PNAS Nexus

*PNAS Nexus*, 2022, **1**, 1–15
https://doi.org/10.1093/pnasnexus/pgac141
Advance access publication date: 27 July 2022
**Research Report**

# Evolutionary instability of selfish learning in repeated games

Alex McAvoy [a,b,*], Julian Kates-Harbeck [c], Krishnendu Chatterjee [d] and Christian Hilbe [e]

[a]Department of Mathematics, University of Pennsylvania, Philadelphia, PA, USA
[b]Center for Mathematical Biology, University of Pennsylvania, Philadelphia, PA, USA
[c]Department of Physics, Harvard University, Cambridge, MA, USA
[d]Institute of Science and Technology Austria, Klosterneuburg, Austria
[e]Max Planck Research Group: Dynamics of Social Behavior, Max Planck Institute for Evolutionary Biology, Plön, Germany
*To whom correspondence should be addressed: Email: amcavoy@sas.upenn.edu
**Edited By:** Karen E. Nelson.

## Abstract

Across many domains of interaction, both natural and artificial, individuals use past experience to shape future behaviors. The results of such learning processes depend on what individuals wish to maximize. A natural objective is one's own success. However, when two such "selfish" learners interact with each other, the outcome can be detrimental to both, especially when there are conflicts of interest. Here, we explore how a learner can align incentives with a selfish opponent. Moreover, we consider the dynamics that arise when learning rules themselves are subject to evolutionary pressure. By combining extensive simulations and analytical techniques, we demonstrate that selfish learning is unstable in most classical two-player repeated games. If evolution operates on the level of long-run payoffs, selection instead favors learning rules that incorporate social (other-regarding) preferences. To further corroborate these results, we analyze data from a repeated prisoner's dilemma experiment. We find that selfish learning is insufficient to explain human behavior when there is a trade-off between payoff maximization and fairness.

**Significance Statement:**

A natural first approach to learning is to attempt to improve one's own outcome (e.g. wealth, resources, or reputation), without regard for others. This kind of "selfish" learning, however, can be detrimental in social dilemmas, in which the individuals' incentives are at odds. Here, we study the evolutionary dynamics of different learning rules, demonstrating that selfish learning can be driven to extinction in evolving populations. To this end, we contrast selfish learning with a competing learning rule, which uses simple social preferences. The competing rule attains superior outcomes across a wide range of social interactions, even when interacting with selfish learners, and it ensures that these outcomes are both fair and socially optimal.

## Introduction

Individuals naturally adapt to their environment, either by modifying their existing behaviors or by considering alternative ones when necessary (1, 2). The study of behavioral adaptation is far-reaching, with applications ranging from microbial dynamics (3,4) to social preferences in humans (5–7) to learning algorithms in multiagent systems (8, 9). Evolutionary game theory, a tool for modeling behavioral adaptations (10–17), has been used to describe how people learn to engage in reciprocity (18–30), how social norms evolve over time (31,32), how groups are formed (33), how thriving communities can suddenly be undermined by corruption (34,35), and how artificial agents behave "in silico" (36–39). A key assumption in evolutionary game theory is that individuals might not act optimally from the outset. Rather, adaptation happens over time through either cultural or genetic mechanisms (10–15).

To describe how people learn new behaviors, the respective literature has considered various cognitive processes. Some models stipulate that individuals learn by imitating their peers (40–42), whereas others assume that learning is based on aspiration levels (43,44), reinforcement (45, 46), or introspection (47, 48). Crucially, however, learning rules are frequently based on the assumption that individuals strive to increase their own immediate payoffs. Although errors, mutations, and chance events may temporarily lead individuals to adopt inferior strategies, better performing strategies are favored on average. We refer to learning rules with this property as "selfish learning."

In reality, many models ask neither whether individuals actually learn based on strict payoff maximization nor whether they have a long-run incentive to do so (40–48). This assumption might be justified when interactions lack any strategic component (as in single-player optimization). However, the rationale for selfish learning is less clear in social dilemmas, where there are conflicts of interest between the individual and the group (49–51). When social dilemmas give rise to multiple equilibria, selfish optimization may easily result in detrimental outcomes that are socially

**Competing Interest:** The authors declare no competing interest.
**Received:** December 6, 2021. **Accepted:** July 22, 2022

inefficient. This drawback is already well-recognized in the field of multiagent reinforcement learning, where selfish learners are considered "naïve" and serve as a benchmark for other learning rules (52). If selection acts upon the learning rules that determine how players choose strategies, then it may select for different learning rules altogether.

The problem of optimal learning is best illustrated with a key model of evolutionary game theory: the repeated prisoner's dilemma (19). In each round of the game, two players independently choose whether to cooperate or defect. Each player has an incentive to defect even though mutual cooperation is in everybody's interest. For repeated games with sufficiently long time horizons, there is a wide range of possible equilibria (53). In particular, players may always cooperate, always defect, or alternate between cooperation and defection (54). After decades of research, much is known about particular strategies that sustain cooperation, such as "tit-for-tat" (55) or "win-stay, lose-shift" (36). Much less is known how such desirable strategies can be learned in the first place, especially when players are learning concurrently. If strategy updating is described by imitation, for example, players may end up defecting for a substantial amount of time (56). Such detrimental outcomes become increasingly likely when the game involves only a few rounds, offers a small benefit of cooperation, or when players commit errors (21).

In this study, we ask whether there is a learning rule that helps individuals find more profitable equilibria, even when the opponent is self-interested. To this end, we imagine two individuals interacting in a repeated game. Each individual has its own learning rule for adapting its strategy, with the ultimate goal of achieving a high payoff. As a baseline, we consider a variant of selfish learning. At regular time intervals, a selfish learner compares the performance of its present strategy with the (hypothetical) performance of a slightly perturbed version of its strategy. The learner adopts the perturbed strategy if it yields a higher payoff, regardless of its effects on others.

We contrast selfish learning with a learning rule we term "fairness-mediated team learning" (*FMTL*). Players with this learning rule balance two objectives, efficiency and fairness. To promote efficiency, *FMTL* favors strategies that increase the total payoff of the group (i.e. the "team" payoff). In this way, *FMTL* aims to avoid equilibria that leave everybody worse off. To prevent themselves from getting exploited, however, *FMTL* players simultaneously aim to minimize payoff differences within their group (i.e. promote "fairness"). The respective weights assigned to efficiency and fairness are dynamically adjusted based on the players' current payoffs. Increasing efficiency is the primary objective when payoff inequalities are negligible.

The definition of *FMTL* has a natural connection to the field of multiagent learning (57–61). Two of the most well-studied areas involve so-called "fully-cooperative" and "fully-competitive" interactions. In fully cooperative interactions, players have identical payoff functions; what is good for one player is equally good for the other. In fully competitive settings, the players' incentives are perfectly opposed. The iterated prisoner's dilemma falls somewhere in between and is often referred to as a "mixed" or "general-sum" game (62,63). As such, it presents a more difficult learning problem. In striving for fairness, *FMTL* forces the players to have approximately equal payoffs, which is reminiscent of the fully cooperative setting. Once payoffs are sufficiently close, *FMTL* views itself and the opponent as a team and attempts to optimize the total payoff. The approach of optimizing a team score is common in cooperative settings, especially when individuals have mostly (but not completely) aligned incentives (9).

When *all* learners are driven by fairness and efficiency, it may not be surprising that the learning dynamics favor socially beneficial outcomes. Remarkably, however, such outcomes already arise when only *one* learner has these objectives. For a wide range of two-player games, we show that *FMTL* players tend to settle at equilibria that are both individually optimal and socially efficient even when interacting with selfish opponents. Based on these observations, we explore the dynamics that arise when the two learning rules themselves are subject to evolution. We consider a process with two timescales. On a short timescale, individuals have a fixed learning rule that they use to guide their strategic choices. On a longer timescale, players can switch their learning rule, based on how successful it proved to be. The resulting evolutionary dynamics depend on which game is played, and on the relative pace at which learning rules are updated (compared to how often strategies are updated). When learning rules and strategies evolve at a similar timescale, selfish learning is favored. However, when learning rules evolve at a slower rate, selfish learning is routinely invaded and ceases to be stable in many classical games.

# Results

## A model of learning in repeated games

To compare different learning rules, we consider two players who interact in a repeated game. In each round, players independently decide whether to cooperate or defect. As a result, each player obtains a payoff (Fig. 1a). After every round, players interact for another round with probability $\lambda$. They decide whether to cooperate based on their strategies. A strategy is a rule that tells the player what to do in the next round, given the history of previous play. In the simplest version of the model, players use memory-one strategies (13). These strategies are contingent on only the last round of play, and they reasonably approximate human behavior in economic experiments (64–67). A memory-one strategy for player $X$ consists of an initial probability of cooperation, $p_0$, together with a four-tuple of conditional cooperation probabilities, $\mathbf{p} = (p_{CC}, p_{CD}, p_{DC}, p_{DD}) \in [0, 1]^4$. Here, $p_{xy}$ is the probability that $X$ cooperates in the next round when $X$ played $x$ and $Y$ played $y$ in the previous round. For example, an unconditional defector is represented by a memory-one strategy with $p_0 = 0$ and $\mathbf{p} = (0, 0, 0, 0)$. Tit-for-tat takes the form $p_0 = 1$ and $\mathbf{p} = (1, 0, 1, 0)$. Given strategies $\mathbf{p}$ and $\mathbf{q}$ of the two players, we can compute how likely they are to cooperate over the course of the game, and which overall payoffs $\pi_X(\mathbf{p}, \mathbf{q})$ and $\pi_Y(\mathbf{p}, \mathbf{q})$ they obtain (Fig. 1b and "Methods").

After each repeated game, players may update their strategies based on their experience with the opponent. They do so by implementing a learning rule. We consider learning rules that consist of four elements: a distribution $\mathcal{D}$ over initial strategies, a sampling procedure $\mathcal{S}$, a set of objective functions $\mathcal{V}$, and a priority assignment $\Omega$. The first element, the distribution over initial strategies, determines the player's default strategy that is used before any learning takes place. The second element, the sampling procedure, determines how to generate an alternative strategy in each learning step. Throughout the main text, we assume that a player $X$ with current strategy $\mathbf{p}$ generates an alternative strategy $\mathbf{p}'$ using local random search (68) (Fig. 1c). After generating an alternative strategy, the player decides whether to accept it. To this end, the set of objective functions $\mathcal{V}$ specifies what the player's strategy ought to maximize. If $X$'s objective is to maximize $V \in \mathcal{V}$, the player switches to $\mathbf{p}'$ if and only if $V(\mathbf{p}', \mathbf{q}) > V(\mathbf{p}, \mathbf{q})$. This decision may involve, for example, a comparison between the player's current payoff, $\pi_X(\mathbf{p}, \mathbf{q})$, and the payoff $\pi_X(\mathbf{p}', \mathbf{q})$ the player could have ob-

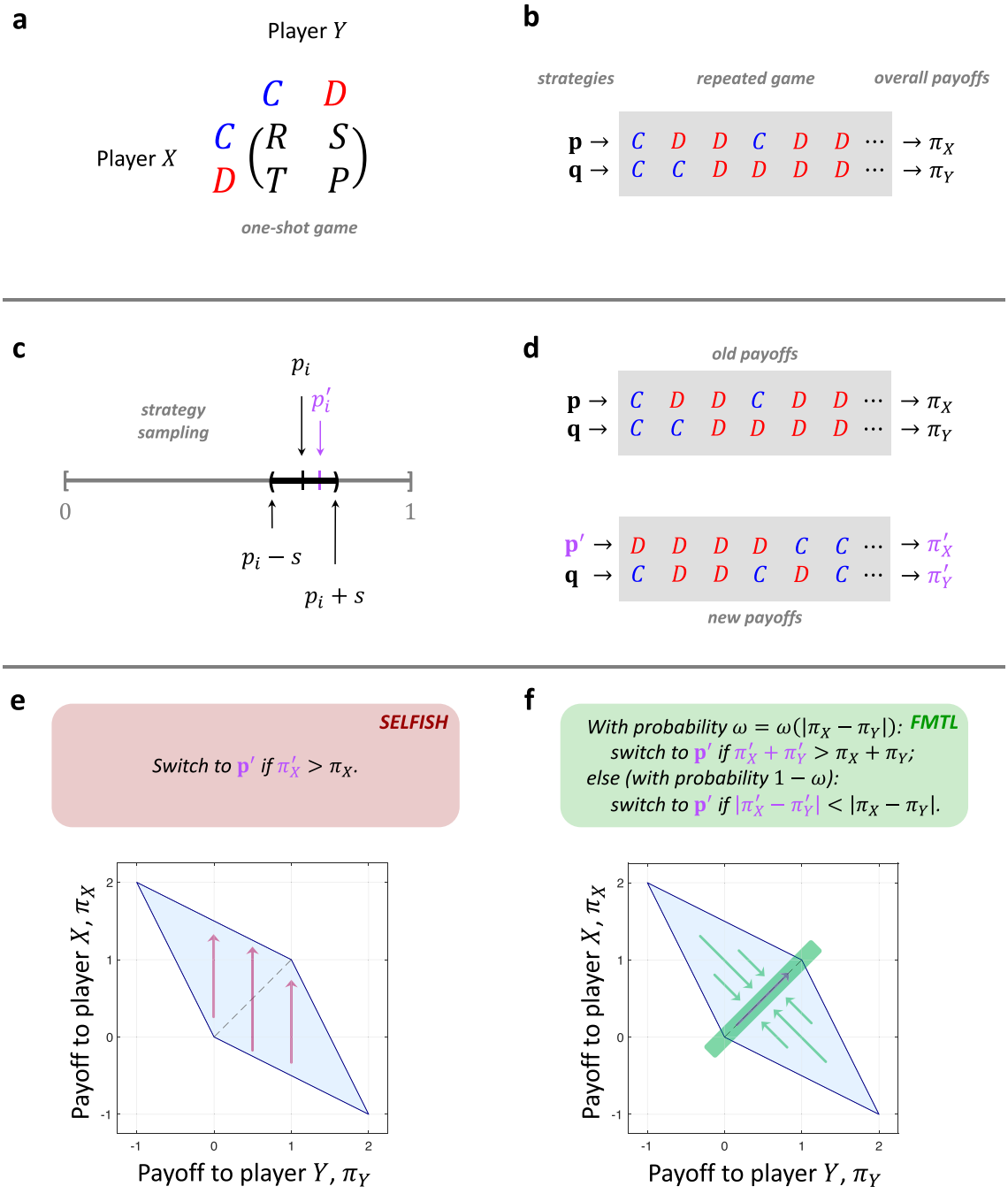


**Fig. 1.** Learning in repeated games and *FMTL*. (a) In each encounter of a repeated game, players engage in a "one-shot" game. Both choose actions, C or D, and receive payoffs, R, S, T, or P. (b) A repeated game consists of a sequence of one-shot games. Both players choose strategies, $\mathbf{p}$ and $\mathbf{q}$, and receive overall payoffs, $\pi_X$ and $\pi_Y$, respectively. (c) To update strategies, players occasionally sample a new, nearby strategy $\mathbf{p}'$ ("local search"). (d) When X uses strategy $\mathbf{p}'$ against Y's strategy, $\mathbf{q}$, the players get new payoffs, $\pi_X'$ and $\pi_Y'$. This new strategy is then evaluated based on a player's learning rule. (e) If X is a selfish learner, $\mathbf{p}'$ is accepted only if it improves X's payoff, i.e. if $\pi_X' > \pi_X$. (f) If X uses *FMTL*, then with probability $1 - \omega$ she takes $\mathbf{p}'$ only if it brings the two players' payoffs closer together (fairness). Otherwise, with probability $\omega$, she takes $\mathbf{p}'$ only if it improves the sum of the two players' payoffs (efficiency). The probability $\omega$ is a decreasing function of the payoff difference so that fairness becomes more important to *FMTL* as one player starts to do better than the other.

tained using the alternative strategy (Fig. 1d). However, a player's priorities over her objectives may change over time. The priority assignment $\Omega(\mathbf{p}, \mathbf{q})$ determines with which probability each objective $V \in \mathcal{V}$ is chosen.

We compare the performance of two learning rules. According to selfish learning, a player switches to the alternative strategy if and only if it increases the player's payoff. Within our framework, selfish learning can be represented by the objective function $V_S(\mathbf{p}, \mathbf{q}) = \pi_X(\mathbf{p}, \mathbf{q})$, which the player strives to maximize (Fig. 1e). We refer to such a player as a selfish learner. The other learning rule is *FMTL* (Fig. 1f). *FMTL* has two objective functions, $\mathcal{V} = \{V_E, V_F\}$. The first objective is to achieve efficiency. With the objective function $V_E(\mathbf{p}, \mathbf{q}) = \pi_X(\mathbf{p}, \mathbf{q}) + \pi_Y(\mathbf{p}, \mathbf{q})$, the player aims to maximize the group's total payoff. The other objective is fairness. Here, a player aims to minimize payoff differences, which is equivalent to maximizing the objective

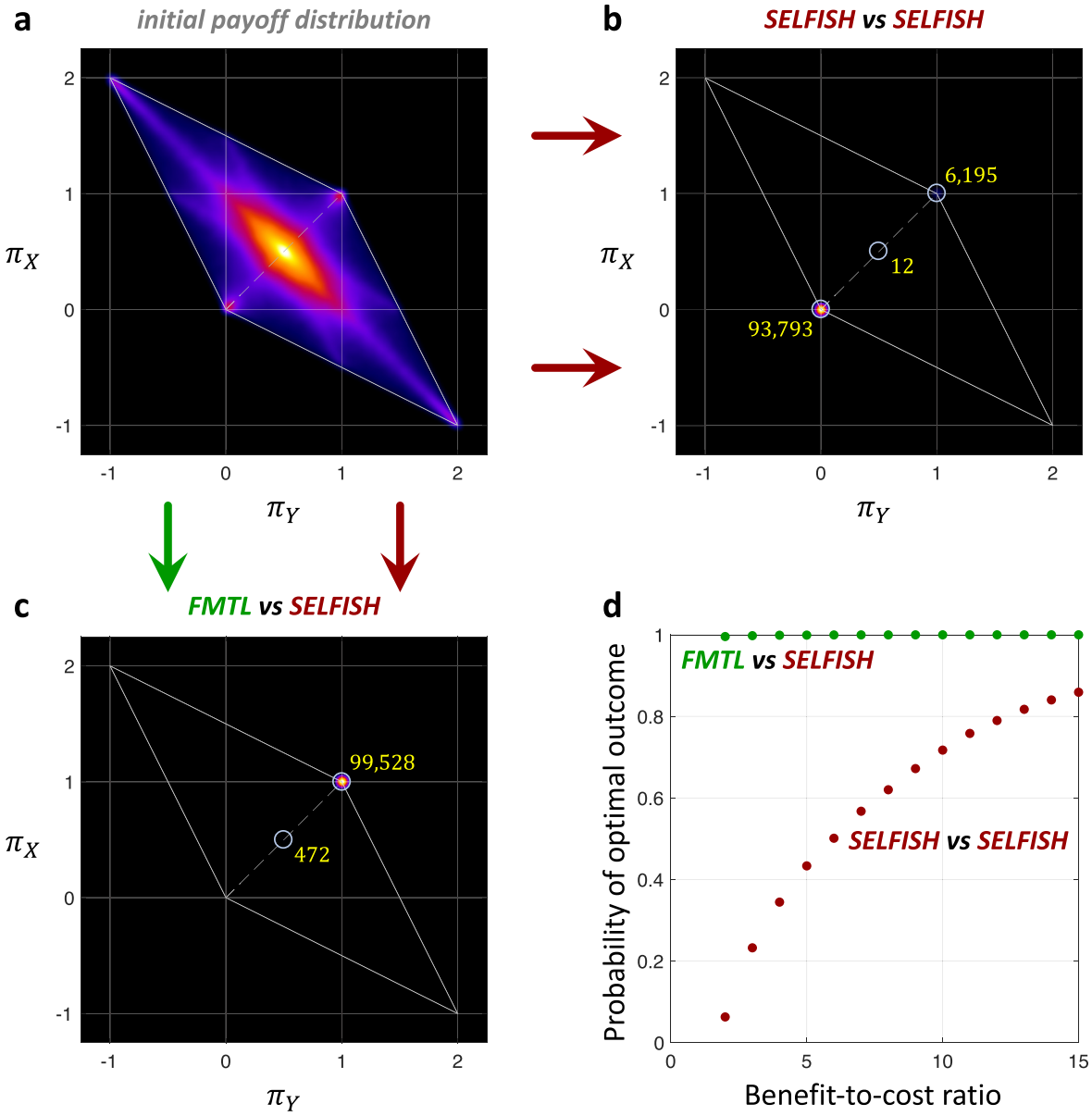


**Fig. 2.** *FMTL* versus a selfish learner in the repeated donation game. (a) $X$ and $Y$ initially choose random strategies and receive payoffs that fall within the feasible region. (b) When $X$ and $Y$ are both selfish learners, we let $X$ and $Y$ update until neither one has accepted a new strategy for $10^4$ update steps. This process is repeated for $10^5$ iterations. The resulting distribution of payoffs is concentrated around the mutual defection payoff. Only a small number of runs (approximately 6%) result in mutual cooperation, and an even smaller number settle at alternating cooperation. (c) When $X$ plays *FMTL* instead, final payoffs are concentrated around the payoff for mutual cooperation. This is also the fair and socially optimal outcome. (d) As one may expect, two selfish learners require a substantial benefit-to-cost ratio to coordinate on the optimal outcome ($b - c$) with high probability. In contrast, *FMTL* against a selfish learner gives excellent outcomes even when $b/c$ is small. The endpoints in (b) and (c) are based on $10^5$ random initial strategy pairs, and for $b = 2$ and $c = 1$.

$V_F(\mathbf{p}, \mathbf{q}) = -\left|\pi_X(\mathbf{p}, \mathbf{q}) - \pi_Y(\mathbf{p}, \mathbf{q})\right|$. *FMTL* prioritizes efficiency if the players' current payoffs are sufficiently close, and it prioritizes fairness if there is substantial inequality. For a precise description of the priority assignment, see "Methods."

## Learning dynamics across different repeated games

To compare the two learning rules, we first assume each player's learning rule is fixed. Across a range of different two-player games, we explore how the players' learning rules affect their payoffs. As the baseline scenario, we consider two selfish learners. We then contrast this scenario with groups where either one or both players use *FMTL*.

### *Prisoner's dilemma*

We start by exploring how players fare in one of the most basic and well-studied social dilemmas, the donation game (13). Here, cooperation means paying a cost $c > 0$ to deliver a benefit of $b > c$ to the coplayer. This results in a prisoner's dilemma: players individually prefer to defect, yet mutual cooperation yields a better payoff than mutual defection. Players start out with random memory-one strategies (Fig. 2a) and the game is repeated for many rounds (see "Methods"). Previous work shows there are four types of

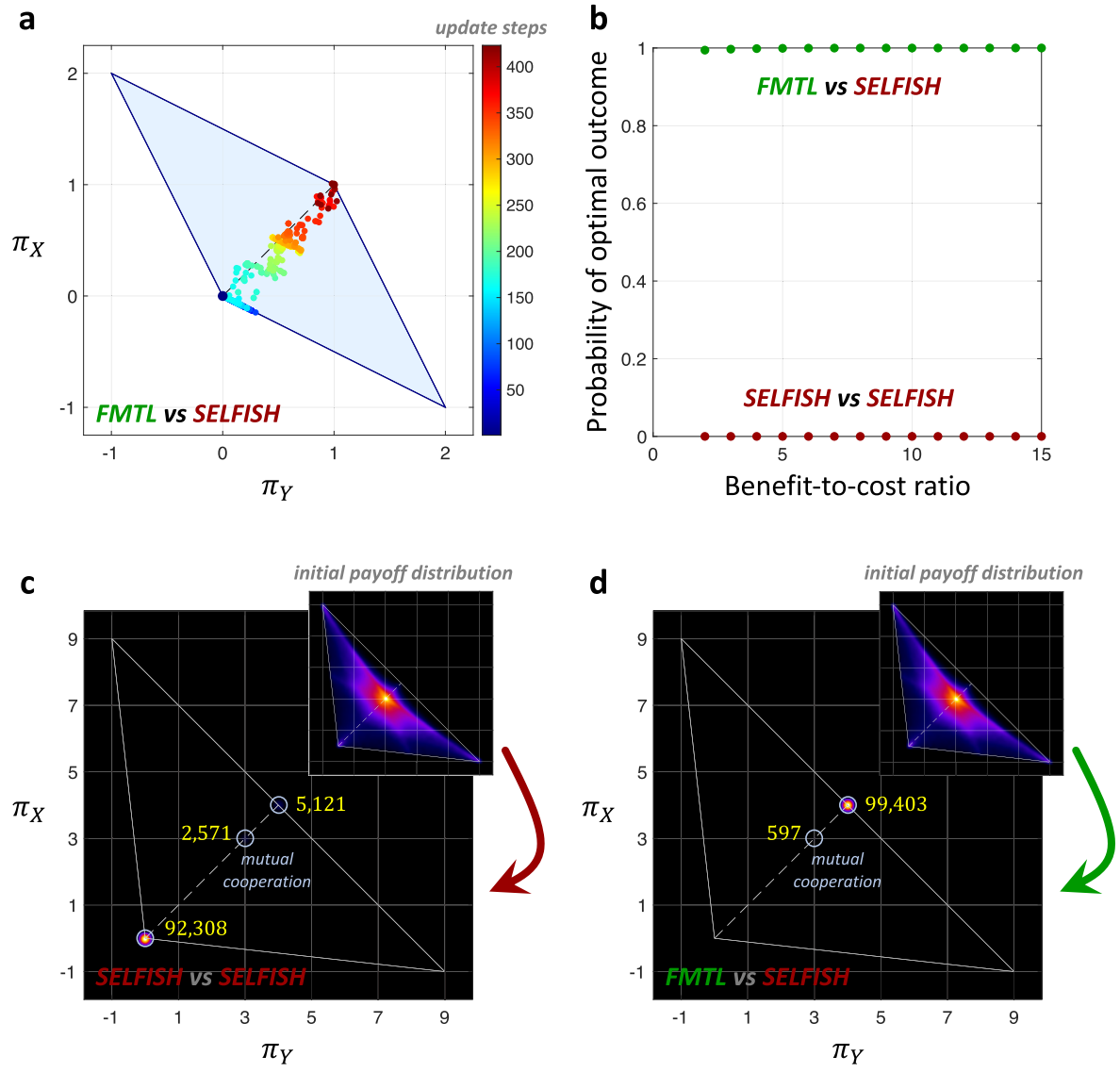


**Fig. 3.** Robustness of *FMTL*. (a) To illustrate how *FMTL* can help players to escape from mutual defection, we consider a scenario where *X* uses *FMTL* and *Y* is a selfish learner. Initially, both *X* and *Y* play ALLD ("always defect"). Even under such hostile initial conditions, the process moves toward mutual cooperation. (b) Across all benefit-to cost-ratios, pairs with an *FMTL* learner reliably escape mutual defection. In contrast, two selfish learners never escape. (c) and (d) In addition to the donation game, we also consider simulations for a prisoner's dilemma with $(S + T)/2 > R$. In this game, optimal play requires alternating cooperation and defection. When *FMTL* is matched with a selfish learner, players are indeed most likely to settle at this outcome, although they have no explicit information about *R* relative to $(S + T)/2$. Game parameters: (a and b) $R = 1$, $S = -1$, $T = 2$, and $P = 0$; (c and d) $R = 3$, $S = -1$, $T = 9$, and $P = 0$.

symmetric equilibria among memory-one players (54): players either both cooperate, both defect, alternate, or they use so-called equalizers (20,69).

We use simulations to explore which of these equilibrium outcomes is eventually realized (if any), depending on which learning rules the players apply. When two selfish learners interact, most of the time they either end up in mutual cooperation or mutual defection (Fig. 2b). Although cooperation is socially optimal, an overwhelming majority of simulations give rise to an all-defection equilibrium with low payoffs. The observed learning dynamics change completely if one (or both) of the two learners switches to *FMTL*. In that case, most simulated pairs of learners end up cooperating (Fig. 2c; see Figure S1 for a depiction of the final strategies). Remarkably, players yield almost full cooperation already for low benefit-to-cost ratios, for which cooperation is usually difficult to establish (21) (Fig. 2d).

To explore which mechanism allows *FMTL* to evade inefficient equilibria against selfish learners, we consider simulations in which initially both players defect (Fig. 3a). In this initial state, the players' payoffs are equal but inefficient. As a result, *FMTL* prioritizes efficiency over fairness, leading the respective learner to cooperate occasionally. While this reduces the payoff to *FMTL*, it also provides some strategic leverage. By cooperating conditionally, the learner can affect how profitable it is for the selfish opponent to cooperate. Once the *FMTL* player adopts a strategy that makes cooperation a best response, even a selfish opponent has an incentive to adapt (Figure S2 and Video S1). In this way, *FMTL* triggers dynamics of ever-increasing cooperation rates. Eventually, players reach an equilibrium in which both players cooperate. Different simulated trajectories vary due to stochasticity, but they almost always lead from defection to cooperation, independent of the magnitude of the benefit-to-cost ratio (Fig. 3b). We observe

similar dynamics when players start out with random strategies (see Figure S3).

Donation games may favor successful learning because efficiency requires only a simple pattern of behavior: both players merely need to cooperate each round. Instead, additional coordination problems might arise if efficiency requires the players to cooperate in turns. This problem occurs, for example, in a prisoner's dilemma with $R < (S + T)/2$. Here, it is socially optimal to agree on a policy of alternation: $X$ cooperates and $Y$ defects in even rounds, whereas $X$ defects and $Y$ cooperates in odd rounds. Again, we explore the dynamics of this game when players use either selfish learning or *FMTL*. When they both use selfish learning, they typically fail to cooperate altogether (Fig. 3c). But if one of them adopts *FMTL*, they overwhelmingly discover the optimal policy of alternation (Fig. 3d). We observe a similar pattern in the context of a harmony game (here, $R > T$ and $S > P$, such that cooperation is dominant (70)). When $R < (S + T)/2$, two selfish learners typically coordinate on mutual cooperation. But if one of the two players switches to *FMTL*, they both achieve the superior alternation outcome. Specifically, when two selfish learners interact in a game with payoffs $R = 1$, $S = 2$, $T = 0.5$, and $P = 0$, we find that over 99% of runs give a payoff of $(R, R) = (1, 1)$. Once one of the learners switches to *FMTL*, more than 99% of runs end up at $((S + T)/2, (S + T)/2) = (1.25, 1.25)$.

### *Generalized social dilemmas: stag hunt and snowdrift games*

While the prisoner's dilemma has been instrumental in modeling human cooperation, there are other natural rankings of the game's payoffs. These alternative rankings result in weaker forms of conflict. To further explore the performance of *FMTL*, we consider generalized social dilemmas, defined by three properties (49–51): (i) the payoff for mutual cooperation exceeds the payoff for mutual defection, $R > P$; (ii) when players choose different actions, the defector obtains a larger payoff than the cooperator, $T > S$; and (iii) irrespective of their own action, players prefer their opponent to cooperate, $R > S$ and $T > P$. In addition to the prisoner's dilemma and the harmony game, there are two more generalized social dilemmas (71): the stag hunt (72) and the snowdrift (73,74) game.

The stag hunt game has the payoff ranking $R > T > P > S$. In particular, mutual defection is an equilibrium of the one-shot game, but so is mutual cooperation. For repeated stag hunt games, we find that players do not need to settle at socially optimal outcomes, even if they both adopt *FMTL* (Figure S4a–f). Such inefficiencies may arise more generally. They can occur in all games with equilibria that are fair but inefficient, and where one player alone is unable to raise the group payoff (see Supporting Information for analytical results). Nevertheless, we find that *FMTL* makes players less likely to settle at such inefficient equilibria in the first place. As a result, each player performs better on average if at least one of them adopts *FMTL*.

In the other generalized social dilemma, the snowdrift game, the payoff ranking is $T > R > S > P$. Mutual defection is no longer an equilibrium because unilateral cooperation is a better outcome for *both* players. When two selfish learners engage in a repeated snowdrift game, they often approach one of these two pure equilibria. Eventually, one player cooperates each round and the other defects (Figures S4g and S5a). Which of the two players ends up cooperating depends on their initial strategies and on chance. In contrast, if one of them switches to *FMTL*, players most likely coordinate on a pattern of play that is both fair and efficient. Similar to the prisoner's dilemma, this pattern requires players to either mutually cooperate (if $(S + T)/2 < R$; Figure S4h), or to cooperate in an alternating fashion (if $(S + T)/2 > R$; Figure S5b). In the latter case, already two selfish learners tend to achieve an efficient (albeit possibly unfair) outcome. The role of *FMTL* here, relative to selfish learning, is to eliminate inequality.

### *Alternative forms of conflict: hero game*

Finally, we consider an example that does not meet the conditions of a social dilemma: the hero game (15). This game, sometimes referred to as a (symmetric) battle of the sexes (75,76), satisfies $S > T > R \geq P$. Here, mutual $C$ is preferred to mutual $D$, but when both players use $C$, a single player can act as a "hero" and improve both players' payoffs by switching to $D$ (77). The one-shot game has two pure equilibria, but players disagree on which equilibrium they prefer. When two selfish learners engage in the repeated game, they frequently converge toward one of these pure one-shot equilibria (Figure S5d). In contrast, groups with at least one *FMTL* player reliably learn to alternate (Figure S5e and f). Selfish learning is able to generate efficient outcomes, but only *FMTL* makes sure the realized outcome is fair, irrespective of the learning rule of the opponent. Table S1 summarizes these results across the different games we study.

## Evolutionary dynamics of learning rules

After analyzing how different learning rules affect adaptation, we study how the learning rules themselves evolve over time. We consider two timescales. In the short run, the players' learning rules are fixed. Players use their learning rule to adapt to their opponent. In the long run, learning rules reproduce, based on how well players with the respective rule perform. This process may reflect cultural or biological evolution (i.e. successful learners are either imitated more often, or they have more offspring). Just as repeated games can be thought of as a "supergame" layered over a one-shot game (53) (Fig. 1a and b), the process describing the evolution of learning rules can be thought of as a supergame layered over the repeated game (Fig. 1c–f).

### *Description of the evolutionary process*

To describe this supergame formally, we consider a population of learners who use either selfish learning (S) or *FMTL* (F). In the short run, players are randomly matched to engage in repeated games with a fixed partner. Over the course of their interactions, they update their strategies according to their learning rules, as in the previous section. As a result, they receive a payoff that depends on the game being considered, the players' learning rules, and on the time that has passed for learning to unfold. For a given game, let $a_{ij}(n)$ denote the expected payoff of a learner $i$ against another learner $j$ after $n$ learning steps (i.e. after the players had $n$ opportunities to revise their strategies). We estimate these payoffs using numerical simulations.

For a given number of learning steps $n$, the four payoffs $a_{SS}(n)$, $a_{SF}(n)$, $a_{FS}(n)$, and $a_{FF}(n)$ can be assembled in a $2 \times 2$ payoff matrix. We interpret the entries of this matrix as the payoff of each learning rule, and we interpret $n$ as the players' learning horizon. Payoff matrices for different values of $n$ reflect different assumptions on how patient players are. For small $n$, players are impatient. They assess the quality of their learning rule by how well they perform after only a few learning steps. In contrast, for large $n$, players assess the quality of their learning rule according to how well it performs eventually (even if it may be ineffective in the short run).

For a given payoff matrix, we use the replicator equation to model the evolutionary dynamics among learning rules (78). The

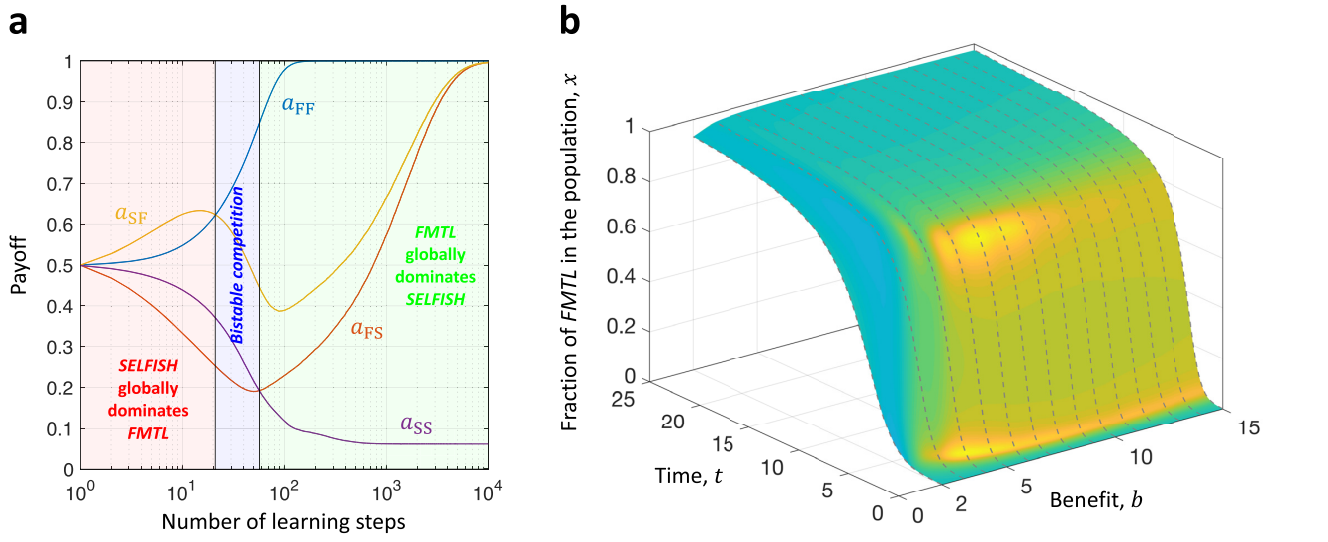


**Fig. 4.** Evolutionary dynamics of *FMTL* in the donation game. (a) In general, the players' payoffs depend on their learning rule and on how many learning steps they had to update their strategies. Here, we show these (supergame) payoffs for the donation game with $b = 2$ and $c = 1$. If the number of learning steps is small (fewer than $\approx 20$ in this example), the payoffs of *FMTL* and selfish learning take the form of a prisoner's dilemma: *FMTL* yields the better payoff when adopted by everyone, but it is dominated by selfish learning. With slightly more learning steps, the dynamics transition into bistable competition. Moreover, after only $\approx 60$ learning steps, *FMTL* globally dominates selfish learning. (b) Starting from an initial frequency of $x_0 = 10^{-3}$, *FMTL* quickly spreads in a population of selfish learners under replicator dynamics. Depicted here are evolutionary trajectories for the donation game when $c = 1$ and $b$ varies from 2 to 15. The dynamics in (b) are shown for the supergame payoffs after convergence of the learning process; but due to (a) similar results hold when the timescale of learning is much shorter.

equation tracks the frequencies of each learning rule over time, and it favors those rules with higher average payoffs (see "Methods"). For 2 × 2 games, as in our case, replicator dynamics can result in four different scenarios (10, 13). Either one learning rule is globally stable (dominance), there is a mixed population that is globally stable (coexistence), each learning rule is locally stable (bistability), or any mixed population is stable (neutrality).

### *Evolutionary dynamics across different games*

The resulting dynamics between the two learning rules depend on the game and on the players' learning horizon. Figure 4(a) shows the possible scenarios for the donation game. If the learning horizon is short, then learning rules are selected according to whether they result in an immediate advantage. As a result, we find that selfish learning is dominant, as one might expect. However, as players become increasingly patient, the dynamics first take the form of bistable competition, and then *FMTL* becomes globally stable. In this parameter regime, which starts after $n \approx 60$ learning steps, already a small initial fraction of *FMTL* learners is sufficient to drive selfish learning to extinction (Fig. 4b). Since clusters of *FMTL* learners perform better than clusters of selfish learners, the evolutionary advantage of *FMTL* is even more pronounced in structured populations (79, 80) (see "Methods" for details).

These patterns generalize to other games. In each case, selfish learning is dominant when players have a very short learning horizon. As players become sufficiently patient, *FMTL* becomes dominant in all variants of the prisoner's dilemma and the stag-hunt game, as well as in the snowdrift game when $(S + T)/2 < R$ (Figures S6 and S7). Only for the snowdrift game with $(S + T)/2 > R$ and the hero game may selfish learning prevail for long learning horizons. Table S2 gives a summary of these results. Interestingly, the games in which *FMTL* becomes globally stable are exactly those in which two selfish learners are at a considerable risk of settling at inefficient equilibria (Table S1). Conversely, the games in which selfish learning can prevail are those in which selfish learners tend to achieve efficient outcomes (Figure S5a and d). For example, in all simulated cases of the hero game, selfish learners settle at equilibria with maximum average payoffs (even though payoffs may be shared unfairly). Because replicator dynamics depend on only the average payoffs (not on the distribution of payoffs), and because selfish learning achieves the maximum average payoff against itself, *FMTL* cannot invade. This conclusion does not require replicator dynamics. Instead, it remains true for all evolutionary dynamics that depend on only a trait's average payoff, including stochastic models of weak selection (81).

Importantly, for *FMTL* to be successful, each of its components, fairness and efficiency, is vital. To illustrate this point, we repeat the previous simulations for the donation game with alternative learning rules (Figure S9): players either value only fairness, only efficiency, or they combine fairness with selfishness. Against a selfish learner, we find that each alternative rule performs worse than *FMTL*. When the focal player values only fairness, payoffs tend to be equal but inefficient (Figure S9a). When the focal player values only efficiency, the focal player tends to cooperate unconditionally, whereas the selfish coplayer defects (Figure S9b). Finally, when the focal player combines fairness and selfish learning, the overall performance is similar to the case of two selfish learners (Figure S9c). These results highlight that *FMTL*'s two components are effective only when combined. Learners who only aim for efficiency are subject to exploitation; learners who only value fairness obtain equal payoffs, but at the price of obtaining low payoffs.

## Selfishness versus fairness in a repeated prisoner's dilemma experiment

The results presented herein cast doubt on the assumption that selfish learning can fully explain human adaptation processes in

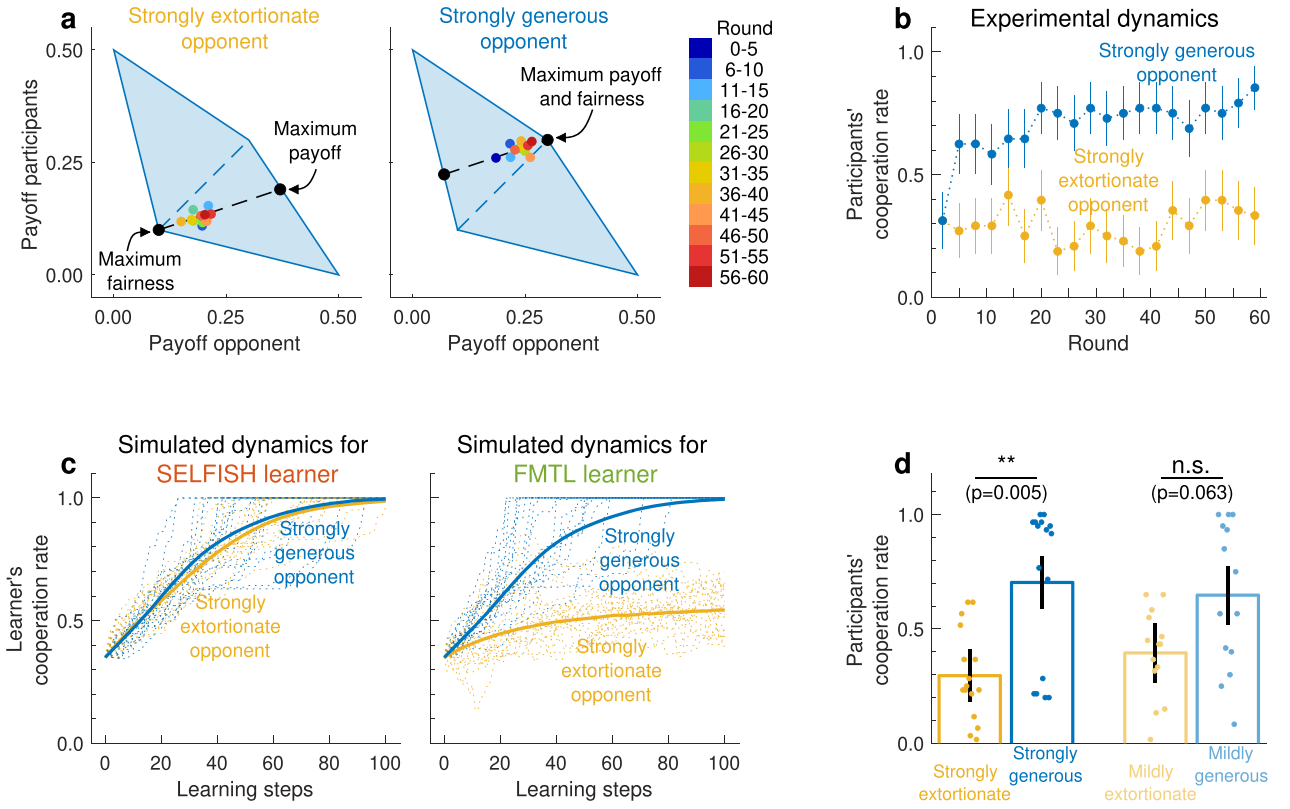


**Fig. 5.** Selfish learning fails to explain human behavior when there is a tension between payoff maximization and fairness. To empirically distinguish between selfish learning and fairness-mediated learning, we reanalyze data from a repeated prisoner's dilemma experiment (84). In this experiment, human participants interact with a computerized opponent who implements a fixed strategy. Human participants are not informed about the nature of their opponent. (a) The computer implements either an extortionate (20) or a generous (25) zero-determinant strategy. Such strategies ensure that there is a linear relationship between the payoffs of the two players (indicated by the black dashed line). If human participants wish to maximize their payoffs, they should learn to cooperate in either case. If participants wish to enhance fairness, they should cooperate against generous strategies but not against extortioners. (b) The experiment finds that humans become more cooperative against generous strategies. There is no trend toward cooperation against extortioners. (c) To shed further light on these observations, we simulate the possible learning dynamics. We consider a single learner, who either adopts selfish learning or *FMTL*. Only for *FMTL* do we recover that subjects fail to fully cooperate against extortioners. (d) The previous results are based on two treatments in which the trade-off between fairness and payoff maximization is strongest. When the computer instead implements so-called mildly generous and extortionate strategies, this trade-off is weaker. As a result, while generous coplayers still tend to induce more cooperation, the effect is now smaller. Error bars indicate standard errors, and any statistical statements are based on nonparametric tests (Mann–Whitney U test and Wilcoxon test). For details, see "Methods."

repeated games. While it has been suggested that selfish learning reasonably approximates cooperative behavior in linear multiplayer games (82,83), such games differ in crucial aspects from the two-player interactions studied herein. For example, in multiplayer games, each individual has less of an impact on other group members. Moreover, each player's defection can only be punished collectively (by withholding cooperation from *all* other group members). In contrast, pairwise interactions provide individuals with more immediate possibilities to affect their coplayer's future behavior. It is pairwise games where learning rules have the strongest strategic impact.

To explore the relevance of selfish learning for pairwise games, we reanalyze data from a prisoner's dilemma experiment (84). In this experiment, human participants play against a computer program (unbeknownst to the human subjects). The computer program either implements an extortionate (20) or a generous (25) "zero-determinant" strategy (20,21). Against both classes of strategies, the payoff-maximizing choice for human participants is to cooperate in every round. However, while full cooperation is also the fairness-maximizing choice against generous opponents, it leads to maximally unequal outcomes against extortioners (Fig. 5a).

There are two reasons why this experimental paradigm allows for a clean comparison between selfish learning and other learning rules based on other-regarding preference: (i) Because the computer's strategy is fixed (and known to the researcher), each human participant's learning behavior can be studied in isolation. (ii) The two competing learning rules make opposing predictions for this experiment: If human participants use selfish learning, they should equally learn to cooperate against either computer strategy. In contrast, if behavior is better described by *FMTL*, we expect more cooperation against the generous strategy (Fig. 5c; see "Methods" for a detailed description of the experimental setup and our predictions).

In Fig. 5(b), we compare experimental data for the "strongly extortionate" and the "strongly generous" computer strategy (for which fairness considerations are most likely to impede cooperation in the extortion treatment). For both treatments, human cooperation rates are similar in the beginning (31.3% during the first three rounds). Against the generous program, humans increase their cooperation rate to 85.4% by the end of the experiment (during the last three rounds). In contrast, against the extortionate program, overall cooperation rates are largely unchanged (33.3%), although the monetary incentives for cooperation are identical.

These experimental results are consistent with learning behavior that is shaped by fairness considerations. In line with this view, when the computer implements a strategy that is only mildly extortionate or generous, there is less of a difference in human cooperation rates (Fig. 5d). Moreover, the difference disappears altogether if individuals are informed ahead of the experiment that they are matched with a computerized opponent, in which case fairness considerations can be expected to be absent (85). Overall, these results suggest that fairness motives are of crucial importance to describe human behavior in pairwise interactions, even though they may be less salient in multiplayer games (82, 83).

## Discussion

Given the extensive effort to explore strategies that sustain cooperation in repeated social dilemmas (18–30), it seems quite remarkable that relatively little is known about how strategies with desirable properties can be learned most effectively. Instead, much of the existing work tends to take the way in which individuals learn as given. While details vary between studies (40–47), most often it is assumed that individuals adopt strategies that enhance their own payoff, abandoning strategies that are personally disadvantageous. This modeling assumption could be justified on theoretical grounds if selfish payoff maximization were indeed an optimal learning policy. Here, we have explored under which conditions selfish learning can be expected to succeed. We show theoretically that selfish learning performs well when individuals wish to optimize their short-run performance. However, if individuals are motivated by how well they fare eventually, selfish learning can be of limited use in navigating conflicts of interest.

To assess the performance of selfish learning, we contrast it with an alternative rule termed *FMTL*. Rather than maximizing just one's own payoff, a learner who adopts *FMTL* strives to enhance either the efficiency of the resulting game outcome or its fairness. By striving to increase efficiency, *FMTL* attempts to evade inferior equilibria that leave all group members worse off. By striving to increase fairness, *FMTL* avoids exploitation when other group members continue to learn selfishly. Using individual-based simulations, we show that *FMTL* can help individuals to settle at better equilibria. These equilibria are either more equitable or more efficient. Moreover, when players select their learning rules according to how well they perform eventually, *FMTL* outcompetes selfish learning for most of the games we study. Qualitatively, the ability of *FMTL* to draw out better outcomes also extends to learning based on imitation as opposed to introspection (Section S4 of Supporting Information; Figures S11 and S12) as well as asymmetric games (Section S5 of Supporting Information; Figure S13).

Our theoretical results are further corroborated by an analysis of human cooperation in a repeated prisoner's dilemma with fixed opponent strategies (84,85). When interactions entail a trade-off between fairness and payoff maximization, selfish learning fails to explain crucial patterns of human behavior (Fig. 5). There is a considerable literature within behavioral economics seeking to describe human behavior that deviates from models of pure self-interest (86). In particular, there is ample empirical evidence that fairness and efficiency are important drivers of human behavior. Humans value fairness starting from a young age (87, 88), and they are often willing to accept substantial reductions in their own income to achieve more egalitarian outcomes (89). Such a demand for fairness can have substantial economic consequences, as it constrains a firm's profit seeking behavior (90) and market prices (91). There is similar experimental evidence about the importance of efficiency (6, 92). In dictator games, human participants often give up some of their own payoff in order to increase that of the pair (86,93). At the same time, however, human decisions to increase efficiency via "gifts" are constrained by fairness considerations (94), consistent with the constraint built-in to *FMTL*.

Our results are related to the "indirect evolutionary approach," which explores the evolution of preferences with game-theoretic methods (95–98). The respective literature distinguishes between objective and subjective payoffs. Objective payoffs include monetary rewards and reproductive success (fitness). Subjective payoffs capture how individuals experience certain outcomes and what they strive to maximize when making decisions. Similar to our model, this approach is "indirect" because preferences that guide behaviors do not need to align with objective payoffs. However, while this literature explicitly models the evolutionary dynamics of preferences (98, 99), it usually does not describe how preferences affect the way individuals learn (see Supporting Information for a more detailed discussion). The learning process is the main focus of our study. We explore how different learning heuristics influence the way in which subjects navigate between equilibria of differing efficiency. To this end, our framework requires a notion of learning rules that is slightly more general than the notion of preferences considered before. Learning rules do not only specify the objectives that players wish to maximize. Instead, they also determine how players choose between different objectives and how they generate alternative strategies. In this way, our framework also applies to learning on shorter time scales, wherein subjects update their learning rules even before the learning dynamics reach an equilibrium (Table S2).

Our results also have implications for objective design in multiagent learning. Previous work has shown that objectives based on a convex combination of the players' payoffs can improve outcomes relative to selfish learning in stag hunt games (100). An alternative approach is to implement a look-ahead into the opponent's learning process in order to shape their future behavior (52) (however, this forward-looking approach requires substantial information about the opponent; see Supporting Information). While the space of possible objective functions is vast, we introduce a learning rule (one of many, perhaps) that can outcompete naïve selfish learning. Compared to the rules considered previously, *FMTL* has the advantage of being comparably simple, and its components are natural for both humans and machines to implement. Much in the same way that the most rudimentary strategy (tit-for-tat) won Axelrod's tournaments (55), here too a simple learning rule is able to align incentives with a selfish learner.

Even if individuals are ultimately driven by their own advantage, optimal learning rules may require them to take into account other considerations, such as the well-being of others. Of course, the repeated two-player games studied herein cannot capture all realistic interactions in which learning is relevant. However, these simple baseline models can help to understand the general principles at work in more complex settings (101). Delineating which aspects are crucial for successful learning is, in our view, one of the most exciting directions for future research.

## Methods

### Strategies and payoffs in repeated games

All of the strategies we consider for repeated games are "memory-one" strategies, which means that they consist of a five-tuple of probabilities, ($p_0$, $p_{CC}$, $p_{CD}$, $p_{DC}$, and $p_{DD}$), where $p_0$ is the probability of cooperating (action $C$) in the initial round and $p_{xy}$ is the probability a player cooperates after using $x \in \{C, D\}$ in the previous

round against an opponent using $y \in \{C, D\}$. The initial distribution over the outcomes ($CC$, $CD$, $DC$, and $DD$), where the first action is that of $X$ and the second is that of $Y$, is

$$\nu_0(\mathbf{p}, \mathbf{q}) = (p_0 q_0,\ p_0(1-q_0),\ (1-p_0)q_0,\ (1-p_0)(1-q_0)). \quad (1)$$

Following the initial round, transitions between states are described by the stochastic matrix

$$M(\mathbf{p}, \mathbf{q}) = \begin{pmatrix} p_{CC}q_{CC} & p_{CC}(1-q_{CC}) & (1-p_{CC})q_{CC} & (1-p_{CC})(1-q_{CC}) \\ p_{CD}q_{DC} & p_{CD}(1-q_{DC}) & (1-p_{CD})q_{DC} & (1-p_{CD})(1-q_{DC}) \\ p_{DC}q_{CD} & p_{DC}(1-q_{CD}) & (1-p_{DC})q_{CD} & (1-p_{DC})(1-q_{CD}) \\ p_{DD}q_{DD} & p_{DD}(1-q_{DD}) & (1-p_{DD})q_{DD} & (1-p_{DD})(1-q_{DD}) \end{pmatrix}. \quad (2)$$

After $t \geq 0$ rounds, the distribution over states is $\nu_t(\mathbf{p}, \mathbf{q}) = \nu_0(\mathbf{p}, \mathbf{q}) M(\mathbf{p}, \mathbf{q})^t$. With discounting factor $\lambda \in [0, 1)$, the mean payoffs to $X$ and $Y$ when $X$ plays $\mathbf{p}$ and $Y$ plays $\mathbf{q}$ are

$$\pi_X(\mathbf{p}, \mathbf{q}) = (1-\lambda) \sum_{t=0}^{\infty} \lambda^t < \nu_t(\mathbf{p}, \mathbf{q}), (R, S, T, P) >$$
$$= < (1-\lambda)\nu_0(\mathbf{p}, \mathbf{q})(I - \lambda M(\mathbf{p}, \mathbf{q}))^{-1}, (R, S, T, P) >; \quad (3a)$$

$$\pi_Y(\mathbf{p}, \mathbf{q}) = (1-\lambda) \sum_{t=0}^{\infty} \lambda^t < \nu_t(\mathbf{p}, \mathbf{q}), (R, T, S, P) >$$
$$= < (1-\lambda)\nu_0(\mathbf{p}, \mathbf{q})(I - \lambda M(\mathbf{p}, \mathbf{q}))^{-1}, (R, T, S, P) >, \quad (3b)$$

respectively, where $<\cdot, \cdot>$ denotes the standard inner (dot) product on $\mathbb{R}^4$. To approximate the results of an infinite-horizon game, we use a discounted game with a small discounting rate, $\lambda = 1 - 10^{-3}$. In this way, we ensure that the limiting payoffs always exist, even if the players adopt strategies that allow for multiple absorbing states, such as tit-for-tat (13).

## Learning rules

In the following, we describe the four components of a learning rule (the initial strategy distribution, the sampling procedure, the set of objective functions, and the priority assignment) for both selfish learning and *FMTL*.

### *Distribution over initial strategies*

When two individuals are first paired with one another, they each choose an initial strategy for their first interaction, which is to be subsequently revised during the learning process. Two of the most natural choices are (i) to choose each coordinate of the strategy independently from a uniform distribution on [0, 1], and (ii) to choose each coordinate independently from an arcsine (Beta(1/2, 1/2)) distribution on [0, 1]. For the figures presented herein, we use the latter distribution because it is more effective in exploring the corners of the space $[0, 1]^5$ of memory-one strategies (36). However, we obtain similar qualitative results for a uniform initial distribution. In addition, we also explore the learning dynamics that arise when players initially defect unconditionally (Fig. 3a and b).

### *Sampling procedure*

We assume strategy sampling to be local in the following sense (see also Fig. 1c). Let $s \in [0, 1]$ and suppose that $z_i$ is uniformly distributed on $[-s, s]$ (with $z_i$ independent of $z_j$ for $j \neq i$). If $p_i$ is the coordinate being "mutated" at a given time step, then the candidate sample of this coordinate in the next step is $p_i' = \min\{\max\{p_i + z_i, 0\}, 1\}$. We use $s = 0.1$ in our examples, which allows for exploration while ensuring that the candidate strategy is not too distant from the current strategy (so that desirable properties of the current strategy are not immediately discarded). Relatively small values of $s$ make the trajectories of the learning process more interpretable (e.g. Fig. 3a), but they also slow down the learning process. While taking a different value of $s$ can change the overall performance of each learning rule, we did not find a scenario in which it reverses the relative ranking of selfish learning compared to *FMTL*.

### *Objective functions*

Once a candidate strategy is sampled, the respective player decides whether to accept it based on the player's objectives. To this end, each learning rule specifies a set of objective functions,

$$\mathcal{V} = \left\{ V \mid V : [0, 1]^5 \times [0, 1]^5 \to \mathbb{R} \right\}. \quad (4)$$

Each objective function $V$ takes the players' memory-one strategies $\mathbf{p}$ and $\mathbf{q}$ as an input, and returns a value that indicates to which extent the players' objectives are met. Throughout our study, we consider objective functions that depend on only the players' payoffs, $\pi_X(\mathbf{p}, \mathbf{q})$ and $\pi_Y(\mathbf{p}, \mathbf{q})$, but more general formulations are possible . A candidate strategy $\mathbf{p}'$ is accepted if $V(\mathbf{p}', \mathbf{q}) > V(\mathbf{p}, \mathbf{q})$, and it is discarded otherwise.

For the simulations, we assume the candidate strategy is accepted if and only if $V(\mathbf{p}', \mathbf{q}) > V(\mathbf{p}, \mathbf{q}) + \varepsilon$, where $0 < \varepsilon \ll 1$ is a small threshold. This assumption prevents floating point errors from resulting in faulty decisions, particularly when $V(\mathbf{p}, \mathbf{q})$ is extremely close to $V(\mathbf{p}', \mathbf{q})$. While any one such faulty decision might have negligible effects on the learning process, these mistakes can accumulate over many time steps. In all of our numerical examples, the threshold we use is $\varepsilon = 10^{-12}$. The use of such a threshold can also be interpreted in terms of bounded rationality (102). Due to limitations on cognition or information, the learner may not be able to distinguish two values that are sufficiently close together. Similarly, in the presence of noise perturbing the observed payoffs, $\varepsilon$ may be thought of parametrizing the confidence an agent has that there will truly be an improvement in $V$ by switching to the new strategy.

We note that objective functions take both players' strategies as an input. One interpretation of this assumption is that the learner needs to have precise knowledge of the coplayer's strategy in order to forecast whether a given strategy change is profitable. However, we note that this rather stringent assumption is in fact not necessary. Instead, we only need to assume that individuals can reliably assess the sign of $V(\mathbf{p}', \mathbf{q}) - V(\mathbf{p}, \mathbf{q}) - \varepsilon$. That is, players only need to be able to make qualitative assessments.

For selfish learners, the set of objective functions contains a single element, $\mathcal{V} = \{V_S\}$ with $V_S(\mathbf{p}, \mathbf{q}) = \pi_X(\mathbf{p}, \mathbf{q})$. For *FMTL*, the set of objective functions is $\mathcal{V} = \{V_E, V_F\}$. Here, $V_E(\mathbf{p}, \mathbf{q}) = \pi_X(\mathbf{p}, \mathbf{q}) + \pi_Y(\mathbf{p}, \mathbf{q})$ reflects the objective to increase efficiency, whereas $V_F(\mathbf{p}, \mathbf{q}) = -|\pi_X(\mathbf{p}, \mathbf{q}) - \pi_Y(\mathbf{p}, \mathbf{q})|$ corresponds to the objective of enhancing fairness.

### *Priority assignments*

A learning rule's priority assignment $\Omega$ determines which of the players' different objectives is currently maximized. Formally, a priority assignment is a map $\Omega : [0, 1]^5 \times [0, 1]^5 \to \Delta(\mathcal{V})$, where $\Delta(\mathcal{V})$ denotes the space of probability distributions on $\mathcal{V}$. For each of the players' current memory-one strategies $\mathbf{p}, \mathbf{q} \in [0, 1]^5$ it determines the probability with which each possible objective $V \in \mathcal{V}$ is chosen for maximization.

In the case of selfish learning, the priority assignment is trivial, because there is only one possible objective to choose from. In the

case of *FMTL*, the priority assignment takes the form $\Omega(\mathbf{p}, \mathbf{q}) = (\omega, 1-\omega)$, where $\omega = \omega(\mathbf{p}, \mathbf{q})$ is the weight assigned to efficiency. In our formulation of *FMTL*, $\omega$ depends on the current magnitude of the payoff difference (see Figure S10a),

$$\omega(\mathbf{p}, \mathbf{q}) = \exp\left(-\frac{(\pi_X(\mathbf{p},\mathbf{q}) - \pi_Y(\mathbf{p},\mathbf{q}))^2}{2\sigma^2}\right). \quad (5)$$

The parameter $\sigma > 0$ reflects a player's tolerance with respect to inequality. In the limit $\sigma \to 0$, a player always aims to enhance fairness, whereas in the opposite limit $\sigma \to \infty$, the player always aims to improve efficiency. In general, neither efficiency nor fairness alone are sufficient for establishing good outcomes against selfish learners (Figure S9). Finding a proper balance between these two objectives depends on the nature of the interaction, and as a result the optimal value of $\sigma$ can vary from game to game.

We choose $\sigma$ by taking the average payoff for *FMTL* versus a selfish learner over a small number of runs ($10^3$), for each $\sigma \in \{0.01, 0.02, \ldots, 1.00\}$. We then select $\sigma$ based on which value maximizes the average payoff of the *FMTL* individual, except for when this value is comparable for all such $\sigma$, in which case we choose the value that minimizes the runtime. For the donation game (Figs 2c and d, 3a and b, and 4; Figures S1, S3, S9c, and S13; Video S1), we use $\sigma = 0.1$; for the prisoner's dilemma with $(S+T)/2 > R$ (Fig. 3c and d; Figure S6b) and with $(S+T)/2 < P$ (Figure S6a), we use $\sigma = 1$; for the stag hunt game with $(S+T)/2 > P$ (Figures S4a–c and S7a), we use $\sigma = 0.01$; for the stag hunt game with $(S+T)/2 < P$ (Figures S4d–f and S7b), we use $\sigma = 1$; for the snowdrift game with $(S+T)/2 < R$ (Figures S4g–i and S7c), we use $\sigma = 0.1$; for the snowdrift game with $(S+T)/2 > R$ (Figures S5a–c and S8a), we use $\sigma = 1$; and for the hero game (Figures S5d–f and S8b), we use $\sigma = 0.1$. The use of different values of $\sigma$ in different games is a result of treating $\sigma$ as a hyperparameter (103) that can be fine-tuned according to the nature of a given repeated game. However, we note that we get qualitatively similar (even if not completely optimized) results when we use a single value of $\sigma$ across all our main examples (e.g. $\sigma = 0.25$).

Instead of a bell curve, $\omega$ could be a bump function such as

$$\omega(\mathbf{p}, \mathbf{q}) = \begin{cases} e^{-\frac{(\pi_X(\mathbf{p},\mathbf{q}) - \pi_Y(\mathbf{p},\mathbf{q}))^2}{\alpha^2\sigma^2\left(\alpha^2 - (\pi_X(\mathbf{p},\mathbf{q}) - \pi_Y(\mathbf{p},\mathbf{q}))^2\right)}} & |\pi_X(\mathbf{p},\mathbf{q}) - \pi_Y(\mathbf{p},\mathbf{q})| < \alpha \\ 0 & |\pi_X(\mathbf{p},\mathbf{q}) - \pi_Y(\mathbf{p},\mathbf{q})| \geqslant \alpha \end{cases}, \quad (6)$$

(see Figure S10b). In practice, we do not see significant qualitative differences between these two functions provided the parameters are chosen properly. However, relative to Eq. (6), the function in Eq. (5) does have the advantage of depending on only one shape parameter, $\sigma$.

### *Implementation of learning rules*

For two players with given learning rules, we explore the resulting learning dynamics with simulations. The code is available (see "Data Availability" statement). For these simulations, the learning process is terminated after neither learner has accepted a new candidate strategy in a fixed threshold number of update steps (here, $10^4$). All of the examples we consider terminate.

In addition, Fig. 4(a) and Figures S6–S8 illustrate the expected payoffs as a function of the number of learning steps. In these graphs, the horizontal axis shows how many opportunities the two players had to revise their strategies. The vertical axis represents the average payoff over sufficiently many simulations. This payoff depends on the learning rules of the focal player and the opponent.

## Evolutionary dynamics of learning rules

The learning rules considered here may be viewed as strategies for a "supergame." The payoffs of this supergame are given by the players' expected payoffs after $n$ updating steps. Here, we consider $n$ as a fixed parameter of the model, and we refer to it as the player's learning horizon. When $n$ is small, individuals evaluate the performance of their learning rule based on how well they fare after a few learning steps. In contrast, when $n$ is large, learning rules are evaluated according to how well they fare eventually.

Because we consider the competition between two learning rules, this supergame can be represented as a $2 \times 2$ matrix. To this end, we fix a base game $G \in \{HE, PD, SH, SD\}$, where HE is the hero game, PD is the prisoner's dilemma, SH is the stag hunt game, and SD is the snowdrift game. For a given learning horizon $n$, we can write the resulting $2 \times 2$ matrix as

$$\begin{array}{c} \\ FMTL \\ SELFISH \end{array} \begin{array}{c} \begin{array}{cc} FMTL & SELFISH \end{array} \\ \begin{pmatrix} a^G_{FF}(n) & a^G_{FS}(n) \\ a^G_{SF}(n) & a^G_{SS}(n) \end{pmatrix} \end{array}. \quad (7)$$

When the game and the learning horizon is clear (or irrelevant), we sometimes drop the indices $G$ and $n$ for better readability.

### *Evolutionary dynamics in well-mixed populations*

Given the payoff matrix, we explore the evolutionary dynamics between learning rules using the replicator equation (78). The replicator equation describes deterministic evolution in infinite populations. Let $x \in [0, 1]$ denote the frequency of *FMTL* in the population. The frequency of selfish learners is $1 - x$. Using the payoffs of Eq. (7), the fitness values of the two types are

$$f_F(x) = x a_{FF} + (1-x) a_{FS}; \quad (8a)$$

$$f_S(x) = x a_{SF} + (1-x) a_{SS}, \quad (8b)$$

respectively. The average fitness in the population is $\overline{f}(x) = x f_F(x) + (1-x) f_S(x)$. Under replicator dynamics, the frequency of *FMTL* satisfies the ordinary differential equation

$$\frac{dx}{dt} = x\left(f_F(x) - \overline{f}(x)\right). \quad (9)$$

When $a_{FF} \geq a_{SF}$ and $a_{FS} \geq a_{SS}$, we have $f_F(x) \geq f_S(x)$, and thus $f_F(x) \geqslant \overline{f}(x)$ for all $x \in (0, 1)$. Moreover, if one of the inequalities is strict, $a_{FF} > a_{SF}$ or $a_{FS} > a_{SS}$, then $f_F(x) > \overline{f}(x)$. This payoff relationship holds in prisoner's dilemmas and stag hunt games, as well as in the snowdrift game with $(S+T)/2 < R$, when $n$ is sufficiently large. In that case, there are two equilibria. The equilibrium $x = 0$ is unstable, whereas $x = 1$ is globally stable. It follows that *FMTL* evolves from all initial populations with $x > 0$. Figure 4(b) illustrates these dynamics for the donation game as the benefit of cooperation varies. It is worth pointing out that the relative value of $a_{SF}$ compared to $a_{FS}$ does not affect replicator dynamics. More precisely, even though a selfish learner always gets at least the coplayer's payoff in any interaction with *FMTL* (i.e. even though $a_{SF} \geq a_{FS}$ for all games we studied), selfish learning may still go extinct.

### *Evolutionary dynamics in structured populations*

The classical replicator equation describes populations in which all individuals are equally likely to interact with each other. To explore how population structure is expected to affect our evolutionary results, we use the approach of Ohtsuki and Nowak (104). They show that under weak selection, various stochastic evolutionary processes on regular graphs can be approximated by a

replicator equation with modified payoffs. Instead of $a_{ij}$, the payoff of strategy $i$ against strategy $j$ is now given by $\widetilde{a}_{ij} := a_{ij} + b_{ij}$. Here, $b_{ij}$ depends on the game, the evolutionary process under consideration, and the degree $k > 2$ of the network. For example, for death-birth updating,

$$b_{ij} = \frac{(k+1)\,a_{ii} + a_{ij} - a_{ji} - (k+1)\,a_{jj}}{(k+1)\,(k-2)}. \tag{10}$$

When we apply this formula to the four payoffs $a_{FF}$, $a_{FS}$, $a_{SF}$, and $a_{SS}$, the modified payoffs are

$$\widetilde{a}_{FF} = a_{FF}; \tag{11a}$$

$$\widetilde{a}_{FS} = a_{FS} + \tfrac{(k+1)a_{FF} + a_{FS} - a_{SF} - (k+1)a_{SS}}{(k+1)(k-2)}; \tag{11b}$$

$$\widetilde{a}_{SF} = a_{SF} + \tfrac{(k+1)a_{SS} + a_{SF} - a_{FS} - (k+1)a_{FF}}{(k+1)(k-2)}; \tag{11c}$$

$$\widetilde{a}_{SS} = a_{SS}. \tag{11d}$$

For replicator dynamics, only payoff differences matter. These can be written as follows:

$$\begin{aligned}\widetilde{a}_{FF} - \widetilde{a}_{SF} = {} & \left(1 + \frac{1}{(k+1)\,(k-2)}\right)(a_{FF} - a_{SF}) \\ & + \frac{1}{(k+1)\,(k-2)}\,(a_{FS} - a_{SS}) \\ & + \frac{k}{(k+1)\,(k-2)}\,(a_{FF} - a_{SS});\end{aligned} \tag{12a}$$

$$\begin{aligned}\widetilde{a}_{FS} - \widetilde{a}_{SS} = {} & \left(1 + \frac{1}{(k+1)\,(k-2)}\right)(a_{FS} - a_{SS}) \\ & + \frac{1}{(k+1)\,(k-2)}\,(a_{FF} - a_{SF}) \\ & + \frac{k}{(k+1)\,(k-2)}\,(a_{FF} - a_{SS}).\end{aligned} \tag{12b}$$

In a well-mixed population, the condition for *FMTL* to be globally stable is $a_{FF} \geq a_{SF}$ and $a_{FS} \geq a_{SS}$, with at least one inequality being strict. In the examples where we observe *FMTL* to globally stable, the inequality $a_{FF} > a_{SS}$ is also satisfied (Fig. 4a; Figures S6 and S7). For those games, Eq. (12) allows us to conclude that if *FMTL* is globally stable in a well-mixed population, then it is also stable in any regular network. Moreover, especially if the degree $k$ of the network is small, regular networks may require a smaller learning horizon for *FMTL* to become globally stable.

## Empirical analysis of a repeated prisoner's dilemma experiment

### *Experimental methods*

To compare our theoretical predictions to actual human behavior, we reanalyze the experimental data collected by Hilbe et al. (84). In this experiment, human subjects play 60 rounds of a repeated prisoner's dilemma against a computer program. Humans are not told the nature of their opponent; instead, they only learn that they are "matched with an opponent with whom they will interact for many rounds." In each round, participants can either cooperate or defect. The payoffs per round are derived from the payoffs used in Axelrod's tournament (19):

$$R = 0.30\,\text{EUR};\quad S = 0.00\,\text{EUR};\quad T = 0.50\,\text{EUR};\quad P = 0.10\,\text{EUR}. \tag{13}$$

In addition, all participants get a show-up fee (independent of their performance) of 10 EUR. All participants are first-year biology students recruited from the universities of Kiel and Hamburg, Germany.

The experiment consists of four treatments, which only differ in the memory-one strategies implemented by the computer program. The four strategies are referred to as being "strongly extortionate," "mildly extortionate," "mildly generous," and "strongly generous," and they are specified as follows:

$$\begin{aligned}&\text{Strongly extortionate:} && p_0 = 0,\ \mathbf{p} = (0.692,\ 0,\ 0.538,\ 0);\\ &\text{Mildly extortionate:} && p_0 = 0,\ \mathbf{p} = (0.857,\ 0,\ 0.786,\ 0);\\ &\text{Mildly generous:} && p_0 = 1,\ \mathbf{p} = (1,\ 0.077,\ 1,\ 0.154);\\ &\text{Strongly generous:} && p_0 = 1,\ \mathbf{p} = (1,\ 0.182,\ 1,\ 0.364).\end{aligned} \tag{14}$$

For the given payoff values, these four strategies represent so-called zero-determinant strategies (20,21). By using one of these strategies, the computer program ensures that there is an approximately linear relationship between the payoff of the human participant $\pi_H$ and the payoff of the computer program $\pi_C$. The respective linear relationships are (84)

$$\begin{aligned}&\text{Strongly extortionate:} && \pi_H - P = \tfrac{1}{3}\,(\pi_C - P);\\ &\text{Mildly extortionate:} && \pi_H - P = \tfrac{2}{3}\,(\pi_C - P);\\ &\text{Mildly generous:} && \pi_H - R = \tfrac{2}{3}\,(\pi_C - R);\\ &\text{Strongly generous:} && \pi_H - R = \tfrac{1}{3}\,(\pi_C - R).\end{aligned} \tag{15}$$

The interpretation of these payoff relationships is as follows. If the program is extortionate, the human coplayer's surplus (over the mutual defection payoff) is only one-third or two-thirds of the computer's surplus. On the other hand, if the program is generous, the human coplayer's loss (compared to the payoff for mutual cooperation) is only one-third or two-thirds of the computer's loss. In particular, an extortionate computer program always obtains at least as much as the human coplayer, whereas a generous program only obtains at most the payoff of the human coplayer (21). In Fig. 5(a), the payoff relationships for the strong strategy variants are indicated by a black dashed line. In the extortionate case, the black dashed line is always on or below the main diagonal. In the generous case, the black dashed line is always on or above the main diagonal. The colored dots represent experimental data, averaged over all participants of the respective treatment. Each dot indicates the payoff of the human participant and the computer opponent for consecutive five-round intervals. Overall, the study has been conducted with 60 participants (16 participants in each of the two strong treatments and 14 participants in each of the two mild treatments).

### *Theoretical predictions*

For the interpretation of the experimental results with respect to our theoretical framework, three aspects of the strategies are crucial.

(1) According to Eq. (15), there is a positive linear relationship between the payoffs of the computer program and the human participant. In particular, if humans wish to maximize their own payoff, they should strive to maximize their opponent's payoff. That is, they should cooperate in every single round.
(2) In contrast, if humans wish to have equal payoffs $\pi_H = \pi_C$ against the extortionate program, they should defect in every round (in which case $\pi_H = \pi_C = P$). If they wish to have equal payoffs against the generous program, humans should cooperate in every round (in which case $\pi_H = \pi_C = R$).
(3) Importantly, the monetary incentives for humans to become more cooperative are the same in the strongly extortionate and in the strongly generous treatment. In either case, for every cent that they increase the coplayer's payoff by being more cooperative, their own payoff is increased by

one-third of a cent. Similarly, the monetary incentives for increasing cooperation in the two mild treatments are also identical.

Because of the first and the third property, we would predict that participants who wish to increase their own payoffs cooperate equally often, independent of whether they are matched with a strongly extortionate or with a strongly generous opponent. An analogous prediction applies to the two mild treatments. In contrast, because of the second property, we would predict that participants who wish to enhance fairness are more likely to cooperate in the generous treatments. These predictions are general; they do not depend on the exact implementation of the participants' learning rules.

In addition to these qualitative predictions, we have also run simulations for the learning rules studied herein. In Fig. 5(c), we show the average of 1,000 simulations (bold solid lines) as well as a sample of 20 representative simulation runs (thin dotted lines). For each simulation, there is one learner who either implements selfish learning or *FMTL*. Consistent with our theoretical analysis, the learner is restricted to memory-one strategies. Initially, the learner's strategy is unconditional, with $p_0 = 0.35$ and $\mathbf{p} = (0.35,\ 0.35,\ 0.35,\ 0.35)$. In each simulation, the learner is given 100 opportunities to revise its memory-one strategy. New strategies are sampled within an $s = 0.1$-neighborhood of the parent strategy. For *FMTL*, we use a sensitivity parameter of $\sigma = 0.1$, as in our main text analysis of the standard prisoner's dilemma. The learner's opponent applies a fixed memory-one strategy that is either strongly extortionate or strongly generous, as specified by Eq. (14). Consistent with the qualitative predictions above, we find that selfish learners equally learn to cooperate no matter whether the opponent is extortionate or generous. In contrast, the *FMTL* player only learns to fully cooperate against a generous opponent.

### *Statistical analysis*

To test these predictions, we first compare the cooperation dynamics against the strongly extortionate strategy and the strongly generous strategy (participants matched with the strongly extortionate strategy face the strongest trade-off between payoff maximization and fairness). In Fig. 5(b), each dot represents the human participants' cooperation rates, averaged over three rounds. In particular, the panel shows that initially participants are equally likely to cooperate against both computer programs (31.3% during the first three rounds for both). However, only for the strong generosity treatment is there a significant increase in cooperation rates (to 85.4% during the last three rounds; Wilcoxon matched-pairs signed-rank test: $Z = 2.9341$, $P = 0.003$, all tests are two-tailed). In contrast, in the strong extortion treatment, there is no such increase (33.3%; Wilcoxon matched-pairs signed-rank test: $Z = 0.3494$, $P = 0.726$.).

We obtain similar results if we compare overall cooperation rates over all 60 rounds (Fig. 5d). Against the strongly extortionate program, this average cooperation rate is 29.6%, compared to 70.3% against the strongly generous program (Mann–Whitney U test, $Z = 2.789$, $P = 0.005$). For the two mild treatments, the difference in cooperation rates is smaller, and fails to be significant; however, players still tend to cooperate more against the generous program (the cooperation rates are 39.5% and 64.8%, Mann–Whitney U test, $Z = 1.860$, $P = 0.063$). Overall, these results suggest that fairness considerations affect human cooperation rates. Moreover, this effect is more pronounced the stronger the trade-off between payoff maximization and fairness.

## Acknowledgments

The authors are grateful to Jörg Oechssler for many helpful comments.

## Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

## Funding

A.M. was supported by a Simons Postdoctoral Fellowship (Math+X) at the University of Pennsylvania; K.C. was supported by the European Research Council Consolidator Grant 863818 (ForM-SMArt); and C.H. was supported by the European Research Council Starting Grant 850529 (E-DIRECT).

## Authors' Contributions

All authors designed research, performed research, and wrote the paper.

## Data Availability

Supporting code is available at https://github.com/alexmcavoy/fmtl/.

## References

1. Traulsen A, Semmann D, Sommerfeld RD, Krambeck HJ, Milinski M. 2010. Human strategy updating in evolutionary games. Proc Natl Acad Sci USA. 107:2962–2966.
2. Rand DG, Nowak MA. 2012. Human cooperation. Trends Cogn Sci. 117:413–425.
3. Vulic M, Kolter R. 2001. Evolutionary cheating in *Escherichia coli* stationary phase cultures. Genetics. 158(2):519–526.
4. Zomorrodi AR, Segrè D. 2017. Genome-driven evolutionary game theory helps understand the rise of metabolic interdependencies in microbial communities. Nat Commun. 8(1):1563.10.1038/s41467-017-01407-5
5. Fehr E, Schmidt KM. 1999. A theory of fairness, competition, and cooperation. Quart J Econ. 114(3):817–868.
6. Charness G, Rabin M. 2002. Understanding social preferences with simple tests. Quart J Econ. 117(3):817–869.
7. Fischbacher U, Gächter S. 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. Am Econ Rev. 100(1):541–556.
8. Bloembergen D, Tuyls K, Hennes D, Kaisers M. 2015. Evolutionary dynamics of multi-agent learning: a survey. J Artif Int Res. 53:659–697.
9. Zhang K, Yang Z, Başar T. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Handbook of Reinforcement Learning and Control. Vamvoudakis KG, Wan Y, Lewis FL, Cansever D (eds), Cham, Switzerland: Springer International Publishing. pp. 321-384.
10. Hofbauer J, Sigmund K. 1988. The theory of evolution and dynamical systems. Cambridge: Cambridge University Press.
11. Friedman D. 1991. Evolutionary games in economics. Econometrica. 59(3):637.
12. Weibull JW. 1995. Evolutionary game theory. Cambridge (MA): MIT Press.

13. Sigmund K. 2010. The calculus of selfishness. Princeton (NJ): Princeton University Press.
14. McNamara JM. 2013. Towards a richer evolutionary game theory. J Roy Soc Int. 10(88):20130544.
15. Tanimoto J. 2015. Fundamentals of evolutionary game theory and its applications. Japan: Springer.
16. Javarone MA. 2018. Statistical physics and computational methods for evolutionary game theory. Cham: Springer International Publishing.
17. Newton J. 2018. Evolutionary game theory: a renaissance. Games. 9(2):31.
18. Trivers RL. 1971. The evolution of reciprocal altruism. Quart Rev Biol. 46(1):35–57.
19. Axelrod R, Hamilton W. 1981. The evolution of cooperation. Science. 211(4489):1390–1396.
20. Press WH, Dyson FJ. 2012. Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. Proc Natl Acad Sci. 109(26):10409–10413.
21. Hilbe C, Chatterjee K, Nowak MA. 2018. Partners and rivals in direct reciprocity. Nat Human Behav. 2:469–477.10.1038/s41562-018-0320-9
22. Stewart AJ, Plotkin JB. 2012. Extortion and cooperation in the prisoner's dilemma. Proc Natl Acad Sci. 109(26):10134–10135.
23. van Segbroeck S, Pacheco JM, Lenaerts T, Santos FC. 2012. Emergence of fairness in repeated group interactions. Phys Rev Lett. 108:158104.
24. Fischer I, *et al.* 2013. Fusing enacted and expected mimicry generates a winning strategy that promotes the evolution of cooperation. Proc Natl Acad Sci. 110:10229–10233.
25. Stewart AJ, Plotkin JB. 2013. From extortion to generosity, evolution in the iterated prisoner's dilemma. Proc Natl Acad Sci. 110(38):15348–15353.
26. Pinheiro FL, Vasconcelos VV, Santos FC, Pacheco JM. 2014. Evolution of all-or-none strategies in repeated public goods dilemmas. PLoS Comput Biol. 10(11):e1003945.
27. Akin E. 2015. What you gotta know to play good in the iterated prisoner's dilemma. Games. 6(3):175–190.
28. Yi SD, Baek SK, Choi JK. 2017. Combination with anti-tit-for-tat remedies problems of tit-for-tat. J Theor Biol. 412:1–7.
29. Hilbe C, Martinez-Vaquero LA. Chatterjee K, Nowak MA. 2017. Memory-*n* strategies of direct reciprocity. Proc Natl Acad Sci USA. 114:4715–4720.
30. McAvoy A, Nowak MA. 2019. Reactive learning strategies for iterated games. Proc R Soc A Math Phys Eng Sci. 475(2223):20180819.
31. Ohtsuki H, Iwasa Y. 2004. How should we define goodness? – Reputation dynamics in indirect reciprocity. J Theor Biol. 231:107–20.
32. Santos FP, Santos FC, Pacheco JM. 2018. Social norm complexity and past reputations in the evolution of cooperation. Nature. 555:242–245.
33. Javarone MA, Marinazzo D. 2017. Evolutionary dynamics of group formation. PLoS ONE. 12(11):e0187960.
34. Abdallah S, et al. 2014. Corruption drives the emergence of civil society. J R Soc Int. 11:20131044.
35. Lee Y, Iwasa Y, Dieckmann U, Sigmund K. 2019. Social evolution leads to persistent corruption. Proc Natl Acad Sci USA. 116(27):13276–13281.
36. Nowak M, Sigmund K. 1993. A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. Nature. 364(6432):56–58.
37. Zhong F, Kimbrough SO, Wu DJ. 2002. Cooperative agent systems: artificial agents play the ultimatum game. Proceedings of the 35th Annual Hawaii International Conference on System Sciences. Big Island (HI): IEEE Computer Society.
38. Batut B, Parsons DP, Fischer S, Beslon G, Knibbe C. 2013. In silico experimental evolution: a tool to test evolutionary scenarios. BMC Bioinf. 14(Suppl 15):S11.
39. Kiourt C, Kalles D. 2016. Synthetic learning agents in game-playing social environments. Adapt Behav. 24(6):411–427.
40. Szabó G, Tőke C. 1998. Evolutionary prisoner's dilemma game on a square lattice. Phys Rev E. 58:69–73.
41. Traulsen A, Pacheco JM, Nowak MA. 2007. Pairwise comparison and selection temperature in evolutionary game dynamics. J Theor Biol. 246:522–529.
42. Amaral MA, Javarone MA. 2018. Heterogeneous update mechanisms in evolutionary games: mixing innovative and imitative dynamics. Phys Rev E. 97(4):042305.
43. Oechssler J., 2002. Cooperation as a result of learning with aspiration levels. J Econ Behav Org. 49:405–409.
44. Du J, Wue B, Altrock PM, Wang L. 2014. Aspiration dynamics of multi-player games in finite populations. J Roy Soc Int. 11(94):1742–5662.
45. Sandholm TW, Crites RH., 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. BioScience. 37:147–166.
46. Masuda N, Ohtsuki H. 2009. A theoretical analysis of temporal difference learning in the iterated prisoner's dilemma game. Bull Math Biol. 71(8):1818–1850.
47. Hauser O, Hilbe C, Chatterjee K, Nowak MA. 2019. Social dilemmas among unequals. Nature. 572:524—527.
48. Couto MC. Giaimo S Hilbe C 2022. Introspection dynamics: a simple model of counterfactual learning in asymmetric games. New J. Phys.. 24(6):063010
49. Dawes RM. 1980. Social dilemmas. Ann Rev Psychol. 31(1):169–193.
50. Kerr B, Godfrey-Smith P, Feldman MW. 2004. What is altruism?. Trends Ecol Evol. 19(3):135–140.
51. Nowak MA. 2012. Evolving cooperation. J Theor Biol. 299:1–8.
52. Foerster J *et al.* 2018. Learning with opponent-learning awareness. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. p. 122–130. Richland (SC): International Foundation for Autonomous Agents and Multiagent Systems.
53. Friedman J. 1971. A non-cooperative equilibrium for supergames. Rev Econ Stud. 38:1–12.
54. Stewart AJ, Plotkin JB. 2014. Collapse of cooperation in evolving games. Proc Natl Acad Sci USA. 111(49):17558–17563.
55. Axelrod R. 1984. The evolution of cooperation. New York (NY): Basic Books.
56. Stewart AJ, Plotkin JB. 2016. Small groups and long memories promote cooperation. Sci Rep. 6:26889.
57. Bowling M, Veloso M. 2002. Multiagent learning using a variable learning rate. Artif Int. 136(2):215–250.
58. Tuyls K, Hoen PJT, Vanschoenwinkel B. 2005. An evolutionary dynamical analysis of multi-agent learning in iterated games. Auton Agent Multi-Agent Syst. 12(1):115–153.
59. Shoham Y, Powers R, Grenager T. 2007. If multi-agent learning is the answer, what is the question?. Artif Int. 171(7): 365–377.
60. Stone P. 2007. Multiagent learning is not the answer. It is the question. Artif Int. 171(7):402–405.
61. Tuyls K, Weiss G. 2012. Multiagent learning: basics, challenges, and prospects. AI Mag. 33(3):41.
62. Hu J, Wellman MP. 2003. Nash Q-learning for general-sum stochastic games. J Mach Learn Res. 4:1039–1069.

63. Hoen PJ, Tuyls K, Panait L, Luke S, La Poutré JA. 2006. An overview of cooperative and competitive multiagent learning. In: Learning and adaption in multi-agent systems. Berlin Heidelberg: Springer, p. 1–46.
64. Engle-Warnick J, Slonim RL. 2006. Inferring repeated-game strategies from actions: evidence from trust game experiments. Econ Theor. 28(3):603–632.
65. Dal Bó P, Fréchette GR. 2011. The evolution of cooperation in infinitely repeated games: experimental evidence. Am Econ Rev. 101(1):411–429.
66. Bruttel L, Kamecke U. 2011. Infinity in the lab. How do people play repeated games?. Theor Decis. 72(2):205–219.
67. Dal Bó P, Fréchette GR. 2018. On the determinants of cooperation in infinitely repeated games: a survey. J Econ Lit. 56(1):60–114.
68. Solis FJ, Wets RJB. 1981. Minimization by random search techniques. Math Operat Res. 6(1):19–30.
69. Boerlijst MC, Nowak MA, Sigmund K. 1997. Equal pay for all prisoners. Am Math Month. 104:303–307.
70. Martinez-Vaquero LA, Cuesta JA, Sanchez A. 2012. Generosity pays in the presence of direct reciprocity: a comprehensive study of 2x2 repeated games. PLoS ONE. 7(4):E35135.
71. Hauert C, Michor F, Nowak MA, Doebeli M. 2006. Synergy and discounting of cooperation in social dilemmas. J Theor Biol. 239(2):195–202.
72. Skyrms B. 2003. The stag hunt and the evolution of social structure. Cambridge: Cambridge University Press.
73. Sugden R. 1986. The economics of rights, co-operation, and welfare. Oxford: B. Blackwell.
74. Hauert C, Doebeli M. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature. 428(6983):643–646.
75. Maynard Smith J. 1982. Evolution and the theory of games. Cambridge: Cambridge University Press.
76. Luce RD, Raiffa H. 1989. Games and decisions: introduction and critical survey. In: Dover books on mathematics. Mineola (NY): Dover Publications. ISBN 9780486659435.
77. Rapoport A. 1967. Exploiter, leader, hero, and martyr: the four archetypes of the 2 × 2 game. Behav Sci. 12(2):81–84.
78. Taylor PD, Jonker LB. 1978. Evolutionary stable strategies and game dynamics. Math Biosci. 40(1-2):145–156.
79. Nowak MA, Tarnita CE, Antal T. 2010. Evolutionary dynamics in structured populations. Phil Trans R Soc B. 365:19–30.
80. Perc M, Gómez-Gardeñes J, Szolnoki A, Floría LM, Moreno Y. 2013. Evolutionary dynamics of group interactions on structured populations: a review. J R Soc Int. 10(80):20120997.
81. McAvoy A, Allen B, Nowak MA. 2020. Social goods dilemmas in heterogeneous societies. Nat Human Behav. 4(8):819–831.
82. Burton-Chellew MN, Nax HH, West SA. 2015. Payoff-based learning explains the decline in cooperation in public goods game. Proc R Soc B. 282(1801):20142678.
83. Burton-Chellew MN, West SA. 2021. Payoff-based learning best explains the rate of decline in cooperation across 237 public-goods games. Nat Human Behav. 5:1330–1338.10.1038/s41562-021-01107-7
84. Hilbe C, Röhl T, Milinski M. 2014. Extortion subdues human players but is finally punished in the prisoner's dilemma. Nat Commun. 5:3976.
85. Xu B, Zhou Y, Lien JW, Zheng J, Wang Z. 2016. Extortion can outperform generosity in iterated prisoner's dilemma. Nat Commun. 7:11125.
86. Fehr E, Schmidt KM. 2006. The economics of fairness, reciprocity and altruism – experimental evidence and new theories. In: Handbook of the economics of giving, altruism and reciprocity. Amsterdam: Elsevier, p. 615–691.
87. Fehr E, Bernhard H, Rockenbach B. 2008. Egalitarianism in young children. Nature. 454(7208):1079–1083.
88. McAuliffe K, Blake PR, Steinbeis N, Warneken F. 2017. The developmental foundations of human fairness. Nat Human Behav. 1(2):0042.
89. Dawes CT, Fowler JH, Johnson T, McElreath R, Smirnov O. 2007. Egalitarian motives in humans. Nature. 446:794–796.
90. Kahneman D, Knetsch JL, Thaler R. 1986. Fairness as a constraint on profit seeking: entitlements in the market. Am Econ Rev. 76(4):728–741.
91. Fischbacher U, Fong CM, Fehr E. 2009. Fairness, errors and the power of competition. J Econ Behav Org. 72(1):527–545.
92. Engelmann D, Strobel M. 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Am Econ Rev. 94(4):857–869.
93. Andreoni J, Miller J. 2002. Giving according to GARP: an experimental test of the consistency of preferences for altruism. Econometrica. 70(2):737–753.
94. Güth W, Kliemt H, Ockenfels A. 2003. Fairness versus efficiency: an experimental study of (mutual) gift giving. J Econ Behav Org. 50(4):465–475.
95. Güth W. 1995. An evolutionary approach to explaining cooperative behavior by reciprocal incentives. Int J Game Theor. 24(4):323–344.
96. Güth W, Kliemt H. 1998. The indirect evolutionary approach: bridging the gap between rationality and adaptation. Ration Soc. 10(3):377–399.
97. Huck S, Oechssler J. 1999. The indirect evolutionary approach to explaining fair allocations. Games Econ Behav. 28(1):13–24.
98. Heifetz A, Shannon C, Spiegel Y. 2007. The dynamic evolution of preferences. Econ Theor. 32(2):251–286.
99. Akçay E, Van Cleve J, Feldman MW, Roughgarden J. 2009. A theory for the evolution of other-regard integrating proximate and ultimate perspectives. Proc Natl Acad Sci. 106(45):19061–19066.
100. Peysakhovich A, Lerer A. 2018. Prosocial learning agents solve generalized stag hunts better than selfish ones. Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. Richland (SC): International Foundation for Autonomous Agents and Multiagent Systems, p. 2043–2044.
101. Smaldino PE. 2017. Models are stupid, and we need more of them. In: Computational social psychology. London: Routledge, p. 311–331.
102. Simon HA. 1957. Models of man: social and rational; mathematical essays on rational human behavior in a social setting. Hoboken (NJ) Wiley.
103. Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. J Mach Learn Res. 13(10):281–305.
104. Ohtsuki H, Nowak MA. 2006. The replicator equation on graphs. J Theor Biol. 243:86–97.