitions. We show how the success of cooperation depends on the transition structure of the stochastic game. We give an analytical condition for cooperation to evolve based on the stability of Win-Stay Lose-Shift. We demonstrate that our results are robust with respect to changes in evolutionary parameters. In **Section 4**, we apply our framework to four different scenarios with probabilistic or state-dependent transitions. In the appendices, we present some of the more technical aspects of our study. In **Appendix A**, we discuss the feasibility of cooperation in one-shot social dilemmas. **Appendix B** provides the proofs of our mathematical results. **Appendix C** gives our MATLAB algorithm for calculating payoffs in stochastic games with two states and n players.

1 Previous work

We have studied the evolutionary dynamics of strategies in stochastic games. Several papers in evolutionary game theory have explored the effects of changing payoffs. However, none of those papers use stochastic games; there are no repeated interactions, and the players' strategies do not allow for a targeted reaction to the current game. On the other hand, there is a rich literature on stochastic games in general, but previous studies do not consider evolution of strategies in stochastic games.

1.1 Previous work on evolutionary dynamics in games with variable payoffs

Traditionally, evolutionary game theory studies the dynamics of strategies if players interact in games with fixed payoffs[1–4]. In recent years there has been a growing interest in exploring the dynamics of strategies when the game's payoffs are allowed to vary in time.

In the simplest case, these studies assume that changes in the game's payoffs are exogenously driven, and therefore independent of the composition of the population. For example, Assaf *et. al.*[5] consider social dilemmas where some parameters, such as the cost-to-benefit ratio of cooperation, are subject to extrinsic noise. Keeping the expected value of these model parameters constant, they find that noise can increase the fixation probability of a rare cooperator. Ashcroft *et. al.*[6] consider a model in which the players' environment randomly switches between different states. The states in turn affect which 2×2 game is played. They show that such transitions between different states can facilitate the invasion of a rare mutant. Gokhale and Hauert[7] explore the impact of seasonal changes. In their model, the multiplication factor of a public good game and a synergy factor change periodically in time. They find that the survival of cooperation critically depends on the timescale at which these fluctuations occur. If environmental changes are slow compared to the strategy dynamics, cooperation typically goes extinct even if cooperators would be able to survive in an average environment.

Other studies have considered scenarios in which the players' strategies co-evolve with their environment. In these models, cooperators improve an environmental parameter, which in turn affects the incentives to cooperate. Hauert *et. al.*[8] have used such a setup to explore the coexistence of cooperators and defectors in public good games. In their model, the total population size depends on the number of cooperators. When the population is large, defection is more beneficial. As a consequence, the number of cooperators as well as the total population size decreases. However, in smaller populations individuals interact in smaller groups, in which cooperators are at an advantage. Cooperation can re-invade. For some parameter constellations, demographic feedback can thus lead to persistent oscillations, such that both the fraction of cooperators as well as the population size fluctuate periodically[9].

Weitz *et. al.*[10] consider a more general framework for the coevolution of strategies and the environment. In their model, members of an infinite population interact in a 2×2 game. The game's payoff matrix $A(n)$ depends on some environmental parameter $n \in [0, 1]$. This environmental parameter in turn depends on how many players cooperate. If defection is a dominated strategy in the payoff matrix $A(0)$, and if cooperation is dominated in $A(1)$, populations may experience an "oscillating tragedy of the commons". In addition, Weitz *et. al.*[10] describe all possible evolutionary scenarios depending on the payoffs in $A(0)$. Their classification shows that cooperation goes extinct if $A(n)$ is a prisoner's dilemma for all n.

Another line of research has explored the evolution of cooperation when members of a community share a renewable resource. The amount of resource available at any time depends on its intrinsic growth function, and on the previous extractions by community members. The corresponding models show that cooperation can be sustained if there is sufficient societal pressure for a responsible use of resources[11], or if groups of defectors are more likely to perish because of resource depletion[12].

All of the above models consider one-shot games. While the players' actions may affect the environment, players do not take the long-term consequences of their actions into account. Instead, they revise their strategies only based on the present payoffs. Players do not anticipate the future state of the environment, nor are they able to engage in reciprocal interactions. As a consequence, these models cannot explain which kind of environmental feedback helps to sustain cooperation in strict social dilemmas: if cooperators are at a disadvantage in each environmental state, cooperators also go extinct if the environment co-evolves with the players' strategies (see **Appendix A** for a detailed analysis).

There has also been research on repeated games where players have access to continuous degrees of cooperation[13–19]. Depending on the outcome of previous games, players can decide to increase their contributions to a public good. In contrast to our approach, the game remains the same, irrespective of the players' actions. The players have continuous choices, but the game itself does not change.

Novelty of our approach

Our framework differs from previous evolutionary studies with variable game payoffs in the following aspects:

1. None of the previous papers study repeated interactions.
2. None of the previous papers consider stochastic games.
3. In previous work, players do not tailor their response to the specific environment which they are currently facing.
4. Previous work has suggested that under changing environments, cooperation can only evolve under restrictive conditions; cooperation is not sustained if players face a social dilemma in each environmental state[6–10]. In contrast, the framework of stochastic games shows that cooperation can evolve even if it is disfavored in each state (**Fig. 2**).
5. Our framework is general and can be adapted to many new applications.

1.2 Previous work on stochastic games

Stochastic games have been introduced in the seminal work of Shapley[20] for zero-sum discounted-sum payoff. They were later extended to limit-average payoff[21]. The framework of stochastic games has been applied in computer science to model reactive systems[22–24], to analyze algorithms for poker[25,26], to industrial organization, accumulation of capital, and resource extraction[27]. The literature of stochastic games has a wealth of deep mathematical results, such as determinacy results for discounted-sum payoff[20], and the celebrated result for limit-average payoff[28] that uses results on Puiseux series[29] and results of Shapley[20]. These classical results for stochastic games consider rational players with arbitrarily complicated strategies. Such a general framework implies that many questions are open. For example, the existence of Nash equilibrium in multi-player stochastic games is open for limit-average payoff. Recent results[30,31] establish existence of Nash equilibrium in the special case of two-player nonzero-sum games, but the existence for three of more players remains open. The existence of equilibrium is known for discounted-sum payoff[32]. While these results consider complex strategies and existence of equilibrium rather than the dynamics that can lead to equilibrium, we study evolutionary dynamics of stochastic games with a simpler class of strategies.

Novelty of our approach

1. We bring the framework of stochastic games to evolutionary biology. We make the framework applicable to the broad context of human decision making in social dilemmas.
2. While the classical results for stochastic games take an equilibrium perspective, we introduce evolutionary dynamics and study how strategic behavior evolves.
3. Previous work has considered rational players with arbitrarily complex strategies, for which even existence of equilibrium is not always known. We study simpler classes of strategies (such as reactive or memory-1 strategies), which make a computational analysis feasible.
4. While our analysis is motivated by problems in evolutionary biology, our new dynamic perspective opens up a new research area, with many new exciting questions (e.g., on the dynamical stability of equilibria). These questions are relevant to biologists, mathematicians, economists and computer scientists.

We believe that the combination of evolutionary game theory with stochastic games provides an important new impetus to the study of social interactions with dynamic incentives. Our framework allows us to analyze how cooperation emerges when individuals affect their environment, and when they are able to react to new environmental conditions and to their co-players' past actions.

The framework of stochastic games opens up many new directions for evolutionary game theory. Future work could explore stochastic games in structured populations[33,34], where cooperation with one player allows the possibility of forming new connections with other players. In such a scenario, the entire network structure could co-evolve with the players' actions[35]. Alternatively, defection (but not cooperation) could allow the possibility of ostracism or punishment, such that much of the previous work on the evolution of incentives[36,37] may be recast in terms of stochastic games.

2 Model and Methods

In the following, we first give a full description of stochastic games. Then we formally introduce memory-one strategies, and we show how the assumption of bounded memory can be used to calculate payoffs explicitly. Finally, we also specify the details of the evolutionary process that we have used to explore the dynamics of cooperation in stochastic games.

2.1 General setup of stochastic games

To define a stochastic game, we need to specify (*i*) the set of players; (*ii*) the set of possible states (*iii*) the set of possible actions that each player can take in each possible state; (*iv*) a transition function that describes how states change over time; (*v*) the payoff function, which describes how the players' actions and the current state affect the players' payoffs. In the following, we introduce these objects formally.

(*i*) **The set of possible players.** As the set of possible players, we take the set $\mathcal{N} = \{1, \dots, n\}$. Throughout this work we interpret this set as a group of individuals, interacting in various social dilemmas subject to environmental feedback.

(*ii*) **The set of possible states.** We define $S = \{s_1, \dots, s_m\}$ to be the finite set of possible states. Herein, we interpret these states as the different environmental conditions that the players may face. The players' environment is considered to represent the sum of all exogenous factors that may affect the group of individuals. In particular, the environment includes all ecological and social constraints the players are subjected to. The set of states may also include states that merely encode the previous history of play (for example, by introducing a state that is reached if and only if all players have cooperated in the last k rounds). We illustrate this possibility in Section 4.4, where we model a game in which mutual defection has delayed consequences.

Without loss of generality, we will assume that as the stochastic game begins, players find themselves in state s_1.

(*iii*) **The set of possible actions.** In each state, players independently need to choose which action they would like to take. Herein, we only explore the case where the game being played in each state is a social dilemma with two possible actions. Therefore, player j's action set in any given state takes the form $A_j = \{C, D\}$ for all $j \in \mathcal{N}$. We interpret these two actions as cooperation and defection, respectively. The outcome of a given round of the stochastic game is then described by an n-tuple $\mathbf{a} = (a_1, \dots, a_n)$ where $a_j \in \{C, D\}$ is the action taken by player j. For example, the n-tuple $\mathbf{a} = (C, C, \dots, C)$ denotes the outcome of a round in which all players cooperated. The set $A = A_1 \times \ldots \times A_n$ of all such n-tuples is called the set of action profiles.

(*iv*) **The transition function.** States can change from one round to the next. The state of the stochastic game in the next round depends on the present state, on the players' present actions, and on chance. Formally, these changes of the states are described by a transition function $Q : S \times A \to \Delta^S$, where Δ^S is the set of probability distributions over the set of states S,

$$\Delta^S = \left\{ x = (x_1, \dots, x_m) \in \mathbb{R}^m \;\middle|\; x_i \geq 0 \text{ for all } i, \text{ and } x_1 + \ldots + x_m = 1 \right\}. \tag{1}$$

That is, $Q(s, \mathbf{a}) = (x_1, \dots, x_m)$ gives the probability to be in each of the m states, given that the previous state is $s \in S$, and the actions in the previous round are $\mathbf{a} = (a_1, \dots, a_n) \in A$. As an example we can take the stochastic game depicted in **Fig. 2a** of the main text; for that example, the transition function Q takes the form

$$\begin{aligned}
Q\big(s_1, (C, C)\big) &= (1, 0), \quad & Q\big(s_2, (C, C)\big) &= (1, 0),\\
Q\big(s_1, (C, D)\big) &= (0, 1), \quad & Q\big(s_2, (C, D)\big) &= (0, 1),\\
Q\big(s_1, (D, C)\big) &= (0, 1), \quad & Q\big(s_2, (D, C)\big) &= (0, 1),\\
Q\big(s_1, (D, D)\big) &= (0, 1), \quad & Q\big(s_2, (D, D)\big) &= (0, 1).
\end{aligned} \tag{2}$$

That is, if both players cooperated in the present round, the players will find themselves in State 1 in the next round with certainty; after all other outcomes, players will necessarily be in State 2.

In the above example, the entries of Q did only take the values 0 and 1; moreover, Q did only depend on the players' actions, but not on the present state of the environment. We call such transition functions *deterministic* and *state-independent*, respectively. We note that even if all transitions are deterministic, the game is still called a "stochastic game" because of two reasons. First, games with deterministic transitions represent a special case of the general framework. Second, even if transitions are deterministic, the next round's state may still depend on chance events if the players' strategies do (i.e., if players use strategies that randomize between different actions).

If we want to refer to the j-th element in $Q(s, \mathbf{a})$, i.e. the probability to move to state s_j after observing outcome $(s, \mathbf{a})$, we will sometimes use the notation $Q(s_j|s, \mathbf{a})$. Herein, we only consider *symmetric* transitions: the next state may depend on the number of cooperators in the present round, but it does not depend on who of the players cooperated. Instead of writing $Q(s_j|s, \mathbf{a})$ we can thus use the notation $Q(s_j|s, k)$, where k is the number of C's in the n-tuple $\mathbf{a}$. Moreover, for state-independent transition functions Q, we will drop the dependence on the previous state s, and write $Q(s_j|k)$ to denote the probability to move to state s_j after k players have cooperated.

When there are only two possible states, we can also represent the transition function by a vector $\mathbf{q} = (q_k^i)$. Each entry q_k^i denotes the probability that the players find themselves in State 1 in the next round, given that the previous state was s_i and that k of the players have cooperated. When the stochastic game is state-independent, we drop the upper index i. As an example, using this notation we can represent the transition functions of **Fig. 2** as $\mathbf{q} = (q_n, q_{n-1}, \ldots, q_1, q_0) = (1, 0, \ldots, 0, 0)$. That is, players find themselves in state 1 if and only if all n group members have cooperated in the previous round.

(*v*) **The payoff function.** The (stage game) payoff function $u : S \times A \to \mathbb{R}^n$ describes how the players' payoffs in a given round depend on the present state and on the players' joint actions. Symmetric games between two players can be represented by the respective payoff matrix,

$$U^i = \begin{pmatrix} u^i_{CC} & u^i_{CD} \\ u^i_{DC} & u^i_{DD} \end{pmatrix}. \tag{3}$$

where $u^i_{a\tilde{a}}$ refers to a player's payoff in state s_i, given that the focal player chooses the action a and that the co-player chooses the action $\tilde{a}$. For symmetric games between n players, we use $u^i_{a,j}$ to denote a player's payoff in state s_i if the focal player chooses action $a \in \{C, D\}$ and if j of the other players cooperate. Importantly, we point out that herein we focus on stochastic games where each stage game corresponds to a comparable social dilemma; allowing for arbitrary payoffs in some of the stage games can give rise to somewhat trivial results (for example, if mutual defection leads to a state in which payoffs are strongly negative irrespective of the players' further actions).

As an example for the payoff function in a stochastic game with more than two players, we have considered a group of players that engages in a public good game in each state s_i. That is, each player can decide whether she wants to contribute an amount $c > 0$ towards a common pool. Total contributions are multiplied by a factor r_i (which may depend on the state of the environment), and equally shared among all participants. If j is the number of cooperating co-players in a given round, the payoff of a player is

$$u^i_{a,j} = \begin{cases} \dfrac{j+1}{n} r_i c - c & \text{if } a = C \\ \dfrac{j}{n} r_i c & \text{if } a = D. \end{cases} \tag{4}$$

The prisoner's dilemma that we have considered for 2-player interactions can be considered as a special case of a public goods game with $n=2$. In that case, the effective cost of cooperation is $c(1-r_i/2)$, and the benefit of cooperation to the co-player is $cr_i/2$.

The framework of stochastic games assumes that players interact over infinitely many rounds, but future payoffs may be discounted by a discount factor δ with $0 < \delta < 1$ (equivalently, one may interpret δ as the continuation probability of having a further interaction after the present round). To define the players' overall payoff for the stochastic game, let $s(t)$ denote the state in which the players find themselves in round t, and let $\mathbf{a}(t)$ denote the action profile played in that round. The payoff π of the stochastic game is the weighted sum

$$\pi = (1-\delta) \sum_{t=0}^{\infty} \delta^t \cdot u\big(s(t), \mathbf{a}(t)\big). \tag{5}$$

In the main text, we have often focused on the limit of no discounting, $\delta \to 1$, in which case payoffs are given by the limit of the players' average payoffs per round,

$$\pi = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} u\big(s(t), \mathbf{a}(t)\big). \tag{6}$$

In the examples used in the main text, the limit in (6) did always exist, due to our assumption that players employ simple memory-one strategies with small errors[1]. If that limit exists, the two definitions (5) and (6) coincide as $\delta \to 1$.

Repeated games, which have been extensively studied in evolutionary game theory, can be considered as a special case of a stochastic game with only one possible state, $S = \{s_1\}$. Conversely, any stochastic game between n players may be interpreted as an (asymmetric) repeated game with $n+1$ players, in which the additional player chooses the state of the next round according to the transition function Q. Despite these connections, stochastic games can be very different from repeated games in terms of the equilibria that are possible. For infinitely repeated games, the folk theorem applies: any feasible payoff combination can be supported as an equilibrium, as long as all players yield more than their maxmin payoff[38]. In particular, mutual cooperation can always be achieved in the repeated prisoner's dilemma.

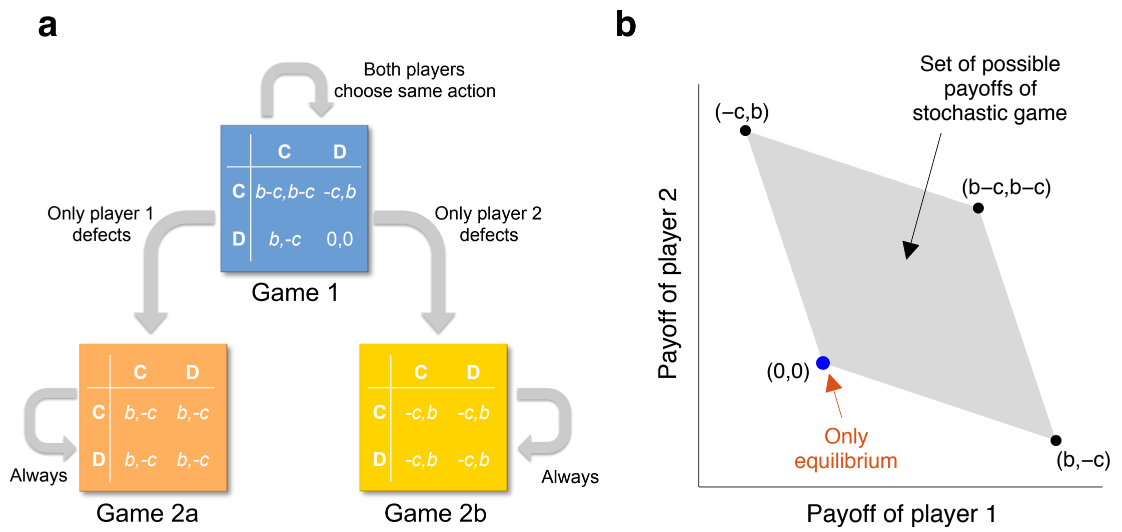

Figure S1: In stochastic games, an analogous result to the Folk theorem of repeated games does not need to hold. For infinitely repeated games, the Folk theorem states that any feasible payoff combination that gives each player more than her minmax payoff can be supported in an equilibrium. An analogous result does not need to hold for stochastic games. **a,** As an example, we can consider a stochastic game in which the initial State 1 takes the form of a prisoner's dilemma. Moreover, we assume that both players have a strong incentive to be the first to defect (because the first unilateral defector has a lasting advantage in all subsequent rounds). Note that in this game, transitions are not symmetric: the next state may depend on who of the two players defects. **b,** In this stochastic game, mutual defection is the only outcome consistent with equilibrium considerations, although mutual cooperation would make both players better off.

For stochastic games, it is well known that an analogous result does not need to hold; **Fig. S1** provides a simple counter-example in the context of social dilemmas. This counter-example is based on a game with absorbing states; once such states are reached, they cannot be left any more. If the stochastic game is instead "communicating" (such that one can get from any state to any other state with positive probability by playing an appropriate strategy), then a Folk theorem result is feasible, see for example Ref. 39.

In the main text we have highlighted another difference between stochastic games and repeated games: even if a stochastic game and the corresponding repeated games all allow for stable mutual cooperation, the likelihood of reaching a cooperative equilibrium through evolutionary processes can be vastly different.

To conclude this section, let us comment on a few questions that arise when stochastic games are applied to specific examples. First, we note that the framework of stochastic games does not specify how much time passes between two consecutive rounds. In some applications, such as in Hardin's example of farmers sharing a joint pasture[40], this timespan might be best thought of as a year. In other applications, like when teams in companies are assigned better tasks the more productive they have been in the past, this time depends on the length of the specific task. In applications, one may thus wish to reflect the amount of time that passes between rounds by choosing an appropriate discount factor δ on future payoffs.

Second, in some applications players may change their actions at a faster rate than the game payoffs change. For example, when all players of a group defect, the corresponding changes in the players' environment may not happen immediately, but only with one round delay. Similarly, it may take several consecutive rounds of mutual defection for the environment to deteriorate. Such scenarios can be studied when further states are added to the stochastic game's state space, which keep track of how many consecutive rounds of mutual defection have occurred in the past. We illustrate this approach with a simple example in Section 4.4.

Finally, the framework of stochastic games does not make any restrictions on the payoffs the players face in any given state of the environment. To apply the framework to a specific example, one only needs to formulate which payoff consequences the players' actions have. In some applications, the players' actions may have drastic consequences on the next round's payoffs, whereas in other applications the payoff feedback may be more gradual. We illustrate different ways to introduce payoff feedback in Section 4.4.

2.2 Calculation of payoffs when players use memory-one strategies

Strategies for stochastic games can be arbitrarily complex; in general they take the whole previous history of the players' actions and of the previously visited states as an input, and they return a value in the interval [0,1] as an output (the player's cooperation probability in the next round). To make an evolutionary analysis feasible, we will focus here on the subset of memory-one strategies. A player with such a strategy bases her decision only on the current state, and on the actions played in the previous round. Formally, a memory-one strategy is a map $P : S \times \tilde{A} \to [0,1]$, where $\tilde{A} = A \cup \{\emptyset\}$. The value of $P(s, \mathbf{a})$ corresponds to the players' probability to cooperate in the next round, given that the present state is $s \in S$ and that the players' actions in the previous round are represented by $\mathbf{a} \in A$ (the empty set is included in $\tilde{A}$ to encode the players' move in the very first round in which no previous history of actions is available). In Table S1, we present a few simple examples of memory-one strategies for stochastic games between n players. In particular, this table includes the strategies *AllD*, proportional Tit-for-Tat ($pTFT$), and Win-stay Lose-shift ($WSLS$).

Restricting our attention to memory-one players is useful for two reasons. First, memory-one strategies are comparably simple to interpret, which facilitates an intuitive understanding of the resulting evolutionary dynamics. At the same time, memory-one strategies are sufficiently general to cover a wide array of interesting behaviors that have been shown to be important in the context of cooperation in repeated games, see for example Refs. 18,41–51, and Chapter 3 in Ref. 1. Second, if all players make use of a memory-one strategy, payoffs according to Eqs. (5) and (6) can be calculated explicitly. In that case, the dynamics of play in the stochastic game can be described as a Markov chain, as we describe below.

The states of the Markov chain are all possible combinations $(s, \mathbf{a})$ of environmental states $s \in S$ and action profiles $\mathbf{a} \in A$, which fully describe the outcome of a given round. In particular, if there are m

Name	Definition	Description
AllD	$P(s,\mathbf{a})=0$ for all s and $\mathbf{a}$	Strategy that defects in every round, independently of the present state and of the actions of the co-players.
AllC	$P(s,\mathbf{a})=1$ for all s and $\mathbf{a}$	Strategy that always prescribes to cooperate.
Grim	In the first round $P(s_1,\emptyset)=1$, then $P(s,\mathbf{a})=1$ if and only if $\mathbf{a}=(C,\dots,C)$, otherwise $P(s,\mathbf{a})=0$.	Trigger strategy that prescribes to cooperate as long as everybody has cooperated in the previous round.
pTFT	In the first round $P(s_1,\emptyset)=1$, then $P(s,\mathbf{a})=k/(n-1)$ with k being the number of cooperating co-players	Conditionally cooperative strategy; in a game with only two players, $pTFT$ simplifies to the classical Tit-for-Tat strategy[41].
WSLS	In the first round $P(s_1,\emptyset)=1$, then $P(s,\mathbf{a})=1$ if all players used the same action in previous round, otherwise $P(s,\mathbf{a})=0$	Generalization of the corresponding strategy that has proven to be successful in the repeated prisoner's dilemma[42].
*Only*1	$P(s,\mathbf{a})=1$ for all $\mathbf{a}$ if $s=s_1$, $P(s,\mathbf{a})=0$ otherwise	Simple example of a state-dependent strategy – cooperates if and only if players find themselves in the first state.

Table S1: Some examples of memory-one strategies for stochastic games.

environmental states and n players who can choose between cooperation and defection, then that Markov chain has $m\cdot 2^n$ possible states. To construct the transition matrix of the Markov chain, let us assume that the transition function between states is given by Q, and that the players' memory-one strategies are given by $P_1,\dots,P_n$. Then the probability that players move from $(s_i,\mathbf{a})$ in one round to $(s_j,\mathbf{a}')$ in the next has the form

$$M_{(s_i,\mathbf{a})\to(s_j,\mathbf{a}')}=Q(s_j|s_i,\mathbf{a})\cdot\prod_{k=1}^{n}y_k, \tag{7}$$

where the y_k are defined as

$$y_k=\begin{cases} P_k(s_j,\mathbf{a}) & \text{if } a'_k=C\\ 1-P_k(s_j,\mathbf{a}) & \text{if } a'_k=D.\end{cases} \tag{8}$$

That is, the transition probability is a product of $n+1$ factors; the first factor represents the transition towards the next environmental state, whereas the other n factors represent the transitions in the players' behaviors. Similarly, the probability to be in one of the $m\cdot 2^n$ states in the very first round is given by

$$v^0_{(s,\mathbf{a})}=\begin{cases} \prod_{k=1}^{n}z_k & \text{if } s=s_1\\ 0 & \text{else,}\end{cases} \tag{9}$$

where

$$z_k = \begin{cases} P_k(s,\emptyset) & \text{if } a_k = C \\ 1 - P_k(s,\emptyset) & \text{if } a_k = D. \end{cases} \tag{10}$$

To calculate the players' payoffs, let $\mathbf{v}^0$ be the row-vector that contains all the initial probabilities according to (9) and let M be the corresponding transition matrix with entries as defined in (7). When future payoffs are discounted, we compute the vector

$$\mathbf{v} = (1-\delta)\mathbf{v}^0 \sum_{t=0}^{\infty} (\delta M)^t = (1-\delta)\mathbf{v}^0 (I - \delta M)^{-1}. \tag{11}$$

The entries $v_{(s,\mathbf{a})}$ of this vector can be interpreted as the expected frequencies to observe the outcomes $(s, \mathbf{a})$ over the course of the stochastic game. The players' payoffs according to Eq. (5) can then be computed by

$$\pi = \sum_{s \in S, \mathbf{a} \in A} v_{(s,\mathbf{a})} \cdot u(s,\mathbf{a}). \tag{12}$$

In the limit of no discounting on future payoffs, $\delta \to 1$, the vector $\mathbf{v}$ according to Eq. (11) approaches a left eigenvector of the matrix M with respect to the eigenvalue 1. In the main text, we have assumed that the players use pure memory-one strategies subject to small errors, such that the probability to cooperate is either ε or $1-\varepsilon$. Under this assumption, the limiting distribution $\mathbf{v}$ for $\delta \to 1$ is unique, and it is independent of the players' cooperation probabilities in the very first round.

In the following, due to the assumed symmetries of the stochastic game, we will often restrict ourselves to symmetric memory-one strategies. If a player applies a symmetric memory-one strategy, her action only depends on the present state, on her own previous action, and on the number of cooperators among the co-players (but not on the identity of the cooperating co-players). Symmetric memory-1 strategies can be written as a vector $\mathbf{p} = (p_0;\, p^i_{a,j})$. The entry p_0 is the player's cooperation probability in the very first round. The entries $p^i_{a,j}$ represent the player's cooperation probability in all subsequent rounds, given that the present state is s_i and that the player has chosen action $a \in \{C, D\}$ in the previous round, while j other players have cooperated. If there are m states and n players, the space of symmetric memory-1 strategies is $(2mn+1)$–dimensional; in the limiting case of $\delta \to 1$ we can ignore the entry p_0 and the space becomes $2mn$–dimensional.

2.3 Evolutionary dynamics

Herein, we do not presume that players act rationally from the outset. Rather we consider a population of players, and we assume that individuals learn to adopt more profitable strategies over time. To model this process of strategy adaptation, we use a simple pairwise imitation process[2,52]. As is typical in models of direct reciprocity, we thereby assume a separation of timescales: players change their strategies at a slower rate than they make their decisions in the stochastic game. Equivalently, we may also assume that strategy updating occurs at a similar timescale, but when computing their own payoff, players take into account how well their strategies would perform over the entire course of the game.

Specifically, we consider a population of constant size N. Each individual is equipped with a symmetric memory-one strategy that tells the individual how to play the stochastic game under consideration. The strategies of the individuals can change over time, due to imitation events and exploration events (these two events correspond to selection and mutation in biological models). We assume that in each evolutionary time step, individuals randomly form groups to interact in the stochastic game (in other words, we assume that the population is *well-mixed*). Payoffs within each group can be calculated using Eq. (12). Depending on the players' own strategy, and on the strategy distribution in the remaining population, we can thus compute expected payoffs $\bar{\pi}$ for all players. To incorporate imitation events, we assume that after payoffs have been computed, two individuals are randomly drawn from the population, a learner and a role model. The learner compares her own payoff $\bar{\pi}_L$ with the payoff of the role model $\bar{\pi}_R$, and she decides to adopt the role model's strategy with probability[53]

$$\rho = \frac{1}{1+e^{-\beta(\bar{\pi}_R-\bar{\pi}_L)}}. \tag{13}$$

The parameter $\beta \geq 0$ is called the *strength of selection*. In the limiting case $\beta = 0$, the imitation probability simplifies to $\rho = 1/2$, independently of the players' payoffs. In that case, players imitate each other essentially at random. As β increases, imitation decisions are increasingly biased towards strategies that lead to higher payoffs; in the limiting case $\beta \to \infty$, the role model's strategy has only a positive chance of being adopted if $\bar{\pi}_R \geq \bar{\pi}_L$.

To incorporate random strategy exploration, we assume that in each evolutionary time step there is a probability $\mu > 0$ that the learner decides not to look for a role model; instead, she simply picks a new strategy from the set of all possible strategies (all memory-one strategies have the same probability to be picked). These two elementary updating events, imitation and exploration, are then iterated over many evolutionary time steps. This generates a stochastic process on the space of all population compositions. Due to our assumptions on the updating events, this process is ergodic. In particular, we can use evolutionary simulations over many time periods to estimate how often the members of the population choose cooperative strategies in the selection-mutation equilibrium (**Fig. 2** of the main text shows sample runs that illustrate the resulting cooperation dynamics).

When the considered strategy set is finite, and when mutations are sufficiently rare, the abundance of each strategy in the selection-mutation equilibrium can be computed exactly[54,55]. The assumption of rare mutations implies that most of the time, populations are homogeneous. Only occasionally a new mutant strategy arises, and this mutant strategy either fixes or goes extinct before there is a new mutation. As a consequence, the evolutionary process can be described as a Markov chain, where the states correspond to all possible homogeneous populations. By calculating the invariant distribution of this Markov chain, we can compute how often each strategy is played in the long run, as described by Fudenberg and Imhof.[54].

If there are infinitely many strategies (for example, when considering all stochastic memory-one strategies), the above method is not directly applicable any longer. Nevertheless, the assumption of rare mutations can still be useful to simulate the evolutionary process more efficiently, using the approach of Imhof and Nowak[56]. If mutations are sufficiently rare, we can assume that at any point in time there are at most two different strategies present in the population, the resident strategy and a mutant strategy. For such a competition between two strategies only, analytical expressions for a strategy's fixation probability are available[57–59]. Hence, we can explicitly calculate the probability that a randomly chosen mutant strategy fixes (in which case the mutant strategy becomes the new resident strategy), or that it goes extinct (in which case the resident strategy remains). Overall, simulating this process leads to a sequence of subsequent resident strategies. Based on this sequence, we can calculate the average cooperation rate and the average payoff of the population over time.

3 Stochastic games with state-independent and deterministic transitions

3.1 A useful result to reduce the dimension of the strategy space

In the following section, we will discuss a few results that apply to the case when the players' actions uniquely determine the next state of the stochastic game. To this end, let us recall some definitions. We say a stochastic game is deterministic if transitions are independent of chance,

$$Q(s_j|s,\mathbf{a})\in\{0,1\} \quad \text{for all states } s_j, s \text{ and action profiles } \mathbf{a}. \tag{14}$$

The stochastic game is state-independent if the next state only depends on the players' previous actions, but not on the previous state,

$$Q(s_j|s,\mathbf{a})=Q(s_j|s',\mathbf{a}) \quad \text{for all states } s_j, s, s' \text{ and all action profiles } \mathbf{a}. \tag{15}$$

Finally, we call a memory-1 strategy P state-independent if

$$P(s_i,\mathbf{a}) = P(s_j,\mathbf{a}) \quad \text{for all states } s_i, s_j \text{ and all action profiles } \mathbf{a}. \tag{16}$$

There is the following relationship between state-independent memory-1 strategies and stochastic games with state-independent and deterministic transitions.

Proposition 1. *Consider a stochastic game between players with memory-one strategies $P_1,\ldots,P_n$.*

1. *Suppose that for a given action profile $\mathbf{a}$ we have $Q(s_j|s,\mathbf{a})=0$ for all states $s\in S$ (i.e., if the players have chosen action profile $\mathbf{a}$, it is impossible to reach state s_j in the next round). Then the transition matrix M, and hence the players' payoffs, are independent of the values of $P_k(s_j,\mathbf{a})$ for all players k.*
2. *In particular, if the stochastic game has state-independent and deterministic transitions, then for any memory-1 strategy P_i there is an associated state-independent memory-1 strategy P_i' such that the payoffs of all players are unchanged if P_i is replaced by P_i'.*

All proofs are provided in the appendix of this SI. For state-independent and deterministic games, Proposition 1 is useful for two reasons. First, Proposition 1 allows us to reduce the dimension of the strategy space: without loss of generality we can restrict our attention to players with state-independent memory-1 strategies. In case of symmetric memory-1 strategies $\mathbf{p} = (p^i_{a,j})$, we can thus drop the dependence on the previous state s_i, and write $\mathbf{p} = (p_{a,j})$. If m is the number of states and n the number of group members, the new strategy space is only $2n$-dimensional (instead of $2mn$-dimensional). Second, Proposition 1 allows us to better compare a stochastic game with an associated repeated game in which players always remain in the same state s_i. In each case, the relevant set of symmetric memory-one strategies has the same dimension $2n$. Thus if cooperation evolves for the stochastic game (but not in the repeated games), this difference cannot be attributed to differences in the complexity of the two strategy sets.

3.2 Comparing the evolutionary dynamics of stochastic games and repeated games

In **Fig. 2** of the main text, we have considered the evolutionary dynamics of a state-independent stochastic game with two states for either 2 players (**Fig. 2c**) or n players (**Fig. 2b**) . The transition function for this stochastic game can be represented by the vector $\mathbf{q} = (q_n, q_{n-1}, \ldots, q_0) = (1, 0, \ldots, 0)$. We have compared this game with the two associated repeated games in which players always remain in the same state. Using our framework, these repeated games can be represented by the transition functions $\mathbf{q} = (1, 1, \ldots, 1)$ (players always remain in state 1, independent of the number of cooperators) and $\mathbf{q} = (0, 0, \ldots, 0)$ (players never visit state 1). Due to Proposition 1, we have used symmetric and state-independent memory-1 strategies, $\mathbf{p} = (p_{C,n-1}, \ldots, p_{C,0};\ p_{D,n-1}, \ldots, p_{D,0})$. **Fig. 2** shows simulation results of the process described in Section 2.3, assuming that players only apply pure strategies with errors, such that each entry $p_{a,j}$ is either ε or $1-\varepsilon$. This assumption leads to a finite strategy space of size 2^{2n}. For finite strategy spaces we can use the method of Fudenberg and Imhof[54] to calculate exact strategy frequencies in the limit of rare mutations. **Fig. S2a** shows the corresponding cooperation frequencies in the selection-mutation equilibrium for $n=2$ as a function of the benefit parameter b_1. The stochastic game always yields higher cooperation rates than the two repeated games, but the difference is most pronounced for intermediate benefits to cooperation in State 1. If b_1 is too small, cooperation evolves in none of the scenarios; if b_1 is sufficiently large, cooperation can already be achieved through repeated interactions in State 1 alone.

To show that our results are robust when players have access to stochastic memory-one strategies, we have run additional simulations using the method of Imhof and Nowak[56]. The corresponding simulation results shown in **Fig. S2b** are qualitatively similar to our results for pure memory-one strategies. Again, cooperation rates increase as the benefit b_1 of cooperation in State 1 increases, but for the stochastic game smaller values of b_1 are necessary to sustain substantial cooperation.

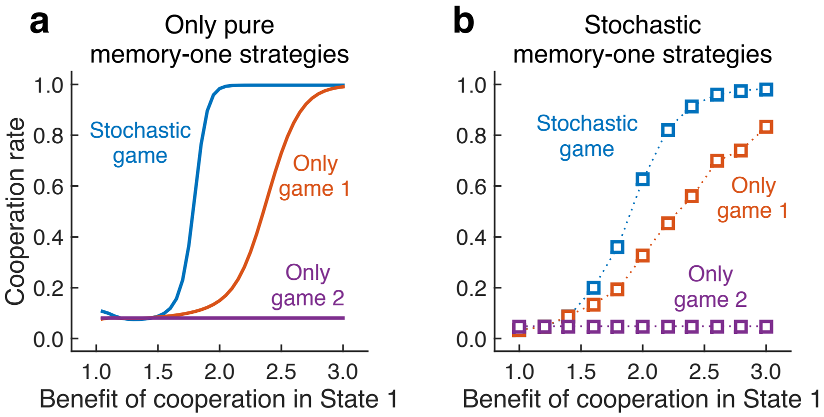

Figure S2: The stochastic game is most favorable to cooperation when the benefit-to-cost ratio in game 1 is intermediate. a, For the 2-player game depicted in **Fig. 2**, we have used the method of Fudenberg and Imhof[54] to calculate exact cooperation frequencies for various values of b_1 when players are restricted to pure memory-one strategies with rare errors. The advantage of the stochastic game is most pronounced when this benefit is intermediate, $1.5 \leq b_1 \leq 2.5$. **b,** To explore whether stochastic strategies allow for different outcomes, we have run simulations according to the protocol of Imhof and Nowak[56]. The two sets of results are in good qualitative agreement. However, stochastic memory-one strategies lead to somewhat smoother transitions between almost full defection and almost full cooperation as the parameter b_1 increases along the x-axis. All parameters are the same as in **Fig. 2a** of the main text: $N = 100$, $b_2 = 1.2$, $c = 1$, $\beta = 1$, $\varepsilon = 0.001$, in the limit of no discounting $\delta \to 1$.

3.3 Robustness with respect to changes in evolutionary parameters

For the results shown so far, we have kept several model parameters at a fixed value. In the following, we aim to demonstrate that our results are robust with respect to changes in these parameters. To this end, we have re-calculated the evolving cooperation rates for the example discussed in **Fig. 2a** and **Fig. S2**, and independently varied the following four parameters:

1. The error rate ε, which determines how often players misimplement their intended move (in the examples shown in the main text, we have used $\varepsilon = 0.001$).

2. The discount rate δ, which measures how relevant future rounds are for calculating a player's overall payoff (in the main text we have considered the limiting case of no discounting, $\delta \to 1$).

3. The strength of selection parameter β, which determines how strongly players take a strategy's payoff into account when updating their strategies (in the main text we have used an intermediate value $\beta = 1$).

4. The mutation rate μ, which gives the rate at which players randomly explore new strategies (to obtain exact numerical results, the data plotted in **Fig. S2** has been derived for the limit of rare mutations, $\mu \to 0$).

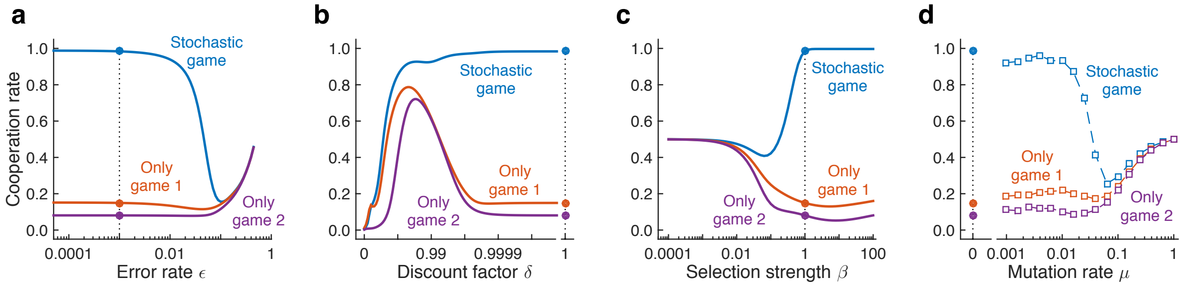

Figure S3: The presented findings are robust with respect to various parameter changes. To test the robustness of our findings, we consider the stochastic game introduced in **Fig. 2a** of the main text, and we independently vary four different parameters, (**a**) the error rate ε, (**b**) the discount factor δ, (**c**) the strength of selection β, and (**d**) the mutation rate μ. Solid lines indicate exact results using the method of Fudenberg and Imhof[54] in the limit of rare mutations, whereas square symbols are used to represent simulation results. The dotted line in each panel indicates the respective parameter value that we have used in **Fig. S2a**. In all four panels, the stochastic game leads to more cooperation than each of the repeated games in isolation, independent of the exact value of the parameter under consideration. All unspecified parameters are the same as in **Fig. S2a**.

Results for the error rate ε are shown in **Fig. S3a**, suggesting there are three parameter regimes. For large error rates $\varepsilon \to 1$, cooperation decisions occur essentially at random, and the overall cooperation rate approaches 50%, independent of the specific game being played. For smaller error rates, $0.1 \leq \varepsilon \leq 1$, mistakes occur too often to allow a targeted punishment of $AllD$ players, and hence defectors prevail in the stochastic game and in the two associated repeated games. Once $\varepsilon \leq 0.1$, we recover the results from the main text, showing that stochastic games can generate cooperation although the associated repeated games cannot.

The effect of the discount factor on evolving cooperation rates is depicted in **Fig. S3b**. Also here, we can identify three parameter regions. When δ is small, cooperation cannot be sustained at all because there is a too high discount on future payoffs to incentivize cooperation in the present round. When δ is intermediate $(0.9 \leq \delta \leq 0.99)$, substantial cooperation can be achieved in all three games. In this parameter region, we observe that the most abundant strategy is $Grim$, which can sustain cooperation as long as the expected number of rounds $1/(1-\delta)$ is small compared to the time $1/\varepsilon$ at which one of the players can be expected to defect by mistake. If δ is beyond that threshold, cooperation can only be sustained in the stochastic game for the given parameter values.

The effects of the selection parameter β and the mutation rate μ are similar (**Fig. S3c,d**). When β is small or when μ is comparably large, the evolutionary process is mainly governed by noise, such that the evolving cooperation rates approach 50%. As β increases and μ decreases, a strategy's performance becomes increasingly important for the strategy's survival, eventually favoring cooperation in the stochastic game and defection in the two repeated games. Overall, **Fig. S3** suggests that the results presented in the main text are reasonably robust. When the transition structure of the stochastic game itself incentivizes cooperation, then the stochastic game leads to higher evolving cooperation rates than

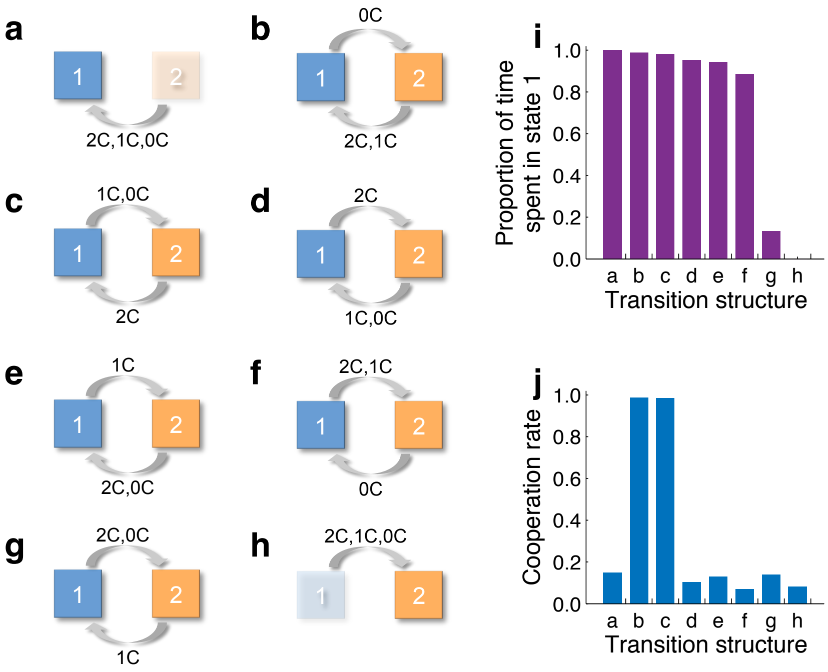

Figure S4: Whether cooperation evolves in 2-player games critically depends on the form of the environmental feedback. Keeping the payoffs fixed at the values used in **Fig. 2a**, we have explored how the evolution of cooperation depends on the underlying transition structure of the stochastic game. **a – h,** We have calculated the selection-mutation equilibrium for all possible stochastic games with two states when transitions are state-independent and deterministic. **i**, Overall, 6 of the 8 transition structures lead players to spend more time in the more profitable State 1, in which mutual cooperation has a higher benefit (as $b_1 > b_2$). **j,** However, only in two out of these six transition structures cooperation evolves. These two structures have in common that mutual cooperation always leads to the beneficial State 1, whereas mutual defection leads to the detrimental State 2. Thus cooperation is most likely to evolve if the environmental feedback itself incentivizes mutual cooperation and disincentivizes mutual defection. The transitions after unilateral defection play a less prominent role.

the associated repeated games, irrespective of the exact value of the other parameters.

3.4 Effect of transition structure on the evolution of cooperation

In the previous figures, we have compared evolving cooperation rates for three different transition structures, $\mathbf{q} = (1,0,0)$ (the "stochastic game"), $\mathbf{q} = (1,1,1)$ ("Only game 1") and $\mathbf{q} = (0,0,0)$ ("Only game 2"). To explore the role of transitions more systematically, we have calculated the cooperation rate in the mutation-selection equilibrium for all $2^3 = 8$ games with state-independent and deterministic transitions $\mathbf{q} = (q_2, q_1, q_0)$, with $q_i \in \{0,1\}$. The results are shown in **Fig. S4**. We find that in 6 out of the 8 cases, players succeed in predominantly being in the first state, in which cooperation is more beneficial as $b_1 > b_2$ (**Fig. S4i**). However, only in two out of these cases, players actually achieve substantial cooperation rates (**Fig. S4j**). These two stochastic games, depicted in **Fig. S4b,c**, have in common that

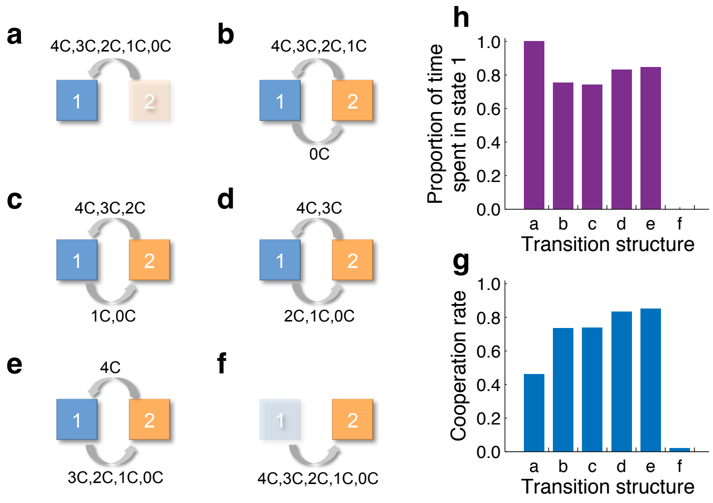

Figure S5: Impact of transitions on cooperation in 4-player public goods games. Here we explore the role of different transition structures on the evolution of cooperation for a stochastic game with two states (with a PGG being played in each state). State 1 is again more beneficial since $r_1 > r_2$, but to be in State 1 it takes a minimum number k of cooperators in the previous round. **a – f,** For a four-player PGG, there are 6 possible monotonic configurations of the stochastic game, as k can be any number between 0 (players always move to first state) and 5 (players never move to first state). **h,** There is a non-monotonic relationship between the 6 transition structures and the time spent in the more beneficial State 1. **g,** The evolving cooperation rate becomes maximal when any deviation from mutual cooperation leads players to State 2 (case e). Parameters are as in **Fig. 2b**, with the multiplication factor in the first state being fixed to $r_1 = 2$; to derive exact results, we have applied the framework of Fudenberg and Imhof[54], considering the limit of rare mutations $\mu \to 0$.

mutual cooperation leads to the beneficial State 1, whereas mutual defection leads to the inferior State 2 (that is, $q_2 = 1$ and $q_0 = 0$). When these two conditions are satisfied, cooperation can evolve independent of the transition after unilateral cooperation (that is, independent of q_1).

Similarly, we have also explored how the evolving cooperation rates depend on the exact shape of the evolutionary feedback when a group of $n > 2$ players is engaged in a public good game. In **Fig. S5**, we consider all “monotonic” transition functions, for which players find themselves in the more beneficial State 1 if at least k of the players have cooperated in the previous round, with $k \in \{0, \ldots, n+1\}$. Formally, these are exactly the state-independent and deterministic transition functions $\mathbf{q} = (q_n, \ldots q_0)$ for which $i < j$ implies $q_i \leq q_j$. As expected, cooperation is most prevalent when the stochastic game exhibits the most strict response to single players defecting. Evolving cooperation rates are highest if already one defector is sufficient to let the group move towards the less beneficial State 2 (**Fig. S5f**). Interestingly, all stochastic games in which players can find themselves in both states (**Fig. S5b–e**) lead to higher cooperation rates than the two repeated games in which players always remain in the same state (**Fig. S5a,f**).

3.5 Numerical analysis of the evolving strategies

So far we have seen that for natural transition structures, stochastic games can favor the evolution of cooperation. In this section, we aim to understand this process on the level of the evolving strategies. To this end, we have again considered all eight state-independent and deterministic stochastic games between two players. For these games, we have recorded how often the 16 memory-one strategies $\mathbf{p}=(p_{C,1},p_{C,0},p_{D,1},p_{D,0})$ with $p_{a,j}\in\{\varepsilon,1-\varepsilon\}$ are played in the selection-mutation equilibrium. We let the benefit of cooperation in the first state vary between $1\le b_1\le 3$. To be able to make comparisons across strategy sets of different size, we have normalized these frequencies. To this end, let $\lambda_\beta(\mathbf{p})$ be the frequency of strategy $\mathbf{p}$ in the selection-mutation equilibrium when the strength of selection is β. We define the strategy's relative abundance by $\lambda_\beta(\mathbf{p})/\lambda_0(\mathbf{p})$, where $\lambda_0(\mathbf{p})$ is the abundance of $\mathbf{p}$ under neutral selection. We call a strategy *favored by selection* if $\lambda_\beta(\mathbf{p})/\lambda_0(\mathbf{p})>1$, i.e., if the strategy is more abundant than expected under neutrality.[60,61]

Fig. S6 shows the relative abundance for the two most important strategies $AllD$ and $WSLS$ across the eight different stochastic games (together with the strategy $Grim$, which plays similar to $AllD$ in an infinitely repeated game with errors, these three strategies are typically played for more than 80% of the evolutionary time). We note that only in four out of the eight stochastic games in **Fig. S6**, almost full cooperation can be achieved as the benefit of cooperation in State 1 approaches $b_1=3$ (this observation remains true if the value of b_1 was further increased). The four cases in which cooperation can evolve are exactly those cases in which mutual cooperation leads the players to remain in the more profitable State 1. Interestingly, the evolution of cooperation is strongly tied to the success of $WSLS$; in all four stochastic games, cooperation evolves exactly when $WSLS$ is favored by selection.

In **Fig. S7**, we demonstrate that a similar result also applies in groups of more than two players. Again, the most abundant strategy is either $AllD$ (when the multiplication factor r_1 is too low for cooperation to evolve) or $WSLS$ (when r_1 is sufficiently large).

3.6 Analytical conditions for the stability of cooperation

The previous numerical results suggest that we can predict when full cooperation can evolve based on the stability properties of $WSLS$. In this subsection, we therefore aim to explore when $WSLS$ is an equilibrium, depending on the game's transition function and on the payoffs of the stage game. The following Proposition considers arbitrary state-independent games with m states and n players.

Proposition 2. *Consider a state-independent stochastic game between n players with discount rate $0<\delta<1$. Let $Q(s_j|k)\in[0,1]$ denote the probability to move to state $s_j\in\{s_1,\ldots,s_m\}$ after a round in which k players have cooperated. The strategy $WSLS$ is a subgame perfect equilibrium if and*

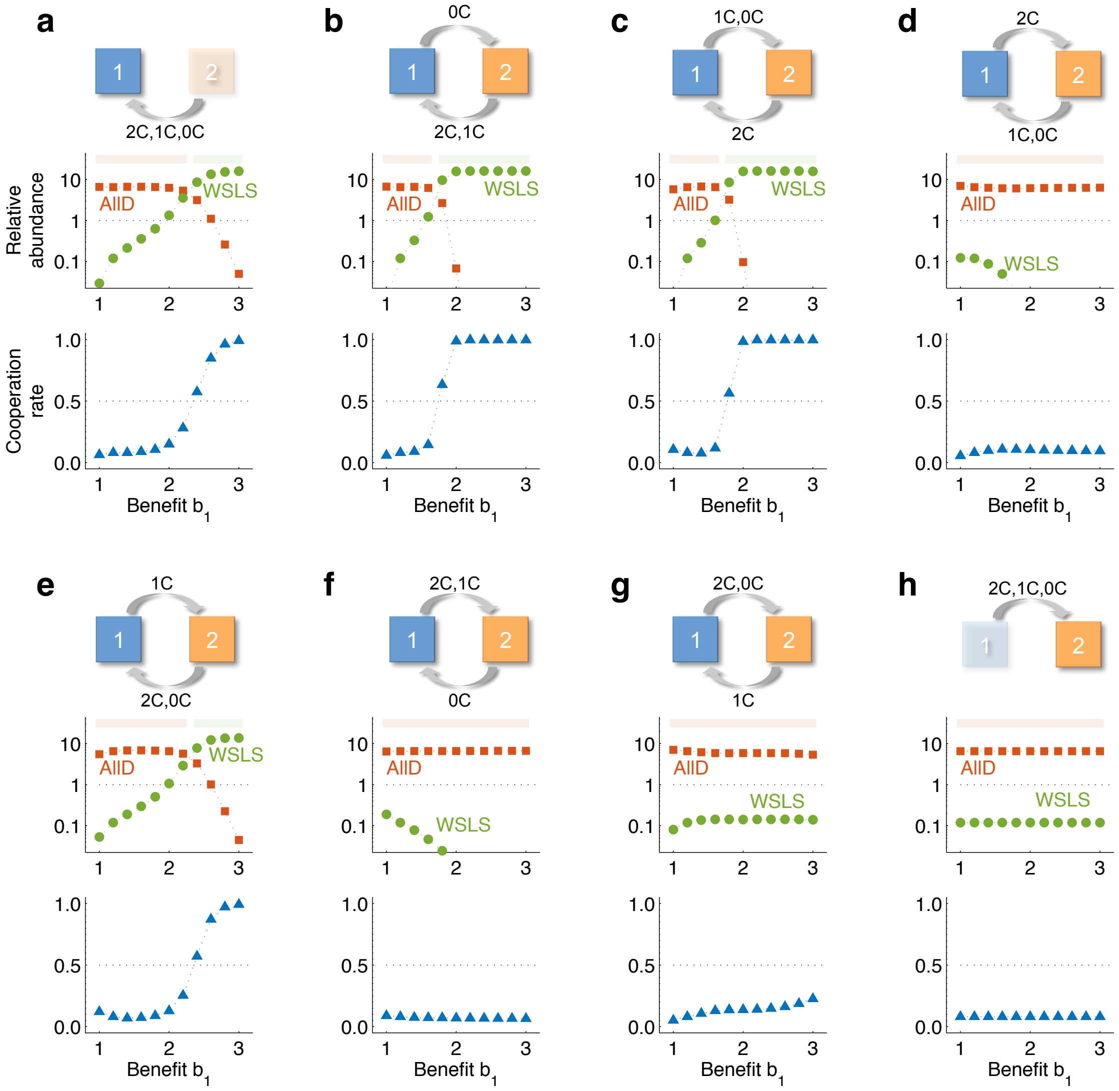

Figure S6: An analysis of the evolving strategies for all state-invariant and deterministic stochastic games with two states and two players suggests that the evolution of cooperation hinges on the success of WSLS. For each of the eight games depicted in **Fig. S4**, we have recorded the evolving cooperation rate (lower subpanel) and the relative abundance of each memory-one strategy (upper subpanel) for different values of b_1. For clarity, we only depict two memory-one strategies explicitly, $AllD$ and $WSLS$; the color-shaded bars on top of the upper panels show parameter regimes in which either $AllD$ or $WSLS$ is most abundant among all 16 strategies. In four out of the eight cases, we observe that full cooperation can evolve as the benefit to cooperation in State 1 approaches $b_1 = 3$. These are exactly the cases in which mutual cooperation leads players towards the more beneficial State 1. Moreover, in these four cases the upper subpanels show that cooperation emerges due to the success of WSLS, which is the predominant strategy whenever cooperation prevails. Except for the value of b_1, all other parameter values are the same as in **Fig. S4**.

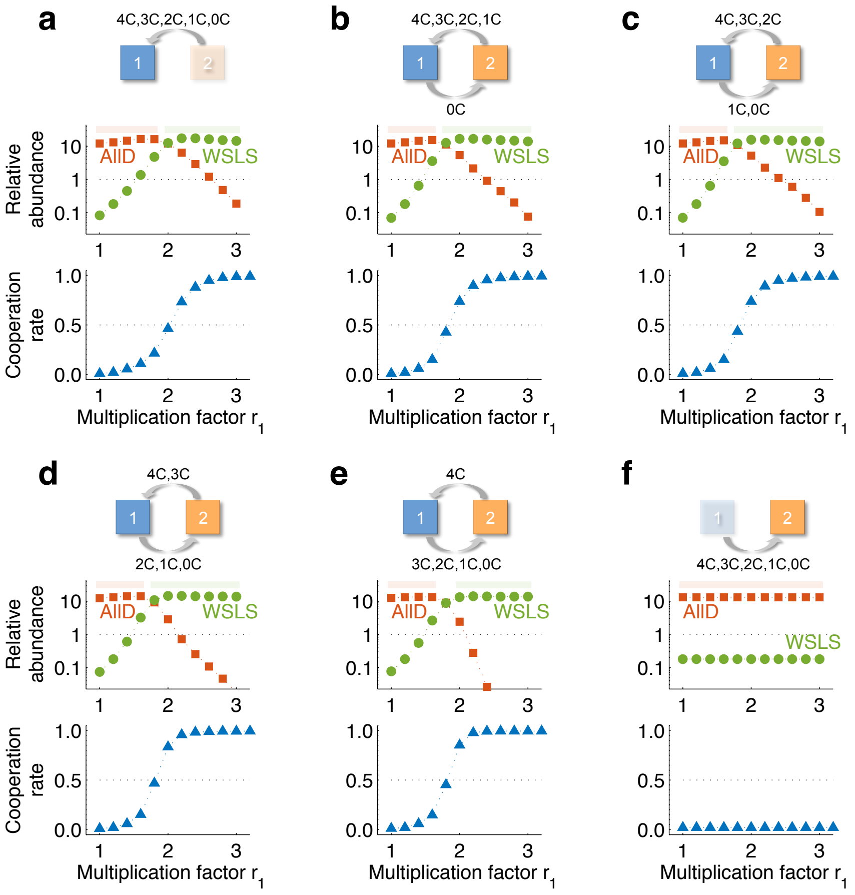

Figure S7: Win-Stay Lose-Shift sustains cooperation in multiplayer public goods games. This figure represents the analogue of **Fig. S6** for the case of multiplayer interactions. Again, we show evolving cooperation rates and the relative abundance of $AllD$ and $WSLS$ for the six state-independent and deterministic games in which transitions are monotonic. In five of these games, cooperation emerges once the multiplication factor r_1 becomes sufficiently large; in all of those $WSLS$ is the most abundant strategy when cooperation evolves. Except for r_1 all parameters are the same as in **Fig. S5**.

only if the following conditions are met in all states s_i*,*

$$\begin{aligned} u^i_{C,n\text{–}1} - u^i_{D,n\text{–}1} + \delta \sum_{j=1}^{m} \Big(Q(s_j|n)u^j_{C,n\text{–}1} - Q(s_j|n-1)u^j_{D,0}\Big) + \delta^2 \sum_{j=1}^{m} \Big(Q(s_j|n) - Q(s_j|0)\Big)u^j_{C,n\text{–}1} &\geq 0. \\ u^i_{D,0} - u^i_{C,0} + \delta \sum_{j=1}^{m} \Big(Q(s_j|0)u^j_{C,n\text{–}1} - Q(s_j|1)u^j_{D,0}\Big) + \delta^2 \sum_{j=1}^{m} \Big(Q(s_j|n) - Q(s_j|0)\Big)u^j_{C,n\text{–}1} &\geq 0. \end{aligned} \tag{17}$$

It is worth to note that the above result neither requires the stochastic game to be deterministic, nor that deviating players choose among the memory-1 strategies. In particular, it implies that if conditions (17) are satisfied, then $WSLS$ is a Nash equilibrium, and no single mutant can have a higher payoff than the residents. Due to continuity, this result remains true in the limiting case of no discounting, $\delta \to 1$, provided that the payoff of the mutant and of $WSLS$ is well-defined (which always holds when the mutant applies a memory-k strategy for some finite k).

For the special case of a pairwise game where a prisoner's dilemma is played in each state, $n=2$, $u^i_{C,1}=b_i-c$, $u^i_{C,0}=-c$, $u^i_{D,1}=b_i$, $u^i_{D,0}=0$, condition (17) simplifies to

$$\delta(1-\delta)\sum_{j=1}^{m} Q(s_j|n)\cdot b_j - \delta^2\sum_{j=1}^{m} Q(s_j|0)\cdot b_j \geq (1+\delta)c \tag{18}$$

In particular, if we assume without loss of generality that states are ordered such that $b_1 > b_2 > \ldots > b_m$, condition (18) is most easily satisfied if $Q(s_1|n)=1$ and $Q(s_m|0)=1$. Thus, we recover the result that cooperation is most likely to evolve if the stochastic game itself reflects the players' behavior, with mutual cooperation always leading to the best state and mutual defection leading to the worst. Moreover, it follows from (18) that all intermediate transition probabilities $Q(s_j|k)$ with $0<k<n$ are irrelevant for the stability of $WSLS$. Even for general games, condition (17) suggests that only the four transitions $Q(s_j|n)$, $Q(s_j|n-1)$, $Q(s_j|1)$ and $Q(s_j|0)$ affect the stability of $WSLS$. In the limiting case $\delta\to 1$, condition (18) becomes

$$\sum_{j=1}^{m}\Big(2Q(s_j|n)-Q(s_j|0)\Big)\cdot b_j \geq 2c. \tag{19}$$

In the case of only two states, $m=2$, we can write the transitions as $Q(s_1|k)=q_k$ and $Q(s_2|k)=1-q_k$. In that case, condition (19) further simplifies to

$$\big(2q_2-q_0\big)\cdot b_1 + \big(1-(2q_2-q_0)\big)\cdot b_2 \geq 2c. \tag{20}$$

This is condition (1) in the main text.

Similarly, we can analyze the stability of $WSLS$ if the game played in each state is an n-player public goods game with decreasing multiplication factors $r_1 > r_2 > \ldots > r_m$. In that case, condition (17) translates into

$$\delta(1-\delta)\sum_{j=1}^{m} Q(s_j|n)\cdot r_j - \delta^2\sum_{j=1}^{m} Q(s_j|0)\cdot r_j \geq 1-r_m/n+\delta, \tag{21}$$

which for $\delta\to 1$ becomes

$$\sum_{j=1}^{m}\Big(2Q(s_j|n)-Q(s_j|0)\Big)\cdot r_j \geq 2-r_m/n. \tag{22}$$

In the case of games with two states only, this condition reads

$$\big(2q_n-q_0\big)\cdot r_1 + \big(1-(2q_n-q_0)\big)\cdot r_2 \geq 2-r_m/n, \tag{23}$$

For the parameter values used in **Fig. S7b–e** ($n=4$, $r_2=1.2$, $q_n=1$, $q_0=0$), this condition becomes $r_1 \geq 1.45$, which is roughly the value where we observe $WSLS$ to become favored by selection.

3.7 An example with many states

In **Fig. 4** of the main text, we have introduced a stochastic game with $n=4$ players and $m=5$ states. In each state s_i players interact in a public goods game with multiplication factor r_i and cost $c=1$. States are ordered such that $r_1 \geq \ldots \geq r_m$. Transitions are deterministic and state-independent. Players move towards State 1 if all four players have cooperated in the previous round, they move towards State 2 if three players have cooperated, etc. That is, for all previous states s_i we have

$$\begin{aligned} &Q(s_i,4)=(1,0,0,0,0), \quad Q(s_i,3)=(0,1,0,0,0), \quad Q(s_i,2)=(0,0,1,0,0),\\ &Q(s_i,1)=(0,0,0,1,0), \quad Q(s_i,0)=(0,0,0,0,1). \end{aligned} \tag{24}$$

Due to Proposition 1, we can thus restrict ourselves to state-independent memory-1 strategies of the form

$$\mathbf{p}=(p_{C,3},p_{C,2},p_{C,1},p_{C,0};\ p_{D,3},p_{D,2},p_{D,1},p_{D,0}), \tag{25}$$

where $p_{a,j}$ is a player's cooperation probability if in the previous round the focal player used action a and j of the co-players have cooperated. Players are assumed to apply pure strategies with errors, such that $p_{a,j} \in \{\varepsilon, 1-\varepsilon\}$. In total, there are $2^8=256$ such strategies. For the payoffs in each state we have considered four scenarios. These scenarios differ in how the multiplication factors depend on the number of previous defectors. Specifically, the multiplication factors r_i for each scenario are defined as follows (see also **Fig. 4c**) :

	State 1	State 2	State 3	State 4	State 5
Scenario with immediate consequences	1.6	1	1	1	1
Scenario with gradual consequences	1.6	1.45	1.3	1.15	1
Scenario with delayed consequences	1.6	1.6	1.6	1.6	1
Scenario with no consequences	1.6	1.6	1.6	1.6	1.6.

We have assumed there is no discounting on the future, $\delta \to 1$. The model was analyzed by calculating exact strategy frequencies and cooperation rates in the selection-mutation equilibrium as mutations become rare[54].

We observe that only the first three scenarios yield substantial cooperation rates (**Fig. 4**), among which the scenario with immediate consequences yields the highest cooperation rate. This can be understood on the level of evolving strategies. Applying condition (22) to this example, we find that $WSLS = (1,0,0,0;\ 0,0,0,1)$ is stable in the first three scenarios, whereas it is unstable in the last scenario. This stability result is reflected in strategy abundances in the selection-mutation equilibrium. $WSLS$ is most abundant in the first three scenarios, whereas $ALLD$ is the most abundant strategy in the last scenario. As a consequence, although in each state the multiplication factor in the scenario with immediate consequences is the lowest across all scenarios, players in this scenario earn the highest payoffs (the payoffs are 0.537, 0.459, 0.401, and 0.089, respectively). Intuitively, although $WSLS$ is an equilibrium in all of the first three scenarios, the payoff of a single $ALLD$ mutant in a $WSLS$ population is lowest when consequences are immediate. It follows that also the fixation probability of $ALLD$ in a $WSLS$ population becomes minimal in this scenario.

3.8 Exploring the impact of different payoff constellations on the game dynamics

In most of our analysis so far we have explored whether stochastic games allow for cooperation even if cooperation yields lower payoffs in each stage game. In the case of pairwise games with two states, we have thus assumed that players interact in a prisoner's dilemma in any given round. In the following, we weaken this assumption.

We consider a stochastic game with two players and two states. As in the main text, players employ pure memory-1 strategies subject to rare errors (with the error rate again being $\varepsilon = 0.001$). For the game in State 2, we suppose players face the same prisoner's dilemma game as in **Fig. 2a**. However, for the game in State 1, we allow more general payoff configurations. Specifically, we assume that the payoffs in game 1 are given by the following matrix,

$$U^1 = \begin{pmatrix} 1 & S^1 \\ T^1 & 0 \end{pmatrix}, \tag{26}$$

where S^1 and T^1 are the sucker's payoff and the temptation payoff in the first state, respectively. Depending on the values of T^1 and S^1, we can distinguish four possible cases:

(PD) If $T^1 > 1$ and $S^1 < 0$, game 1 is a *prisoner's dilemma*. Mutual defection is the only equilibrium in the corresponding one-shot game.

(SH) If $T^1 < 1$ and $S^1 < 0$, game 1 corresponds to a *stag-hunt game*. In that case, the one-shot game corresponds to a coordination game with three equilibria, both cooperate, both defect, and a mixed equilibrium.

(SD) If $T^1 > 1$ and $S^1 > 0$, game 1 corresponds to a *snowdrift game*. In the one-shot case, this game has one symmetric equilibrium according to which players randomize between cooperation and defection.

(HG) If $T^1 < 1$ and $S^1 > 0$, mutual cooperation is the only equilibrium. This game is thus sometimes called *harmony game*.

We are interested in how the evolution of cooperation and the dynamics of play is affected by the payoff values and the game's transition structure. To this end, we have systematically varied the two

payoff parameters S^1 and T^1, and we have considered all eight state-independent and deterministic transitions $\mathbf{q} = (q_2, q_1, q_0)$ with $q_i \in \{0, 1\}$. For each combination of S^1, T^1 and $\mathbf{q}$, we calculate exact strategy frequencies in the selection-mutation equilibrium as the mutation rate becomes rare[54]. Based on these strategy frequencies, we compute (*i*) how often players cooperate on average and (*ii*) how often they switch between states over the course of a game (i.e., in which fraction of rounds players move from State 1 to State 2 or vice versa). The corresponding results are shown in **Fig. S8**.
We observe the following regularities:

1. High cooperation rates are only feasible in those stochastic games in which mutual cooperation leads to the first state (i.e., for which $q_2 = 1$, as in **Fig. S8a,b,c,e**). In those four cases, players tend to be fully cooperative in the harmony game and in the stag-hunt game. In the prisoner's dilemma, players are most likely to establish cooperation if the sucker's payoff S^1 is large and the temptation T^1 is small. Moreover, cooperation in the prisoner's dilemma is more likely to evolve when mutual defection leads players to the second state (as in **Fig. S8b,c**). In this regime, the most abundant strategy is Win-Stay Lose Shift, which is in line with the results reported in Section 3.6. Finally, for the snowdrift game we observe that players may learn to cooperate alternatingly, provided that unilateral cooperation leads to the more profitable State 1 (**Fig. S8a,b**). Otherwise, if unilateral cooperation leads to the second state, players are either fully cooperative or fully defective (**Fig. S8c,e**).

   Interestingly, in the four cases in which mutual cooperation leads to the first state, players rarely switch between the two states, independent of the particular game being played in State 1. That is, players either adopt strategies that make them stay in the first state, or they adopt strategies that make them stay in the second state.

2. The only cases when players learn to alternate between the two states occur when mutual cooperation leads to the second state whereas mutual defection leads to the first state (**Fig. S8d,f**). In those two cases there are regions in the (T^1, S^1)–plane where players cooperate in the first game and defect in the second. These switches between states resemble the oscillating tragedy of the commons reported by Weitz *et al*[10]. In their analysis of one-shot games with environmental feedback, they have observed persistent oscillations as the game payoffs switch between a harmony game and a prisoner's dilemma (see also **Appendix A**). In contrast, **Fig. S8d,f** suggests that similar oscillations can occur in stochastic games even if the first game is a stag-hunt game, provided that the sucker's payoff S^1 is sufficiently large.

3. Finally, if mutual cooperation and mutual defection both lead to the unprofitable second state, cooperation rarely evolves (**Fig. S8g,h**). Only if the game under consideration is a snowdrift game and if unilateral cooperation leads to the first state, players may sometimes learn to cooperate asynchronously (**Fig. S8g**).

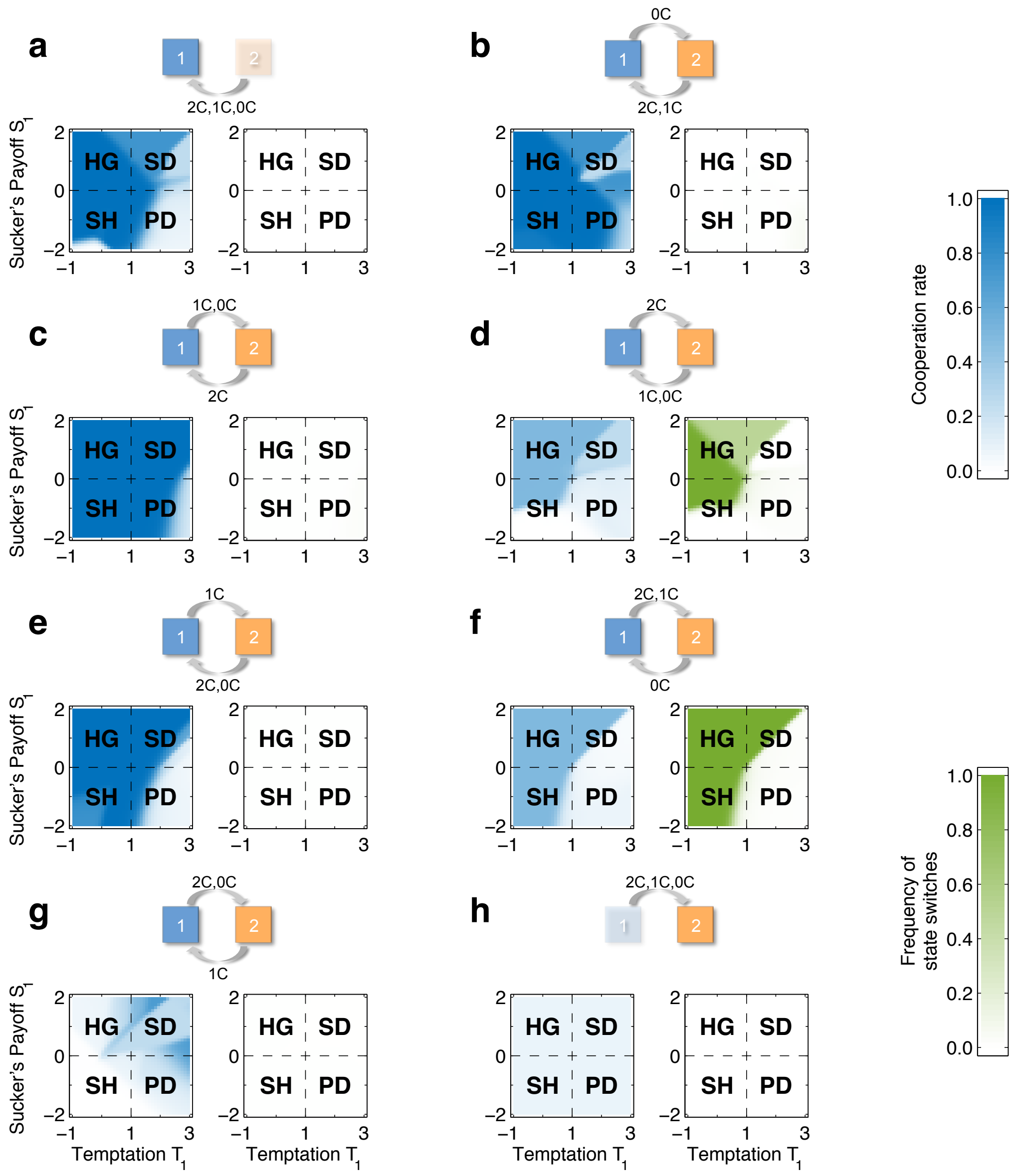

Figure S8: A systematic analysis of the expected game dynamics for different game payoffs. For each of the eight possible state-independent transitions **q**, we have systematically varied the temptation payoff T^1 (x-axis) and the sucker's payoff S^1 (y-axis) in the first state. For each combination of T^1, S^1, and **q** we have computed how often players cooperate in the selection-mutation equilibrium (left panel) and in which fraction of rounds they switch from one state to the other (right panel). Depending on T^1 and S^1, game 1 can be one of four different types: harmony game (HG), snowdrift game (SD), stag-hunt game (SH), and prisoner's dilemma (PD). **a,b,c,e,** Full cooperation can evolve when players find themselves in State 1 after mutual cooperation. **d,f** Players learn to switch between states only when mutual cooperation leads to State 2 and mutual defection leads to State 1. **g,h** In the remaining cases, players hardly cooperate. Parameters: The payoffs in game 2 are the same as in **Fig. 2a**, a prisoner's dilemma with $b_2 = 1.2$ and $c = 1$. For the evolutionary parameters we have considered population size $N = 100$ and selection strength $\beta = 1$.

4 Further applications

In the following, we discuss several further examples where the framework of stochastic games can be used to shed light on cooperation in specific scenarios. All examples have in common that they are state-dependent: transitions do not only depend on the players' previous actions but also on the present state. Examples 1–3 explore the impact of probabilistic transitions on cooperation; the fourth example illustrates the dynamics in a state-dependent stochastic game with more than two states.

4.1 A stochastic game with probabilistic return to State 1

Here we consider the stochastic game illustrated in **Fig. 3a** of the main text. There are two players, $\mathcal{N} = \{1, 2\}$ and the game has two states $S = \{s_1, s_2\}$. Players remain in the first state after mutual cooperation, but they move to the second state if at least one player defects. Once players are in the second state, they return to the first state after mutual cooperation with probability q; otherwise they remain in the second state. That is, the transition function is given by

$$\mathbf{q} = (q_2^1, q_1^1, q_0^1;\; q_2^2, q_1^2, q_0^2) := (1, 0, 0;\; q, 0, 0), \tag{27}$$

where q_k^i is the probability to move to the first state, given that the present state is s_i and that k players have cooperated. In each state i, players interact in a prisoner's dilemma; the payoffs are $u_{CC}^i = b_i - c$, $u_{CD}^i = -c$, $u_{DC}^i = b_i$, and $u_{DD}^i = 0$. Cooperation in State 1 is more profitable, since $b_1 > b_2 > c$. Players are assumed to use memory-one strategies; that is, admissible strategies are 5-tuples of the form

$$\mathbf{p} = (p_{CC}^1;\; p_{CC}^2, p_{CD}^2, p_{DC}^2, p_{DD}^2). \tag{28}$$

The entries $p_{a\tilde{a}}^i$ represent a player's cooperation probability in state s_i, given that the focal player's previous action was a and that the co-player's action was $\tilde{a}$. Note that we can omit the cooperation probabilities p_{CD}^1, p_{DC}^1 and p_{DD}^1 due to Proposition 1 (players never find themselves in the first state if a player has defected in the previous round). Moreover, since we consider the limit of no discounting on future payoffs, $\delta \to 1$, we can also omit a player's initial cooperation probability p_0. In the following, we assume that players use pure strategies with errors. That is, the entries are taken from the set $p_{a\tilde{a}}^i \in \{\varepsilon, 1-\varepsilon\}$. Hence, the strategy space contains $2^5 = 32$ different strategies. For given parameter values b_1, b_2, c, q, we use the method of Fudenberg and Imhof[54] to calculate the abundance of each strategy in the selection-mutation equilibrium. As we have illustrated in **Fig. 3a**, there are parameter combinations in which it takes an intermediate value of q to achieve maximum payoffs in the population.

Here, we discuss this result in terms of the strategies that evolve. For the most abundant strategy, we find there are three different regimes depending on the transition probability q (see **Fig. S9c**). When q is close to zero, the most abundant strategy is $AllD = (0; 0, 0, 0, 0)$; as q increases, the most abundant strategy becomes $WSLS = (1;\; 1, 0, 0, 1)$; and for sufficiently large values of q, the most abundant strategies is a more ambitious version of win-stay lose-shift, which we call $AWSLS = (1;\; 0, 0, 0, 1)$.

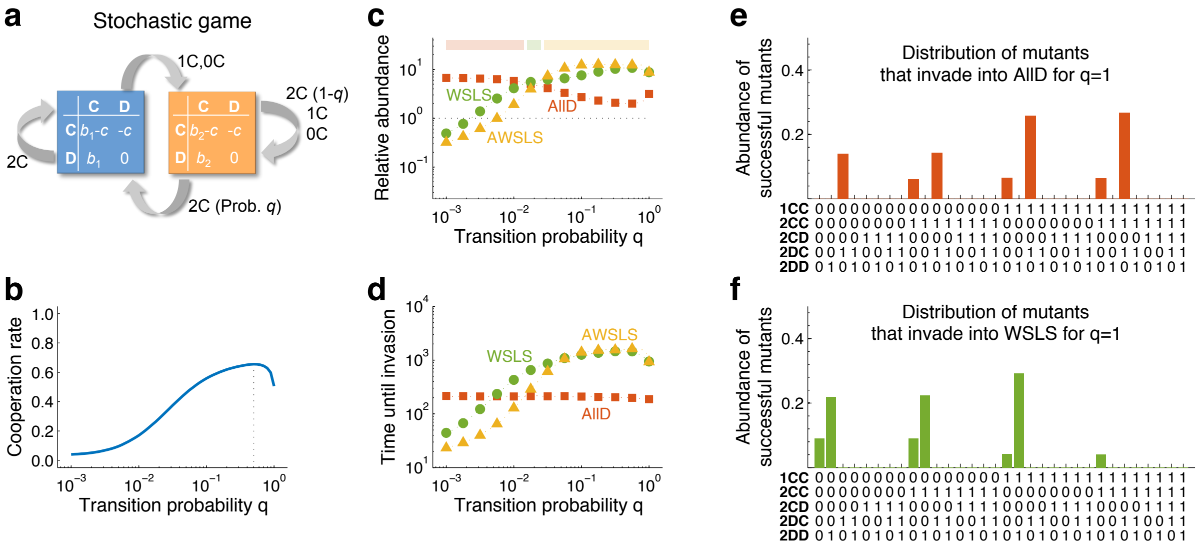

Figure S9: Probabilistic transitions can further enhance cooperation. a, We consider a stochastic game in which any defection leads to State 2. After mutual cooperation in State 1, players remain in State 1 with certainty. After mutual cooperation in State 2, players move towards State 1 with probability q. **b,** Calculating the cooperation rate in the selection-mutation equilibrium in the limit of rare mutations shows that the highest cooperation rate is achieved for intermediate values of q. **c,** We have recorded the abundance of all 32 memory-one strategies in the selection-mutation equilibrium. The most abundant strategy is either $AllD$ (for small values of q, as indicated by the red squares), $WSLS$ (for small but positive values of q, green circles), or $AWSLS$ (for all other q-values, yellow triangles). **d,** To estimate the time it takes each resident strategy to be invaded, we have randomly introduced other mutant strategies, and we have recorded how long it takes until a mutant successfully fixes. To get a reliable estimate, we have performed 10,000 runs for each resident strategy. **e,f,** In addition, we have also recorded which strategy eventually reaches fixation if the resident either applies $AllD$ or $WSLS$ when $q\!=\!1$. Parameters: $b_1\!=\!1.9$, $b_2\!=\!1.4$, $c\!=\!1$, $\beta\!=\!1$, $N\!=\!100$.

An $AWSLS$ player continues with her previous action if and only if the previous payoff was at least b_1-c; otherwise she takes the opposite action in the next round (an ordinary $WSLS$ player follows the same principle, but uses the more modest threshold b_2-c).

To explore why the highest cooperation rate evolves for intermediate values of q (**Fig. S9b**), we have simulated how long it takes other mutant strategies to invade one of these three resident populations. For a resident population consisting of $AllD$ players, we find that it takes approximately 200 mutant invasions until the first mutant successfully reaches fixation, with little dependence on q (**Fig. S9d**). $WSLS$ and $AWSLS$ are less robust for small values of q. For $q = 0$, $AWSLS$ is on average invaded after 17 mutants, and $WSLS$ fares only slightly better, surviving on average 24 mutants. This is intuitive – for $q\!=\!0$ the second state of the stochastic game becomes absorbing. Hence, the stochastic game degenerates to a repeated game in the second state, in which the benefit b_2 is too low for $WSLS$ or $AWSLS$ to be stable. As q increases, both strategies become stable, which is also reflected by a considerably longer invasion time. In the limiting case $q = 1$, the stochastic game becomes state-independent, and the two strategies $WSLS$ and $AWSLS$ are equivalent (since they only differ in their value of p^2_{CC} which now is irrelevant). For $q=1$, we hence find for both strategies that it takes approximately 930 mutant invasions until a mutant takes over the population.

To explore further why the value for the optimal transition probability is smaller than one, we have also recorded which strategies typically succeed in invading $AllD$ (**Fig. S9e**) and $WSLS$ (**Fig. S9f**) when $q=1$. $AllD$ is typically invaded by TFT-like strategies. Successful mutants defect if the co-player has defected, and they cooperate if the co-player has cooperated. In this way, TFT can be shown to have a selective advantage among $AllD$ players; its invasion probability is above the neutral probability $1/N$ for all values of q. On the other hand, for $WSLS$ we find that the most successful mutant usually applies $AWSLS$, and vice versa $AWSLS$ is typically invaded by $WSLS$ (since they both are neutral to each other when $q=1$). However, the second most successful mutant strategy turns out not to be $AllD$, but rather the slightly more cooperative strategy $\mathbf{p} = (0;\ x, 0, 0, 1)$ with $x \in \{0, 1\}$ arbitrary (**Fig. S9f**). For $x = 0$, the payoff of such a strategy in a $WSLS$ population can be calculated as

$$\pi = \frac{b_2 - c + qb_1 + (1-q)b_2}{3} + \mathcal{O}(\varepsilon) \tag{29}$$

Hence, for $q = 1$, this strategy achieves the higher benefit b_1 in every third round (while it receives the mutual defection payoff 0 and the mutual cooperation payoff $b_2 - c$ in the other two rounds). By decreasing the value of q, the mutant's payoff π is reduced, which explains why smaller values of q enhance the stability of $WSLS$ and $AWSLS$.

For q close to one, we observe that $WSLS$ is slightly less abundant than $AWSLS$ (**Fig. S9c**). When the population consists of a mixture of these two strategies, everyone gets the mutual cooperation payoff b_1-c as the error rate goes to zero $\varepsilon \to 0$. However, for positive error rates, the competition between $WSLS$ and $AWSLS$ becomes a coordination game, in which $AWSLS$ is risk-dominant[57]. Thus, when q is close to but below one, we find that $AWSLS$ is favored by selection in a $WSLS$ population (it has a fixation probability above $1/N$), whereas $WSLS$ is disfavored in an $AWSLS$ population.

We have also explored for which payoffs probabilistic transitions are particularly favorable to cooperation. To this end, we have simultaneously varied the benefit in the second state $b_2 \in [1, 2]$, and we have varied how much the first state is preferred, $b_1 - b_2 \in [0, 1]$, see **Fig. S9**. We observe that for probabilistic transitions to be beneficial, the value of b_2 should be intermediate, and the relative advantage of the first state $b_1 - b_2$ should decrease as b_2 increases. Intuitively, if b_2 and $b_1 - b_2$ are too small, cooperation is unlikely to evolve even with probabilistic transitions. On the other hand, as either b_2 or $b_1 - b_2$ become large, almost full cooperation can already be established with deterministic transitions.

Overall, this example illustrates how probabilistic transitions can further enhance the prospects of cooperation. When the transition towards State 1 depends on exogenous chance events, it takes on average more time to reach the more beneficial state after one player has defected. This renders deviations from mutual cooperation more costly.

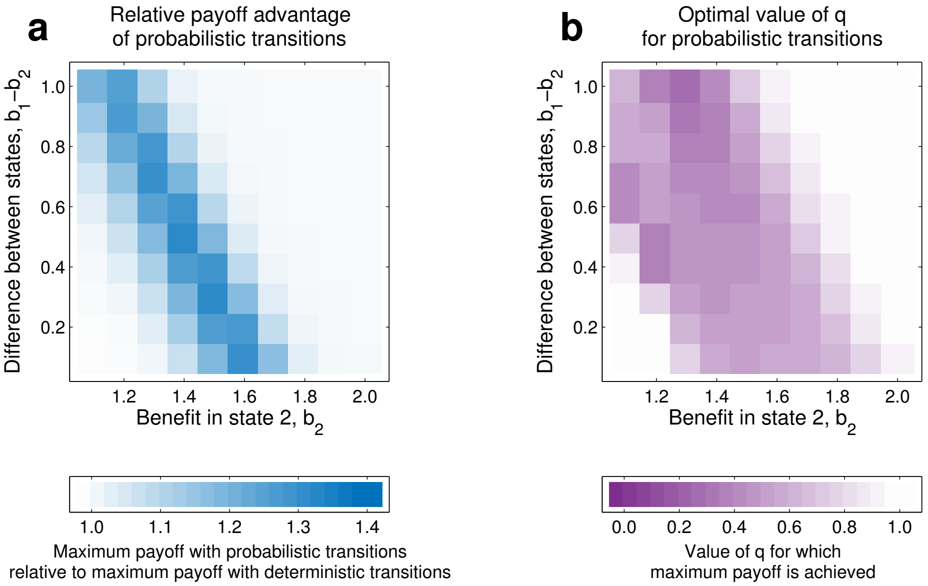

Figure S10: Probabilistic transitions are favorable for intermediate benefit values. For this graph, we have varied the benefit of cooperation in the second state (x-axis) and the relative payoff advantage of the first state (y-axis). We have compared two scenarios: the payoff that players obtain in the selection-mutation equilibrium when players always return to the first state after mutual cooperation ($q = 1$), and the maximum payoff that players obtain when q can take arbitrary values between 0 and 1. The figure shows **a,** the relative payoff advantage of probabilistic transitions, and **b,** the value of q that guarantees the highest payoff to the population. Both panels indicate that probabilistic transitions are most favorable if b_2 is between 1 and 1.8, and b_1 is between 1.8 and 2.2 (such that there is an inverse relationship between b_2 and $b_1 - b_2$). Except for the benefit values, all parameters are the same as in the previous figure.

4.2 A repeated game with endogenous stopping time

As another example, we have considered a scenario in which defections of a player may lead a repeated game to end earlier (as introduced in **Fig. 3b** of the main text). Again there are two players $\mathcal{N} = \{1, 2\}$ and two states $S = \{s_1, s_2\}$. After mutual cooperation, players remain in the first state. However, if one of the players defects, players move to the second state with probability q. State 2 is absorbing. Hence, the transition function is represented by the vector

$$\mathbf{q} = (q_2^1, q_1^1, q_0^1;\ q_2^2, q_1^2, q_0^2) := (1, 1-q, 1-q;\ 0, 0, 0), \tag{30}$$

While in State 1, players interact in a prisoner's dilemma with payoffs $u_{CC}^1 = b_1 - c$, $u_{CD}^1 = -c$, $u_{DC}^1 = b_1$, and $u_{DD}^1 = 0$. In State 2, there are no profitable interactions and $u_{a\tilde{a}}^2 = 0$, irrespective of the players' actions. As the second state is absorbing, this stochastic game is only meaningful for discount factors $\delta < 1$ (otherwise the payoff of any strategy with errors will eventually approach 0). Thus, memory-one

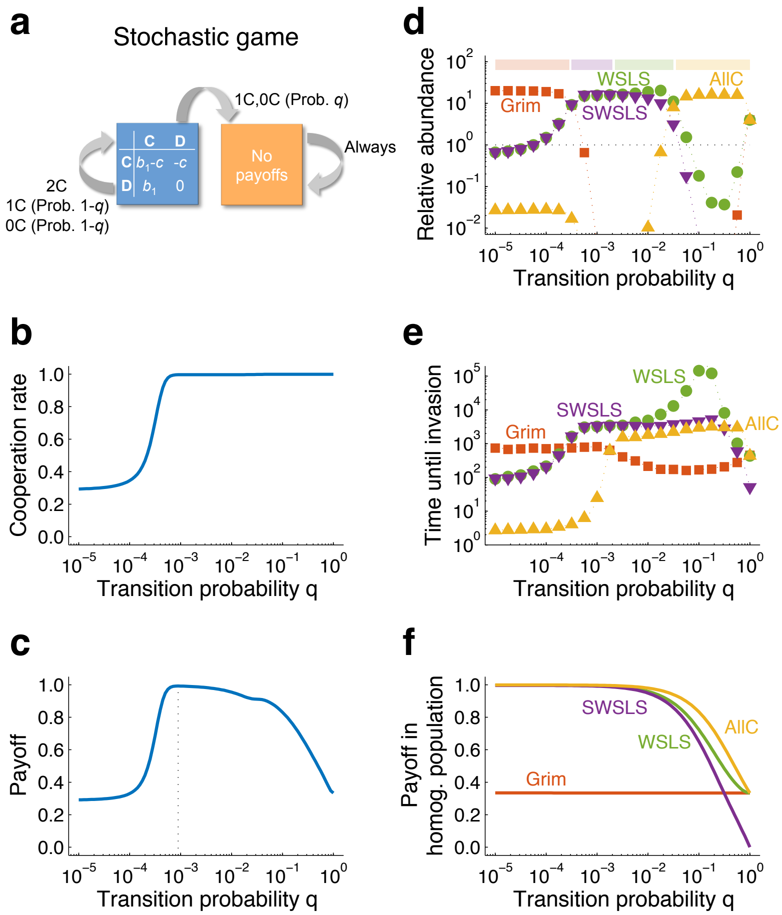

Figure S11: Players benefit from a small endogenous risk that the game stops early. a, We consider a stochastic game in which players remain in State 1 after cooperation, but move towards State 2 with transition probability q if one of the player defects. In State 2 no profitable interactions are possible, the game is over. **b,** According to our evolutionary simulations, a higher transition probability leads to more cooperation. **c,** However, a higher probability q also makes players move to the second state if one of the players merely defected due to an error; hence the dependence of payoffs on q is non-monotonic. **d,e,** When q is small, $Grim$ is the predominant strategy. As q increases, $WSLS$ strategies take over. As q approaches one, unconditional cooperation becomes most successful. **f,** For the given parameter values, a homogeneous $Grim$ population only achieves 1/3 of the maximum payoff possible, since any error leads to relentless defection. The other three strategies get the maximum payoff $b_1 - c$ for $q=0$, but this payoff decreases with q. Parameters: $b_1=2$, $c=1$, $\delta=0.999$, $\varepsilon=0.001$, $\beta=1$, $N=100$.

strategies for this example have the following form,

$$\mathbf{p} = (p_0;\ p^1_{CC}, p^1_{CD}, p^1_{DC}, p^1_{DD}). \qquad (31)$$

Again, we focus on pure memory-one strategies with errors, yielding $2^5 = 32$ different strategies in total.

In the main text (**Fig. 3b**) we have shown that evolving payoffs reach an optimum for intermediate values of the transition probability q. Herein, we aim to explain this finding on the level of evolving strategies. To this end, we again consider the limit of small mutation rates and we use the method of Fudenberg and Imhof[54] to calculate, for each value of q, how abundant each strategy is in the selection-mutation equilibrium. For small values of q, we find that $Grim = (1; 1, 0, 0, 0)$ is the memory-1 strategy most favored by selection (**Fig. S11d**) . However, in $Grim$ populations, cooperation is unstable: once a player defects by error, cooperation breaks down and from then on both players receive a payoff of zero (independent of whether they remain in State 1 or move towards State 2). We can use Eq. (12) to calculate payoffs in a homogeneous $Grim$ population. For $q = 0$ we obtain

$$\pi_G = \frac{1 - \delta(1-\varepsilon)(1 + \delta\varepsilon(1-2\varepsilon))}{1 - \delta(1-2\varepsilon)}(b_1 - c), \tag{32}$$

which for our parameter values yields $\pi_G \approx 0.334 \cdot (b_1 - c)$ (see also **Fig. S11f**) . That is, homogeneous $Grim$ populations only achieve an average cooperation rate of 33%. The predominance of $Grim$ for $q \approx 0$ thus explains the relatively low cooperation rates (**Fig. S11b**) and payoffs (**Fig. S11c**) in the selection-mutation equilibrium.

As q increases, the two strategies $WSLS = (1; 1, 0, 0, 1)$ and its more suspicious counterpart $SWSLS = (0; 1, 0, 0, 1)$ become favored by selection. To get some intuitive insight into why these strategies emerge, we use Eq. (12) to calculate when $WSLS$ becomes stable. The respective condition reads

$$q \geq 1 - \sqrt{\frac{\delta b_1 - c}{\delta^2(b_1 - c)}} + \mathcal{O}(\varepsilon). \tag{33}$$

For our parameter values, the expression under the square root is approximately one, and hence the critical q value is of the order of ε, which explains the rise of $WSLS$ in **Fig. S11**d as the transition probability approaches $q = 0.001$. For such q, the payoff of homogeneous $WSLS$ populations is close to the optimal $b_1 - c$ (**Fig. S11e**) , and hence we also see sharp increase of population payoffs (**Fig. S11c**) .

As the transition probability q increases even further, cooperative strategies become increasingly successful. Using Eq. (12) we find that *AllC* can resist invasion of *AllD* if

$$q \geq \frac{1-\delta}{\delta}\frac{c}{b_1 - c} + \mathcal{O}(\varepsilon). \tag{34}$$

However, despite the success of cooperative strategies, average payoffs begin to diminish once q passes a critical threshold (**Fig. S11c**) . Even if all players are unconditional cooperators, occasional failures to cooperate will often terminate the interaction prematurely, with negative effects on the expected payoffs (**Fig. S11f**) .

Overall, this example illustrates the mechanisms that promote cooperation if the length of a game is not exogenously given, but endogenously determined by the players. If defective relationships have a higher risk to be terminated, even unconditional cooperation can become a stable outcome, provided that future interactions are sufficiently valuable.

4.3 A model with time-out

As our next application, we consider a model with time-out (illustrated in **Fig. 3c** of the main text). There are again two players, $\mathcal{N} = \{1, 2\}$, and two states, $S = \{s_1, s_2\}$. Players remain in State 1 after mutual cooperation and they move towards State 2 otherwise. Once in State 2, they return to State 1 with probability q, independent of the players' actions. The transition function is thus

$$\mathbf{q} = (q_2^1, q_1^1, q_0^1;\ q_2^2, q_1^2, q_0^2) := (1, 0, 0;\ q, q, q), \tag{35}$$

In State 1, players interact in a conventional prisoner's dilemma with payoffs $u_{CC}^1 = b_1 - c$, $u_{CD}^1 = -c$, $u_{DC}^1 = b_1$, and $u_{DD}^1 = 0$. In the time-out State 2, payoffs are zero $u_{a\tilde{a}}^2 = 0$ for all $a, \tilde{a} \in \{C, D\}$. We can obtain analytical results for the dynamics when we focus on games without discounting ($\delta \to 1$) and when we consider players that only base their action in State 1 on the previous state. They use the regular cooperation probability x if the previous state was s_1, and they use the cooperation probability y after returning to state 1 from the time-out state 2. We can represent such strategies (x, y) in our framework by assuming without loss of generality that players always defect when in State 2. in that case, the strategy (x, y) can be encoded as the memory-one strategy

$$\mathbf{p} = (p_{CC}^1, p_{CD}^1, p_{DC}^1, p_{DD}^1;\ p_{CC}^2, p_{CD}^2, p_{DC}^2, p_{DD}^2) = (x, 0, 0, y;\ 0, 0, 0, 0). \tag{36}$$

When an (x_1, y_1)-mutant interacts with an (x_2, y_2)-resident, we can thus use Eq. (12) to calculate player 1's expected payoff,

$$\pi_1 = q \cdot \frac{(1 - x_1 x_2 + x_2 y_1) y_2 b_1 - (1 - x_1 x_2 + x_1 y_2) y_1 c}{(1 + q)(1 - x_1 x_2) - q y_1 y_2}. \tag{37}$$

To model the evolution of strategies in the population, we assume that the population is of infinite size, and that emerging mutant strategies are close to the respective resident strategy. Considering the adaptive dynamics[62] of the system, we assume that the direction of evolution points towards the mutant with highest invasion fitness. That is, the time evolution of some resident population (x, y) is given by the dynamical system

$$\dot{x} = \frac{\partial \pi_1}{\partial x_1}\bigg|_{x_1 = x_2 = x,\ y_1 = y_2 = y} \quad \text{and} \quad \dot{y} = \frac{\partial \pi_1}{\partial y_1}\bigg|_{x_1 = x_2 = x,\ y_1 = y_2 = y} \tag{38}$$

Plugging Eq.(37) into (38), we obtain the system

$$\begin{aligned} \dot{x} &= q y^2 \cdot \frac{x(x + qx - qy) b_1 - (1 + q - xy + y^2) c}{\big((1 + q)(1 - x^2) - q y^2\big)^2}, \\ \dot{y} &= q(1 - x)^2 \cdot \frac{((1 + q) xy - q y^2) b_1 - (1 + q)(1 - x^2 + xy) c}{\big((1 + q)(1 - x^2) - q y^2\big)^2}, \end{aligned} \tag{39}$$

which is defined on the unit square $[0,1]^2$. In **Fig. 3c**, we show a phase portrait of this system for $b_1 = 3$, $c = 1$ and $q = 1/2$. By solving for $\dot{x} = 0$ and $\dot{y} = 0$, we find a unique interior fixed point which is always on the main diagonal,

$$Q = \left(\sqrt{\frac{(1+q)c}{b_1}}, \sqrt{\frac{(1+q)c}{b_1}}\right). \tag{40}$$

In **Fig. 3c**, the fixed point Q is represented by a purple dot. By linearizing the system around Q, we find that the trace of the Jacobian matrix is positive, and hence Q is unstable.[63] A numerical analysis shows that the global dynamics is bistable. Some orbits converge to the line segment $(x, 0)$ with $0 \leq x < 1$; this line comprises all strategies that defect against themselves. Otherwise orbits converge towards the line segment $(1, y)$ with $0 < y \leq \min\{1, \hat{y}\}$, where

$$\hat{y} = \frac{-q(b_1 - c) + \sqrt{q(b_1 - c)(4c + 3qc + qb_1)}}{2cq}. \tag{41}$$

In particular, if $q \leq (b_1 - c)/c$ we find that $AllC = (1, 1)$ is a stable fixed point of the dynamics. Thus, the longer players stay in the time-out state in expectation, the more likely it becomes that $AllC$ becomes a stable equilibrium. As in the previous example, however, the average payoff in an $AllC$ population with rare errors is a decreasing function of q. Hence, for the evolution of high payoffs in the population it is again optimal that q is sufficiently small (increasing the size of the basin of attraction of cooperation), but positive (preventing the second state with no payoffs to be absorbing).

4.4 A stochastic game with delayed payoff feedback

So far we have considered examples in which players revise their actions at the same timescale at which payoff changes occur. However, in many natural applications it may be more realistic to assume there is some delay in the payoff consequences of an action. For example, when farmers overgraze a shared pasture, the consequences may not be visible after a day but only after a week. Such delayed consequences can be captured by introducing additional states to the stochastic game. In the following, we illustrate this technique with a simple example.

We consider a state-dependent stochastic game with two players $\mathcal{N} = \{1, 2\}$ and three states, $S = \{s_1, s_2, s_3\}$, see **Fig. S12a**. Mutual cooperation always leads to the neighboring state with lower index (unless players are already in state s_1, in which case they stay there after mutual cooperation). Mutual defection always leads to a state with higher index (unless players already find themselves in the third state). If only one player defects, players remain in the same state. Hence, the transition function Q takes the following values

$$\begin{array}{lll}
Q\big(s_1, (C, C)\big) = (1, 0, 0), & Q\big(s_2, (C, C)\big) = (1, 0, 0), & Q\big(s_3, (C, C)\big) = (0, 1, 0) \\
Q\big(s_1, (C, D)\big) = (1, 0, 0), & Q\big(s_2, (C, D)\big) = (0, 1, 0), & Q\big(s_3, (C, D)\big) = (0, 0, 1) \\
Q\big(s_1, (D, C)\big) = (1, 0, 0), & Q\big(s_2, (D, C)\big) = (0, 1, 0), & Q\big(s_3, (D, C)\big) = (0, 0, 1) \\
Q\big(s_1, (D, D)\big) = (0, 1, 0), & Q\big(s_2, (D, D)\big) = (0, 0, 1), & Q\big(s_3, (D, D)\big) = (0, 0, 1).
\end{array} \tag{42}$$

Given these transitions, it is impossible that players find themselves in s_1 after mutual defection, or that they find themselves in s_3 after mutual cooperation. Due to Proposition 1, memory-1 strategies are therefore given by 10-tuples,

$$\mathbf{p} = (p^1_{CC}, p^1_{CD}, p^1_{DC};\ p^2_{CC}, p^2_{CD}, p^2_{DC}, p^2_{DD};\ p^3_{CD}, p^3_{DC}, p^3_{DD}). \tag{43}$$

As in the first two examples, we consider pure strategies with errors, such that $p^i_{a\tilde{a}} \in \{\varepsilon, 1-\varepsilon\}$. In total, there are $2^{10} = 1,024$ such strategies. For the payoffs in each state we consider four scenarios that are similar to the scenarios in Section 3.7. In all scenarios and all states, cooperation leads to a cost c for the cooperator. The benefits of cooperation depend on the scenario and on the state as follows (see also **Fig. S12a**):

	State 1	State 2	State 3
Scenario with immediate consequences	b_1	c	c
Scenario with gradual consequences	b_1	$\frac{b_1+c}{2}$	c
Scenario with delayed consequences	b_1	b_1	c
Scenario with no consequences	b_1	b_1	b_1.

Again, we analyze this model by calculating exact strategy frequencies and cooperation rates in the selection-mutation equilibrium as the mutation rate becomes rare[54].

As shown in **Fig. S12b**, immediate payoff consequences are most favorable to cooperation for all considered benefit values $c < b_1 \leq 3c$. As a consequence, this scenario also yields the highest payoffs, given that benefits exceed a moderate threshold, $b_1 \geq 1.4$ (**Fig. S12c**). However, even when payoff consequences are delayed, cooperation rates and payoffs are still higher than in the case when there are no payoff consequences at all. On the level of evolving strategies, we observe that across all scenarios and all benefit values, five different strategies are most abundant:

$$\begin{array}{rcl}
\mathbf{p_{72}} & = & (0,0,0;\ 1,0,0,1;\ 0,0,0), \\
\mathbf{p_{577}} & = & (1,0,0;\ 1,0,0,0;\ 0,0,1), \\
\mathbf{p_{585}} & = & (1,0,0;\ 1,0,0,1;\ 0,0,1) \\
\mathbf{p_{586}} & = & (1,0,0;\ 1,0,0,1;\ 0,1,0) \\
\mathbf{p_{588}} & = & (1,0,0;\ 1,0,0,1;\ 1,0,0)
\end{array} \tag{44}$$

The index of each memory-one strategy indicates the decimal representation of the binary strategy entries. The strategy $\mathbf{p_{72}}$ is self-defective; if applied by everyone in the population, players find themselves in the third state most of the time, in which they mutually defect on each other. In contrast, the four other

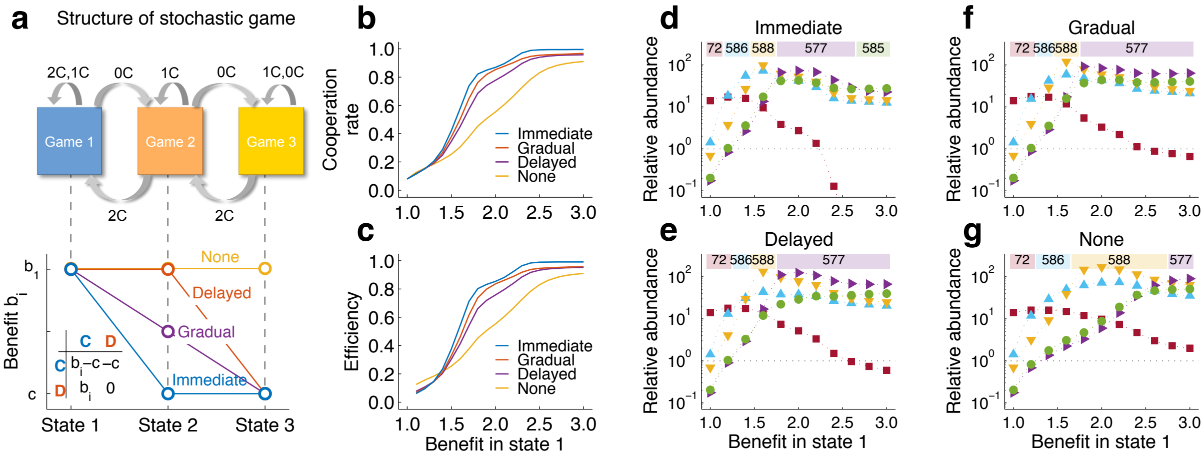

Figure S12: Immediate environmental feedback enhances cooperation. a, We consider a state-dependent stochastic game with three states. Mutual cooperation always leads players to move to a superior state (or to remain in the most beneficial state s_1). Similarly, mutual defection always leads to an inferior state (or players remain in the most detrimental state s_3). After a unilateral defection, players remain in the same state. We consider four different versions of this game, depending on how quickly the payoffs decrease as players move towards an inferior state. **b,** Our numerical results show that an immediate negative response of the environment to defection is most favorable to the evolution of cooperation. **c,** As a consequence, the scenario with immediate consequences also yields the highest average payoffs once the benefit in State 1 exceeds a moderate threshold. **d–g,** On the level of evolving strategies, we find that an immediately responding environment is most favorable to the evolution of $WSLS$ strategies, and strongly selects against defecting strategies. Again, the colored bars on top of each panel indicate the strategy that is most favored by selection for the respective value of b_1. Parameters: $c=1$; b_1 varies between 1 and 3; b_2 is equal to c, $(b_1+c)/2$ and b_1; and b_3 is either equal to c or to b_1 depending on the considered scenario (as depicted in panel a). $N=100$, $\beta=1$, $\delta\to 1$, $\varepsilon=0.001$.

strategies are self-cooperative. Strategy $\mathbf{p}_{585}$ is $WSLS$; the other three strategies represent $WSLS$ with minor adaptations. In **Fig. S12d–g**, we show how well these strategies fare in the four different scenarios considered. When the benefit b_1 is relatively low, we note that the $WSLS$-like strategies evolve most readily in the scenario with immediate consequences of defection.

To check whether our results are sensitive to the assumption that the benefit in the third state was set to c, we have computed the resulting cooperation rates for an alternative scenario. In this alternative scenario, we again set $b_1=1.8$ as in **Fig. S12**, but the benefit in the third state is $b_3 = 1.4$ for the three scenarios in which payoffs change (again, b_2 is either equal to 1.8, 1.6, or 1.4). In that case we obtain the cooperation rates 62.8% (immediate consequences), 57.0% (gradual consequences), 52.6% (delayed consequences) and 45.3% (no consequences, which is the same as before). Again, an immediate feedback proves best for cooperation, but even some delayed feedback can prove beneficial.

The question of delayed payoff feedback may become more consequential if players heavily discount the future. While the numerical results shown in **Fig. S12** assume the limiting case of no discounting, **Fig. S13** systematically explores how the evolving cooperation rates depend on the discount rate and on

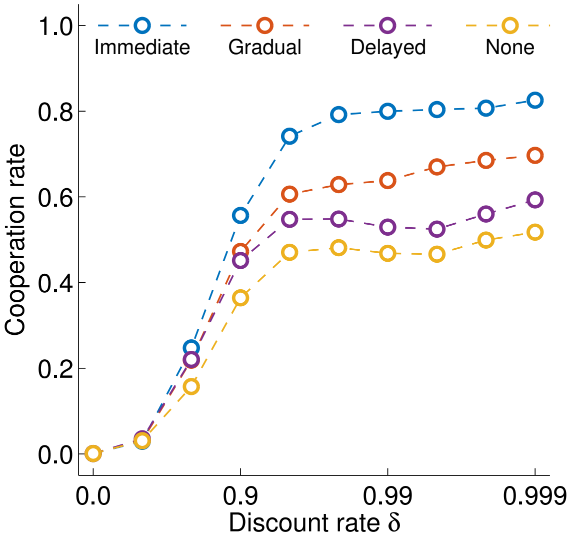

Figure S13: Cooperation in stochastic games requires that players sufficiently take future payoff consequences into account. We have repeated the numerical computations in **Fig. S12** for various discount rates δ. When players entirely focus on the present, such that $\delta = 0$, cooperation evolves in none of the four treatments. As players increasingly take future payoffs into account, cooperation rates increase. Immediate payoff feedback is most conducive to cooperation across all considered values of δ. Parameters are the same as in **Fig. S12**.

the exact shape of the payoff feedback. As one may expect, none of the four treatments allows for the evolution of cooperation when players only value their present payoffs, such that $\delta \to 0$. As we increase the extent to which players take future payoffs into account, all four treatments exhibit more cooperation. The relative ranking of the treatments is the same as before: immediate payoff consequences lead to the highest cooperation rates and no payoff consequences exhibit the lowest cooperation rates. However, the differences between the four treatments are more distinct, as compared to **Fig. S12**. We conclude that stochastic games are most conducive to cooperation if players put sufficient weight on future payoffs, and if the time lag between the players' actions and the resulting payoff feedback is sufficiently short.

On a more technical level, this example highlights that our framework is not restricted to simple state-dependent games with two states only. Instead, it is often possible to calculate exact strategy frequencies in games with multiple states, especially if transitions are deterministic (such that one can often use Proposition 1 to reduce the size of the strategy space).

Appendix A: On the feasibility of cooperation in one-shot social dilemmas

With our study we aim to highlight that the interplay of reciprocity and payoff feedback can be crucial to sustain cooperation. To this end, **Fig. 2** of the main text has illustrated the critical role of payoff feedback: in these examples, players learn to cooperate in the stochastic game (where payoff feedback is present), whereas they predominantly defect in the two associated repeated games (without payoff feedback). In this appendix we wish to illustrate the critical role of repeated interactions: cooperation cannot be sustained if there is only payoff feedback but players lack a long term perspective. Players need to take the future consequences of their actions into account for cooperation to evolve.

To make this point more rigorous, we revisit the model of Weitz *et al*[10] (to allow for an easy comparison, we adopt their notation in the following). They consider an infinite and well-mixed population. Players interact in a pairwise game with two possible actions, C and D. In contrast to our setup, the game is not repeated; if a player is a cooperator, this player is assumed to play C independent of the other players' previous decisions and independent of any environmental cues. The frequency of cooperators in the population is denoted by $x \in [0,1]$. The players' payoffs depend on their own actions, on the frequency of cooperators x, and on their current environment. The current state of the environment is represented by the continuous variable $n \in [0,1]$. The payoff matrix $A(n)$ for a given environment n is assumed to be a linear combination of two borderline cases (see their Eq. **[18]**),

$$A(n) = \begin{pmatrix} R(n) & S(n) \\ T(n) & P(n) \end{pmatrix} := (1-n) \begin{pmatrix} R_0 & S_0 \\ T_0 & P_0 \end{pmatrix} + n \begin{pmatrix} R_1 & S_1 \\ T_1 & P_1 \end{pmatrix}. \tag{45}$$

Using the above payoff matrix, the payoffs of the two player types are defined as

$$\begin{aligned} r_1\big(x, A(n)\big) &= xR(n) + (1-x)S(n), \\ r_2\big(x, A(n)\big) &= xT(n) + (1-x)P(n). \end{aligned} \tag{46}$$

To model how cooperation and the environment co-evolve, they consider the following two-dimensional version of the replicator equation (see their Eq. **[20]**),

$$\begin{aligned} \dot{x} &= x(1-x)\Big[r_1\big(x, A(n)\big) - r_2\big(x, A(n)\big)\Big], \\ \dot{n} &= \varepsilon n(1-n)\big[-1 + (1+\theta)x\big]. \end{aligned} \tag{47}$$

The parameter $\varepsilon > 0$ reflects the relative speed at which the environment changes, as compared to the strategy dynamics. The parameter $\theta > 0$ reflects the recovery of the environment when cooperators are common (i.e., when $x \approx 1$). According to the first equation, the fraction of cooperators increases if and only if cooperators receive a higher payoff than defectors, such that $r_1\big(x, A(n)\big) > r_2\big(x, A(n)\big)$. According to the second equation, the environmental parameter increases if and only if there are sufficiently many cooperators, such that $x > 1/(1+\theta)$. Weitz *et al*[10] observe that if cooperation is a dominant strategy for $n = 0$ and if defection is a dominant strategy for $n = 1$, the dynamics according to (47) can exhibit persistent oscillations.

We note that the replicator equation (47) considers players who maximize their *current* payoff. When cooperators yield a one-shot payoff below average, they tend to switch to the other strategy. This switch of actions may then in turn affect the dynamics of the environmental parameter n and the strategy choices of other players. However, whether such a switch is profitable is exclusively evaluated based on its immediate payoff effects. The subsequent strategy and environmental dynamics is not taken into account when players decide whether to switch to a different strategy.

Due to these differences, the framework of Weitz *et al*[10] is not immediately applicable to the examples we have studied. But to illustrate the differences between the two frameworks, we can naively use their model to analyze the two-player example depicted in **Fig. 2a** of the main text, by studying the solutions of (47) for the payoff values used in **Fig. 2a**,

$$\begin{array}{llll} R_0 = b_1 - c, & S_0 = -c, & T_0 = b_1, & P_0 = 0, \\ R_1 = b_2 - c, & S_1 = -c, & T_1 = b_2, & P_1 = 0. \end{array} \tag{48}$$

For this special case, the replicator equation (47) simplifies to

$$\begin{array}{rcl} \dot{x} & = & -cx(1-x), \\ \dot{n} & = & \varepsilon n(1-n)\big[-1 + (1+\theta)x\big]. \end{array} \tag{49}$$

For the special payoff values considered, the first equation becomes independent of n (because defectors always have the constant payoff advantage c over cooperators, independent of the current environment). Since $c > 0$, it follows for $0 < x < 1$ that $\dot{x} < 0$. As a consequence, $x(t) \to 0$ for any solution of (49) with $0 \le x(0) < 1$ and $0 \le n(0) \le 1$. We conclude that for any initial population in the interior of the state space, cooperators eventually go extinct. **Fig. S14** provides an illustration of this result, and compares the replicator dynamics (47) with the evolutionary dynamics of stochastic games as presented in **Fig. 2a**. This extinction result is not restricted to the specific payoff values in (48). Instead, Weitz *et al*[10] show that when both games constitute a prisoner's dilemma, with $T_i > R_i$ and $P_i > S_i$ for $i \in \{0, 1\}$, then cooperation always goes extinct.

We can slightly generalize this extinction result in the following way. Instead, of Eq. (47), suppose the evolutionary dynamics of x and n is governed by the following system of equations,

$$\begin{array}{rcl} \dot{x} & = & x(1-x)\Big[r_1(x,n) - r_2(x,n)\Big], \\ \dot{n} & = & \varepsilon n(1-n)\varphi(x,n), \end{array} \tag{50}$$

where $r_1(x,n)$, $r_2(x,n)$ and $\varphi(x,n)$ are differentiable functions in x and n. The ordinary differential equation (50) generalizes Eq. (47) in two ways. First, it does not specify the exact form of the payoff functions $r_1(x,n)$ and $r_2(x,n)$. In particular, a player's payoff may be a nonlinear function of the fraction of cooperators in a population, or of the environmental state. In this way, the replicator equation (50) can be used to capture the dynamics of multiplayer games[59]. Second, we do no longer assume a specific

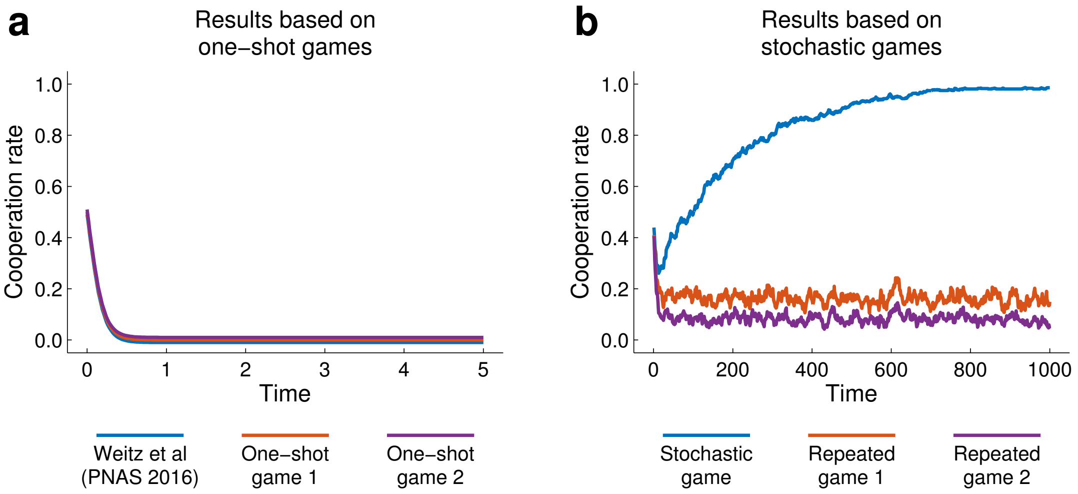

Figure S14: When players face a social dilemma in each state, environmental feedback may not suffice to sustain cooperation. We compare the deterministic dynamics (47) for one-shot games with our results presented in **Fig. 2a** for stochastic games. **a,** The blue curve shows the numerically computed solution of (47) for the initial state $x\!=\!n\!=\!0.5$. For comparison, we additionally show two solutions of the standard replicator equation assuming that the environment is fixed (at either $n=0$ or $n=1$) and payoffs do not co-evolve. All trajectories lead to full defection. **b,** While the repeated game 1 and the repeated game 2 lead to defection, the stochastic game facilitates cooperation. Parameters: **a,** We use the payoffs of **Fig. 2a** of our main text, $b_1=1.8$, $b_2=1.2$, and $c=1$. For the evolutionary parameters we use $\varepsilon=0.1$ and $\theta=2$, as in Fig. 2 of Weitz *et al*[10]. **b,** This panel reproduces our results from **Fig. 2a**. For better comparison with panel a, initially all players use a random initial strategy.

functional form to describe the dynamics of the environment; the function $\varphi(x,n)$ is only assumed to be differentiable, which is merely needed to guarantee existence and uniqueness of solutions of (50). We obtain the following result for the case that players are always engaged in a social dilemma.

Claim. *Suppose one-shot payoffs are such that* $r_1(x,n)\leq r_2(x,n)$ *for all* x *and* n*, and let* $\big(x(t),n(t)\big)$ *be a solution of the replicator equation (50). Then* $x(t)$ *is monotonically decreasing. Moreover, if* $r_1(x,n)<r_2(x,n)$ *for all* x *and* n*, then* $x(t)\to 0$ *for all initial populations with* $x(0)<1$.

Proof. The statement follows directly because $\dot{x} = x(1-x)\Big[r_1(x,n) - r_2(x,n)\Big] \leq 0$ by assumption. If $r_1(x,n)<r_2(x,n)$, then $V(x):=x$ is a Lyapunov function for $0\leq x<1$. □

In all of our examples in the main text we have assumed that players face a social dilemma whenever they are prompted to make a decision, such that the one-shot payoff of cooperation is at most the one-shot payoff of defection. We thus interpret the above result as an indication that for our findings, the interaction of reciprocity and payoff feedback has been essential: in none of the examples we would have observed the emergence of cooperation if players updated their strategies based on their one-shot payoffs only.

Appendix B: Proofs of the Propositions

Proof of Proposition 1.

1. According to Eqs. (7) and (8), the values of $P_k(s_j,\mathbf{a})$ can only affect the transition matrix when calculating entries of the form $M_{(s,\mathbf{a})\to(s_j,\mathbf{a}')}$. However, as $Q(s_j|s,\mathbf{a})=0$ for all $s\in S$, it follows from Eq. (7) that $M_{(s,\mathbf{a})\to(s_j,\mathbf{a}')}=0$, no matter what $P_k(s_j,\mathbf{a})$ is.

2. When the transition function Q is state-independent and deterministic, then for any action profile $\mathbf{a}$ there is a state $s_\mathbf{a}$ such that $Q(s_j|s,\mathbf{a})=0$ for all states $s_j\neq s_\mathbf{a}$, for all previous states $s\in S$. By the first part of the Proposition, the transition matrix M is thus independent of the values of $P_i(s_j,\mathbf{a})$ for all states $s_j\neq s_\mathbf{a}$. In particular, if we define $P_i'(s_j,\mathbf{a}):=P_i(s_\mathbf{a},\mathbf{a})$ for all s_j, then P' is state-independent and neither the transition matrix M nor the players' payoffs change if P_i is replaced by P_i'.

□

Proof of Proposition 2.

For the proof, we make use of the one-shot deviation principle[64]: to show that there is no profitable deviation in a group in which everyone uses $WSLS$, we only need to check all one-shot deviations (where a mutant plays a different action for one round, and returns to $WSLS$ for all subsequent rounds). Due to the definition of $WSLS$, it is useful to distinguish two possible cases, depending on whether or not all players have chosen the same action in the previous round. In the following, we calculate the continuation payoffs for each of these cases, assuming that all players apply $WSLS$.

1. If the present state is $s_i\in S$ and all players have used the same action in the previous round (or players find themselves in the very first round of the game), then all players cooperate in all subsequent rounds, and the continuation payoff π^i_{same} becomes

$$\pi^i_{\text{same}}=(1-\delta)u^i_{C,n-1}+\delta\sum_{j=1}^{m}Q(s_j|n)\cdot u^j_{C,n-1}. \tag{51}$$

2. If players have used different actions, then all players defect in the next round. Hence, *after* the next round players find themselves in the case in which everybody has chosen the same action. The respective continuation payoff π^i_{diff} thus becomes

$$\pi^i_{\text{diff}}=(1-\delta)u^i_{D,0}+\delta\sum_{j=1}^{m}Q(s_j|0)\cdot\pi^j_{\text{same}}. \tag{52}$$

If a player instead decides to make a one-shot deviation in one of these two cases, we can calculate her continuation payoff as follows:

1. If all players have used the same action in the previous round (or if players find themselves in the very first round), a one-shot deviation requires the mutant to defect in the next round, and to play according to $WSLS$ thereafter. The corresponding continuation payoff is

$$\tilde{\pi}^i_{\text{same}} = (1-\delta)u^i_{D,n-1} + \delta \sum_{j=1}^{m} Q(s_j|n-1) \cdot \pi^j_{\text{diff}}. \tag{53}$$

2. Alternatively, if players have used different actions in the previous round, a one-shot deviation from $WSLS$ requires the mutant to cooperate in the next round. Her continuation payoff becomes

$$\tilde{\pi}^i_{\text{diff}} = (1-\delta)u^i_{C,0} + \delta \sum_{j=1}^{m} Q(s_j|1) \cdot \pi^j_{\text{diff}}. \tag{54}$$

$WSLS$ is a subgame perfect equilibrium if and only if $\pi^i_{\text{same}} \geq \tilde{\pi}^i_{\text{same}}$ and $\pi^i_{\text{diff}} \geq \tilde{\pi}^i_{\text{diff}}$. Plugging the definitions (51)–(54) into these inequalities and basic algebraic manipulations yield (17). □

Appendix C: Matlab code for the calculation of payoffs

In the following we provide the code that we have used to calculate the players’ payoffs in stochastic games with n players and two states.

```
function [pivec,cvec]=calcPay(Str,QVec,r1,r2,c);
% Calculates the payoff and cooperation rates in a stochastic game with
% deterministic transitions, playing a PGG in each state
% Str ...  Matrix with n rows, each row contains the strategy of a player
% Strategies have the form (pC,n-1,...pC,0,pD,n-1,...pD,0) where the letter
% refers to the player’s own action and number refers to cooperators among
% co-players.
% QVec = (qn,...,q0) vector that contains the transition
% probabilities qi to go to state 1 in the next round, depending on
% the number of cooperators
% r1,r2 ...  multiplication factors of PGG in each state
% c ...  cost of cooperation

% PART I -- PREPARING A LIST OF ALL POSSIBLE STATES OF THE MARKOV CHAIN,
% PREPARING A LIST OF ALL POSSIBLE PAYOFFS IN A GIVEN ROUND
% A state has the form (s,a1,...an) where s is the state of the
% stochastic game and a1,...,an are the player’s actions.
% Hence there are 2^(n+1) states.

n=size(Str,1);
PossState=zeros(2^(n+1),n+1); % Matrix where each row corresponds to a possible
state
for i=1:2^(n+1)
   PossState(i,:)=sscanf( dec2bin(i-1,n+1), ’%1d’ )’;
end
piRound=zeros(2^(n+1),n); % Matrix where each row gives the payoff of all players
in a given state
for i=1:2^(n+1)
   State=PossState(i,:); nrCoop=sum(State(2:end)); Mult=State(1)*r2+(1-State(1))*r1;
   for j=1:n
      piRound(i,j)=nrCoop*Mult/n-State(j+1)*c;
   end
end

% PART II -- CREATING THE TRANSITION MATRIX BETWEEN STATES

M=zeros(2^(n+1),2^(n+1));
ep=0.001; Str=(1-ep)*Str+ep*(1-Str);
```

```
for row=1:2^(n+1);
   StOld=PossState(row,:); % PreviousState
   nrCoop=sum(StOld(2:end)); EnvNext=QVec(n+1-nrCoop);
   for col=1:2^(n+1);
      StNew=PossState(col,:); %NextState
      if StNew(1)==1-EnvNext;
         trpr=1; % TransitionProbability
         for i=1:n
            iCoopOld=StOld(1+i);
            pval=Str(i,2*n-nrCoop-(n-1)*iCoopOld);
            iCoopNext=StNew(1+i);
            trpr=trpr*(pval*iCoopNext+(1-pval)*(1-iCoopNext));
         end
      else
         trpr=0;
      end
      M(row,col)=trpr;
   end
end
v=null(M'-eye(2^(n+1))); freq=v'/sum(v);
pivec=freq*piRound;
cvec=sum(freq*PossState(:,2:end))/n;
end
```

References

[1] Sigmund, K. *The Calculus of Selfishness* (Princeton Univ. Press, 2010).

[2] Traulsen, A. & Hauert, C. Stochastic evolutionary game dynamics. In Schuster, H. G. (ed.) *Reviews of Nonlinear Dynamics and Complexity*, 25–61 (Wiley-VCH, Weinheim, 2009).

[3] Cressman, R. *Evolutionary Dynamics and Extensive Form Games* (MIT Press, Cambridge, 2003).

[4] Broom, M. & Rychtář, J. *Game-Theoretical Models in Biology* (Chapman and Hall/CRC, 2013).

[5] Assaf, M., Mobilia, M. & Roberts, E. Cooperation dilemma in finite populations under fluctuating environments. *Physical Review Letters* **111**, 238101 (2013).

[6] Ashcroft, P., Altrock, P. M. & Galla, T. Fixation in finite populations evolving in fluctuating environments. *Journal of The Royal Society Interface* **11**, 20140663 (2014).

[7] Gokhale, C. S. & Hauert, C. Eco-evolutionary dynamics of social dilemmas. *Theoretical Population Biology* **111**, 28–42 (2016).

[8] Hauert, C., Holmes, M. & Doebeli, M. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B* **273**, 2565–2570 (2006).

[9] Hauert, C., Holmes, M. & Doebeli, M. Corrigendum – evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B* **273**, 3131–3132 (2006).

[10] Weitz, J. S., Eksin, C., Paarporn, K., Brown, S. P. & Ratcliff, W. C. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences USA* **113**, E7518–E7525 (2016).

[11] Tavoni, A., Schlüter, M. & Levin, S. A. The survival of the conformist: Social pressure and renewable resource management. *Journal of Theoretical Biology* **299**, 152–161 (2012).

[12] Safarzynska, K. The coevolution of culture and environment. *Journal of Theoretical Biology* **322**, 46–57 (2013).

[13] Roberts, G. & Sherratt, T. N. Development of cooperative relationships through increasing investment. *Nature* **394**, 175–179 (1998).

[14] Wahl, L. M. & Nowak, M. A. The continuous prisoner's dilemma: I. Linear reactive strategies. *Journal of Theoretical Biology* **200**, 307–321 (1999).

[15] Killingback, T. & Doebeli, M. 'raise the stakes' evolves into a defector. *Nature* **400**, 518 (1999).

[16] Killingback, T. & Doebeli, M. The continuous Prisoner's Dilemma and the evolution of cooperation through reciprocal altruism with variable investment. *The American Naturalist* **160**, 421–438 (2002).

[17] McAvoy, A. & Hauert, C. Autocratic strategies for iterated games with arbitrary action spaces. *Proceedings of the National Academy of Sciences* **113**, 3573–3578 (2016).

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

[19] Stewart, A. J., Parsons, T. L. & Plotkin, J. B. Evolutionary consequences of behavioral diversity. *Proceedings of the National Academy of Sciences USA* **113**, E7003–E7009 (2016).

[20] Shapley, L. S. Stochastic games. *Proceedings of the National Academy of Sciences* **39**, 1095–1100 (1953).

[21] Gilette, D. Stochastic games with zero stop probabilities. In *Contributions to the theory of Games III*, 179–188 (Princeton University Press, 1957).

[22] Alur, R., Henzinger, T. & Kupferman, O. Alternating-time temporal logic. *Journal of the ACM* **49**, 672–713 (2002).

[23] de Alfaro, L., Henzinger, T. A. & Mang, F. Y. C. The control of synchronous systems. In *CONCUR'00*,

458–473 (Springer, 2000).

[24] de Alfaro, L., Henzinger, T. A. & Mang, F. Y. C. The control of synchronous systems, Part II. In *CONCUR'01*, 566–580 (Springer, 2001).

[25] Miltersen, P. B. & Sorensen, T. B. A near-optimal strategy for a heads-up no-limit texas hold'em poker tournament. *AAMAS'07* 191–197 (2007).

[26] Bowling, M., Burch, N., Johanson, M. & Tammelin, O. Heads-up limit hold'em poker is solved. *Science* **347**, 145–149 (2015).

[27] Neyman, A. & Sorin, S. (eds.) *Stochastic games and applications* (Kluwer Academic Press, Dordrecht, 2003).

[28] Mertens, J. F. & Neyman, A. Stochastic games. *International Journal of Game Theory* **10**, 53–66 (1981).

[29] Bewley, T. & Kohlberg, E. The asymptotic theory of stochastic games. *Math. Oper. Research* 197–208 (1976).

[30] Vieille, N. Two-player stochastic games I: A reduction. *Israel Journal of Mathematics* **119**, 55–91 (2000).

[31] Vieille, N. Two-player stochastic games II: The case of recursive games. *Israel Journal of Mathematics* **119**, 93–126 (2000).

[32] Sobel, M. J. Continuous stochastic games. *Journal of Applied Probability* **10**, 597–604 (1973).

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

[34] Perc, M. *et al.* Statistical physics of human cooperation. *Physics Reports* **687**, 1–51 (2017).

[35] Wardil, L. & Hauert, C. Origin and structure of dynamic cooperative networks. *Scientific Reports* **4**, 5725 (2014).

[36] Sigmund, K. Punish or perish? Retaliation and collaboration among humans. *Trends in Ecology and Evolution* **22**, 593–600 (2007).

[37] Sasaki, T. & Uchida, S. The evolution of cooperation by social exclusion. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20122498 (2013).

[38] Aumann, R. J. Survey of repeated games. In Henn, R. & Moeschlin, O. (eds.) *Essays in game theory and mathematical economics in honor of Oskar Morgenstern*, chap. Survey of repeated games (Wissenschaftsverlag, 1981).

[39] Dutta, P. K. A Folk Theorem for stochastic games. *Journal of Economic Theory* **66**, 1–32 (1995).

[40] Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).

[41] Hilbe, C., Wu, B., Traulsen, A. & Nowak, M. A. Cooperation and control in multiplayer social dilemmas. *Proceedings of the National Academy of Sciences USA* **111**, 16425–16430 (2014).

[42] 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).

[43] Nowak, M. A. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta Applicandae Mathematicae* **20**, 247–265 (1990).

[44] Ohtsuki, H. & Iwasa, Y. The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology* **239**, 435–444 (2006).

[45] 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 B* **264**, 513–519 (1997).

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

[47] Martinez-Vaquero, L. A., Cuesta, J. A. & Sanchez, A. Generosity pays in the presence of direct reciprocity: A comprehensive study of 2x2 repeated games. *PLoS ONE* **7**, E35135 (2012).

[48] Press, W. H. & Dyson, F. D. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. *PNAS* **109**, 10409–10413 (2012).

[49] 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).

[50] Baek, S. K., Jeong, H. C., Hilbe, C. & Nowak, M. A. Comparing reactive and memory-one strategies of direct reciprocity. *Scientific Reports* **6**, 25676 (2016).

[51] Ichinose, G. & Masuda, N. Zero-determinant strategies in finitely repeated games. *Journal of Theoretical Biology* **438**, 61–77 (2018).

[52] Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Physical Review E* **74**, 011909 (2006).

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

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

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

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

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

[58] Kurokawa, S. & Ihara, Y. Emergence of cooperation in public goods games. *Proceedings of the Royal Society B* **276**, 1379–1384 (2009).

[59] Gokhale, C. S. & Traulsen, A. Evolutionary games in the multiverse. *Proceedings of the National Academy of Sciences USA* **107**, 5500–5504 (2010).

[60] Hilbe, C., Nowak, M. A. & Sigmund, K. The evolution of extortion in iterated prisoner's dilemma games. *Proceedings of the National Academy of Sciences USA* **110**, 6913–6918 (2013).

[61] 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).

[62] Hofbauer, J. & Sigmund, K. *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, UK, 1998).

[63] Strogatz, S. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Perseus Books, Cambridge, Massachusetts, 1994).

[64] Blackwell, D. Discounted dynamic programming. *Annals of Mathematical Statistics* **36**, 226–235 (1965).