Learning to cooperate: The evolution of social rewards in repeated interactions

Slimane Dridi and Erol Akçay

Department of Biology, University of Pennsylvania,
433 South University Avenue, Philadelphia, PA 19104, USA

September 8, 2016

Abstract

Understanding the behavioral and psychological mechanisms underlying social behaviors is one of the major outstanding goals of social evolutionary theory. In particular, a persistent question about human–and animal–cooperation is to what extent it is supported by other-regarding preferences–the motivation to increase the welfare of others. In real-world situations, individuals have the opportunity to learn from past interactions, so we may ask how individuals evolve to learn to satisfy their social preferences during the course of an interaction. In this context, the rewards an individual receives from his social behaviors capture his preferences. In this paper, we develop a mathematical model in order to ask whether the mere act of cooperating with a social partner will evolve to be inherently rewarding. In our model, individuals interact repeatedly in pairs and adjust their behaviors through reinforcement learning. Individuals associate to each game outcome a subjective utility, which constitute the reward for that particular outcome. We find that utilities that value mutual cooperation positively but the sucker's outcome negatively are evolutionarily stable. In a reduced model, other-regarding preferences can co-exist with preferences that match the sign of the material payoffs under narrow conditions. In simulations of the full model, we find that selfish preferences that always learn pure defection are also evolutionarily successful.

These findings are consistent with empirical observations showing that humans tend to behave according to distinct behavioral types, and call for further integration of different levels of biological and social determinants of behavior.

1 Introduction

In animals and humans, repeated interactions often lead to mutual cooperation (1–6). However, the psychological basis of cooperative behaviors in humans and animals is still debated: in particular, researchers disagree whether cooperative behaviors are caused by intrinsic preferences for increasing the welfare of others (7). A large body of empirical work has been interpreted as showing that humans do have other-regarding preferences, because cooperation is observed in various experimental games where participants are given the choice between sharing money with other partners, or keeping it for themselves. In many of these experiments, cooperation is observed despite the prediction from standard economic theory that cooperation is not an optimal strategy (8–12). Other researchers (13–15) argue that such results may be explained by the participants not fully understanding the experiment's setup, combined with payoff-based learning during the course of the game.

Traditionally, economists represent the preferences of humans over the possible outcomes of a behavioral interaction using utility functions (16, 17), and assume that humans strive to maximize their utility. It is common to equate the utility with material payoffs (e.g., money in economic experiments), the so-called *homo economicus*, but such utilities frequently fail to predict human behavior, because cooperation in these experiments is not an act that maximizes one's material payoff. In response, economists have proposed that the utility function of an individual may depend not only on the individual's own material payoff, but also on the material payoffs of the other individuals involved in the social interaction. This idea has led to models of preference evolution, where individuals play a given game that has fitness consequences (material payoffs) but where each individual possesses an arbitrary utility function that is genetically determined (18–21). This utility function itself then evolves according to the fitness consequences of the behaviors it generates. Importantly, both the fitness function and the utility function order the outcomes of the social interaction, but these two orderings may be different from each other.

We therefore want to know what the evolutionarily stable utility function is for a given game. In the context of cooperation in particular, can utility functions that assign positive value to others' payoffs, i.e. other-regarding preferences, be evolutionarily stable? The main result from preference evolution models is that if players can observe each other's utility functions before choosing an action, then other-regarding preferences may be evolutionarily stable; otherwise, natural selection leads to an utility function that corresponds exactly to the fitness function (22, 23). This is the same principle that explains the evolution of green-beard genes, where cooperators recognize each other (24). A common way for animals and humans to achieve such recognition is repeated interactions where individuals' responses to each other's behavior is informative of their preferences (20, 25). Interactions between relatives also has been shown to promote other-regarding preferences by interacting with such behavioral responses (25) or recognition of partners (21).

At the same time, most previous theories that model preference evolution or try to explain cooperation in the laboratory do not take into account that the behavior of humans and other animals is modified by learning based on past rewards. Indeed, learning (or initial lack thereof) is usually presented as an alternative to prosocial preferences for explaining behavior in experiments (13). However, as with many social and non-social behaviors consistently produced by a species with a neural system, cooperative behavior must generate positive rewards (in the proximate sense, see below) for an individual (26–29). If cooperation is to be observed in those species, then the temporal sequence of cooperation must be consistent with known principles of learning (30). A theory for the proximate mechanisms of human and animal cooperation is incomplete without accounting for learning at the same time. In a learning context, the question of whether humans have other-regarding preferences thus becomes: can the cooperative act in itself be rewarding?

One may define a reward as an event that generates a particular pattern of activation of neural circuits that induces positive feedback on behavior (26, 29, 31). Essentially, animals tend to repeat actions that are followed by rewards; this phenomenon constitutes the core of instrumental learning. Punishments, on the other hand, are stimuli that generate a negative feedback on behavior, whereby actions followed by punishments tend to be avoided in the future. Certain stimuli act as intrinsic rewards (also called primary, or unconditioned rewards), which allows an animal to build associations between these intrinsic rewards and new actions or stimuli. Glucose is such an intrinsic reward in many animals: an animal can learn to associate glucose with another stimulus (e.g., a particular fruit), or with an action (e.g., in the laboratory, pulling a lever). Once learning has taken place, the associated stimulus (e.g., the fruit), or the associated action (e.g., pulling a lever) become reward predictors (32). Because regions involved in decision-making and social cognition project to the mesolimbic reward system (33), it is possible that the part of the brain responsible for social cognition activates innately the reward system. In other words, cooperation may be intrinsically rewarding in the brain, which may be true not only in humans but also in other primates (34). However, a recent study (that did not observe neural activity) also found that learning in repeated public goods games seems to be driven by material payoffs, and an increase in others' payoffs leads a focal subject to reduce his own contribution to the public good (14). Thus, it is an open empirical question whether rewards other than material payoffs, such as other-regarding preferences, drive social behavior in humans. Moreover, there is little evolutionary theory for how such social preferences should evolve in the context of learning.

From an evolutionary perspective, intrinsic rewards are often viewed as fitness-enhancing stimuli. The idea is that natural selection shapes individuals to innately respond favorably to those stimuli that help increase fitness. This is explicit in many models of the evolution of learning (35–40), where the reinforcement term in the equation describing learning is equated to incremental fitness effects. However, numerous examples show that certain stimuli act as intrinsic rewards without there being a clear relationship between these stimuli and fitness. Certain rewards, when taken in inappropriate amount, are even fitness-detrimental. This is the case for glucose, salt, certain drugs, which can be addictive, and in excessive amount, harmful (41). Hence, the mapping from intrinsic rewards to fitness need not be direct, especially when rewards drive social interactions. Evolutionary models can shed light on how natural selection shapes intrinsic rewards to further individuals' fitness interests in social interactions.

In this article, we present a model of the evolution of such intrinsic rewards when individuals interact in the Prisoner's Dilemma game, where they have the choice between cooperation and defection. To capture general learning processes in humans and animals, we model learning as a basic trial-and-error process where individuals repeat actions followed by rewards and avoid actions followed by punishments. In our model, individuals interact in games whose material payoffs determine fitness. Instead of learning according to the real material payoffs, an individual associates to each game outcome a genetically determined utility, which is used as the intrinsic reward/punishment for that particular outcome. For example, other-regarding individuals may associate positive utilities to outcomes where their partner obtains a positive material payoff, and thus might learn to cooperate as an intrinsically rewarding action. This decoupling of material payoffs and rewards allows us to address the question of how rewards evolve in social interactions. We look for the evolutionarily stable utility functions when individuals interact repeatedly in a game whose material one-shot payoffs determine the 2-person Prisoner's Dilemma game.

2 Model

2.1 Social interactions and rewards

We consider an evolutionary model of repeated pairwise games in a large, well-mixed population of learners with non-overlapping generations. Every generation of the evolutionary process consists of a sequence of interaction rounds, $t = 1, 2, \ldots, T$. At each generation just before $t = 1$, individuals in the population are randomly matched in pairs, and each pair remains together for the entire duration of the game (until $t = T$). Hence, individuals are playing a repeated game with their partner. The one-shot game, played at every time t, is a Prisoner's Dilemma game with two possible actions, cooperate, C (or action 1), or defect, D (or action 2). The one-shot material payoffs for individual i are denoted $\pi_i(C,C) = b - c$, $\pi_i(C,D) = -c$, $\pi_i(D,C) = b$, $\pi_i(D,D) = 0$, where the first element in parentheses of $\pi_i(a_i, a_{-i})$ denotes player i's action (a_i), and the second element denotes his opponent's action (a_{-i}). We assume also that $b > c > 0$. The sequence of material payoffs ultimately determine fitness (see below for details on how fitness is evaluated).

At every interaction round t, each individual in every pair chooses an action. Individual i's action at time t is denoted $a_{i,t}$ and his opponent's action is $a_{-i,t}$. After actions are chosen, both players observe the outcome $(a_{i,t}, a_{-i,t})$ and subjectively evaluate how good the outcome was, which is genetically determined. We call this subjective evaluation the utility function of a player, which may be different than the actual material payoff, $\pi_i(a_{i,t}, a_{-i,t})$, obtained at time t. This utility (rather than the material payoff) determines the reward sensation of a game outcome, and this reward is used by an individual to learn his strategy in the repeated game (see below for details about the learning process). Specifically, the genotype of each individual i associates to each outcome (a_i, a_{-i}) an utility $u_i(a_i, a_{-i})$ that can take any negative or positive real value. We say that the utility is a reward if it is positive $u_i(a_i, a_{-i}) > 0$, while we call it a punishment if $u_i(a_i, a_{-i}) < 0$. Hence a genotype consists of the four utilities $u_i(C,C)$, $u_i(C,D)$, $u_i(D,C)$, $u_i(D,D)$. We can arrange these four utilities in a matrix according to the game outcomes, which we call the utility matrix of individual i (Fig. 1A). However, evolutionarily speaking, it is easier to think of these utilities as the vector $\mathbf{u}_i = (u_i(C,C), u_i(C,D), u_i(D,C), u_i(D,D))$; below we also use the more compact notation $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$, dropping the player's index. The state space is thus $\mathbb{R}^4$. Our interest in this paper is to find the evolutionarily stable utility vector $\mathbf{u}^*$. To do so, we need to know the fitness $f(\mathbf{u}_i)$ of an individual with utility $\mathbf{u}_i$. To arrive there, we first need to specify how the utility vectors of a pair of players determine behavior in the repeated game.

2.2 Learning

We assume that individuals learn to play the game according to a simple trial-and-error procedure. We use a standard model of learning dynamics (30, 40), except that actions are reinforced according to the subjective utilities of a game outcome $u_i(\cdot)$, rather than the objective material payoff $\pi_i(\cdot)$ (see SI text S1). At every time t, an individual i holds in memory action values $V_{i,t}(a_i)$ for both actions $a_i \in \{C, D\}$ and chooses to cooperate at time t with a probability, $p_{i,t}$, that depends on its action values $\{V_{i,t}(C), V_{i,t}(D)\}$. The action values are updated according to the utilities received from the last game outcome (eqs. S1.1–S1.2 in SI text S1).

The behavioral interaction between a reinforcement learner with utilities $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$ and another reinforcement learner with utilities $\mathbf{v} = (v_{11}, v_{12}, v_{21}, v_{22})$ is what we need to analyze in order to compute fitness. By a slight abuse of notation, we denote these two players u and v and their probabilities to cooperate by p_u and p_v respectively. In ref. (40), stochastic approximation theory is used to show that the long-run learning dynamics (eqs. S1.1–S1.2) for a pair of learners can be described as

$$\begin{aligned} \dot{p}_u &= p_u(1-p_u)\xi\left[p_u\{p_v u_{11} + (1-p_v)u_{12}\} - (1-p_u)\{p_v u_{21} + (1-p_v)u_{22}\}\right], \\ \dot{p}_v &= p_v(1-p_v)\xi\left[p_v\{p_u v_{11} + (1-p_u)v_{12}\} - (1-p_v)\{p_u v_{21} + (1-p_u)v_{22}\}\right], \end{aligned} \tag{1}$$

where $\xi > 0$ is an exploration parameter whose role in the original stochastic model is to capture how responsive an individual is to his action values. Eq. 1 displays ten generic behavioral equilibria (Fig. S1). Depending on the specific values of **u** and **v**, one or more of these equilibria may exist. Note that because the original dynamic is stochastic, when the corresponding deterministic system admits several locally stable equilibria, the stochastic dynamics may reach any of these equilibria. It turns out that the theory of stochastic approximations is almost silent about which particular equilibrium will be reached. These lock-in probabilities will however play an important role for the evolutionary stability of the different utility functions we will study below.

Another important fact about the behavioral dynamics is that the stability of the possible behavioral equilibria is very much dependent on the signs of utilities of the individuals involved in an interaction. In particular, one has that a pure equilibrium is locally stable if and only if both players have a positive utility (making it a *reward*) for this outcome. The implication of this is that if at least one player has a negative utility for the outcome, then this outcome is unstable. In other words, if players u and v do not "agree" on preferred outcomes, then a pure behavioral equilibrium cannot be stable. This intuitive result is mathematically true because the eigenvalues of the Jacobian matrix associated to eq. 1 evaluated at a pure outcome (i, j) are simply

$$\lambda_1 = -\xi u_{ij}, \lambda_2 = -\xi v_{ji}. \tag{2}$$

This fact has important evolutionary consequences, as will be detailed below when we analyze interactions between individuals with particular utility functions. In particular, it allows us to classify different utility functions by their sign for each of the four outcomes.

2.3 Fecundity

Assuming that interactions last long enough ($T \to \infty$), we define the fecundity of individual i as being proportional to the average material payoff obtained at equilibrium of the learning process, i.e.

$$f_i = f(u_i) = \sum_{\mathbf{a} \in \mathcal{A}} \hat{\mathbf{p}}(\mathbf{a}) \pi_i(\mathbf{a}), \tag{3}$$

where $\hat{\mathbf{p}}(\mathbf{a}) = \hat{p}_i(a_i)\hat{p}_{-i}(a_{-i})$ is the equilibrium probability of outcome $\mathbf{a} = (a_i, a_{-i})$. The sum in eq. 3 is taken over the set of possible game outcomes, $\mathcal{A} = \{(C,C),(C,D),(D,C),(D,D)\}$. We call $\hat{\mathbf{p}}$ the behavioral equilibrium. Importantly, while the utility function does not appear on the right-hand side of eq. 3, we still defined it as $f(u_i)$ because the equilibrium choice probabilities of a player, $\hat{p}_i(a_i)$, implicitly depend on the utility function of player i, as will become clearer when we derive expressions for the behavioral equilibria below.

The fecundity $f(\mathbf{u})$ depends on the outcome of the learning dynamics, and is therefore not continuous in $\mathbf{u}$, which renders difficult a full analytic treatment of the model. To overcome this problem we adopt two complementary approaches. First, we focus on a smaller number of utility functions that are relevant to our original question of the evolution of other-regarding preferences. Second we run evolutionary simulations of the full model to have a more comprehensive view of our model.

3 Results

3.1 Analytical results for 4-strategy competition

We first consider the evolutionary dynamics of four possible utility functions that are represented in Fig. 1A using the replicator dynamics (for details of the analysis, see S2, S3, and S4):

- The **Realistic** function, which associates to outcomes an utility of the same sign than the real material payoff. This type of utility function is the "default" utility function, used in virtually all models of the evolution of learning (35–40). It takes as a special case the material payoff function, i.e., $u_i = \pi_i$. It is the function that evolves when interactions

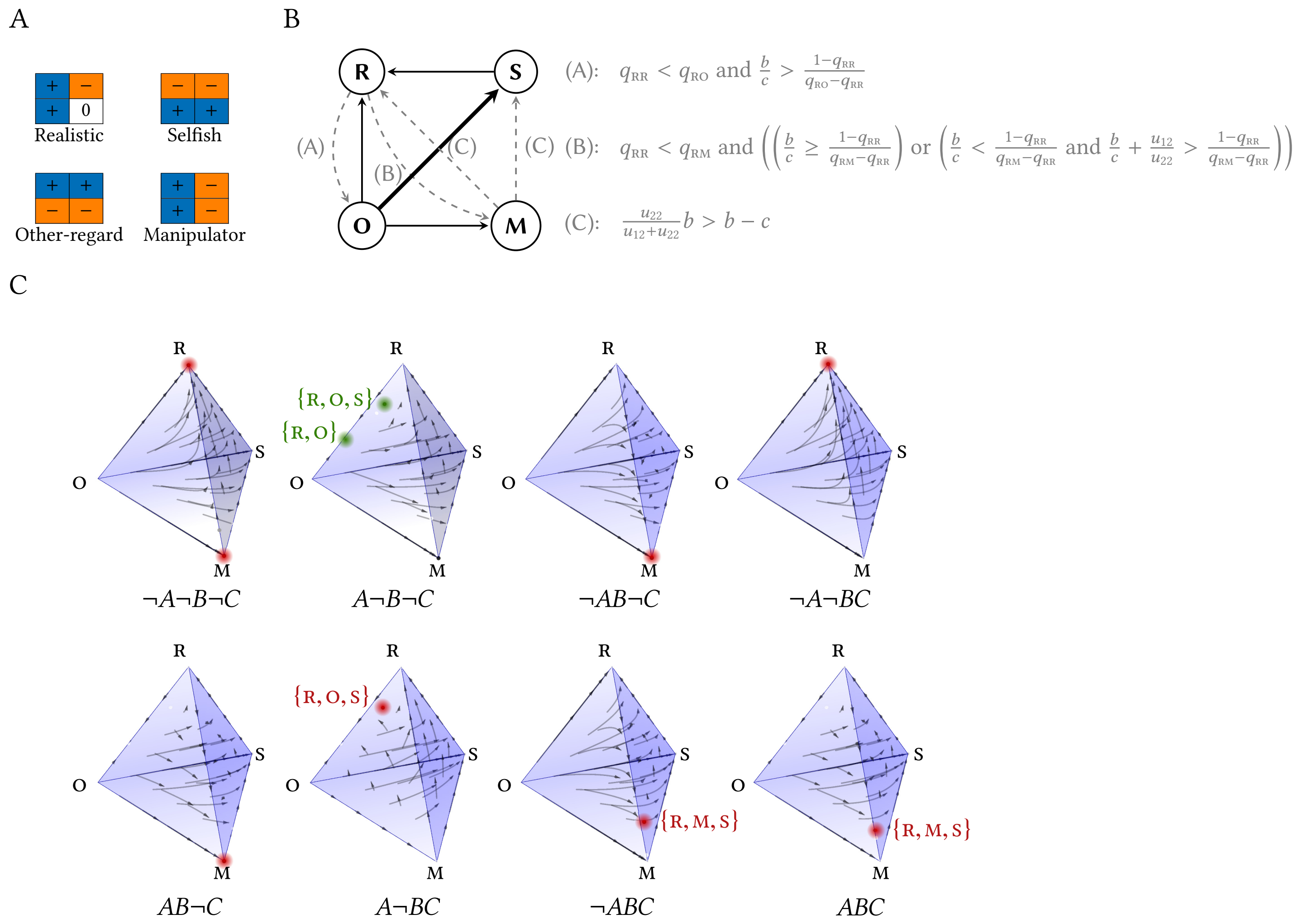

A
+ −
+ 0
Realistic
− −
+ +
Selfish
+ +
− −
Other-regard
+ −
+ −
Manipulator
B
R
S
O
M
(A)
(B)
(C)
(C)
(A): $q_{\text{RR}} < q_{\text{RO}}$ and $\frac{b}{c} > \frac{1-q_{\text{RR}}}{q_{\text{RO}}-q_{\text{RR}}}$
(B): $q_{\text{RR}} < q_{\text{RM}}$ and $\left(\left(\frac{b}{c} \geq \frac{1-q_{\text{RR}}}{q_{\text{RM}}-q_{\text{RR}}}\right) \text{ or } \left(\frac{b}{c} < \frac{1-q_{\text{RR}}}{q_{\text{RM}}-q_{\text{RR}}} \text{ and } \frac{b}{c} + \frac{u_{12}}{u_{22}} > \frac{1-q_{\text{RR}}}{q_{\text{RM}}-q_{\text{RR}}}\right)\right)$
(C): $\frac{u_{22}}{u_{12}+u_{22}} b > b - c$
C
R
S
O
M
$\neg A \neg B \neg C$
{R, O, S}
{R, O}
$A \neg B \neg C$
$\neg A B \neg C$
$\neg A \neg B C$
$A B \neg C$
{R, O, S}
$A \neg B C$
{R, M, S}
$\neg A B C$
{R, M, S}
ABC

Figure 1: Replicator dynamics for the competition between Realistic, Other-regard, Manipulator, and Selfish. (A) The 4 strategies considered in the analytical model. A strategy is defined by the outcomes to which it associates a positive or negative utility. The first row/column corresponds to Cooperate and the second row/column to Defect. Utilities are to row-player. (B) Pairwise invasion diagram between the four strategies Realistic, Other-regard, Manipulator, and Selfish, and associated invasion conditions. A plain directed edge from node X to node Y means that strategy Y always invades a monomorphic population of X (but does not necessarily reach fixation). A dashed directed edge from node X to node Y means that Y can invade X under certain conditions (A, B, and C) on the model parameters. When a given strategy can be invaded by more than one other strategy, a thick edge designates the best response. Note that all combinations of these three conditions are possible. (C) Classification of phase portraits for the replicator dynamics in the 4-strategy game defined by the competition between Realistic, Other-regard, Manipulator, and Selfish. Each subfigure is a drawing of the 4-simplex. At each vertex, one of the four strategies is at frequency 1: Realistic at the top, Manipulator at the bottom-front, Other-regard at the back left, and Selfish at the back right. The letters A, B, and C refer to the conditions in Panel B, where the symbol "¬" denotes logical negation. For instance, the subfigure labeled $\neg ABC$ is drawn for parameter values such that condition A is not true, but conditions B and C are true. Red dots denote locally stable equilibria, i.e. possible outcomes of natural selection. To disambiguate the 3D view, red labels in curly braces indicate the set of strategies present at an equilibrium. In the subfigure for the case $A \neg B \neg C$, the green dots are two alternative outcomes: {R, O} occurs when Selfish cannot invade this polymorphism; if Selfish does invade this polymorphism, {R, O} becomes unstable and {R, O, S} stable. The condition for this to happen is given by eq. S4.7 in Appendix S4.

between players are completely anonymous, one-shot, and there is no assortment in the matching process (22, 23).

- The **Other-regard** function, which associates positive utility to the outcomes where the opponent obtains a strictly positive payoff. In other words, this strategy associates positive utilities only to the outcomes where it cooperates.
- The **Selfish** function, which associates positive utility to the outcomes where it defects.
- The **Manipulator** function, which associates positive utility only to the outcomes where its opponent cooperate. The name of this utility function stems from the fact that it will drive a compliant opponent (who associates positive utility to all outcomes) to cooperate.

We first construct the fitness matrix for the evolutionary game in Table 1 by considering the stable equilibria of learning dynamics of all possible pairwise matchings between the four strategies (described in detail in the SI text S2 and S3). For the four strategies we consider in this section, no more than two behavioral equilibria are locally stable at the same time. It turns out that in all cases where two equilibria are locally stable, one of them is mutual cooperation, $(1, 1)$. Because the underlying learning model is stochastic (eqs. S1.1–S1.2 in SI text S1), the lock-in probability in the cooperative equilibrium $(1, 1)$ will affect the fitness and hence the evolutionary competition between the four strategies we are considering. However, there is no general technique to obtain an expression of the lock-in probability. At this point, we leave these probabilities unspecified, and denote by q_{uv} the probability that an interaction between strategy u and strategy v leads to the cooperative equilibrium $(1, 1)$. For instance, two Realistic individuals can learn mutual cooperation, $(1, 1)$, or mutual defection, $(0, 0)$. The probability that an interaction between two Realistic individuals leads to mutual cooperation is thus denoted q_{RR}; the probability of locking in the defective equilibrium is $1 - q_{\mathrm{RR}}$.

3.1.1 Evolutionary dynamics for the Prisoner's Dilemma

We use the replicator dynamics (42, eq. S4.2 in SI text) to describe the competition between Realistic, Other-regard, Manipulator, and Selfish, with the evolutionary game given in Table 1. Determining the outcome of the replicator dynamics is dependent on the parameters b (benefit to a receiver of a cooperative act), c (cost of cooperating), and the lock-in probabilities in the cooperative equilibrium q_{RR}, q_{RO}, q_{RM}, q_{OM}, for the different behavioral interactions where several equilibria are locally stable.

We first find that, although (Selfish, Selfish) is always a weak Nash Equilibrium (NE) of the evolutionary game between the four strategies, regardless of parameter values, it is never evolutionarily stable (Fig. 1 and Table 1). This is because Selfish gets invaded by Realistic, which learns to defect against Selfish, but cooperates with itself. On the other hand, the strategy Other-regard is also always invaded by every other strategy in pairwise competitions, although it can be part of a mixed equilibrium, as we will see below. All other important results depend on the parameters of the model, and three basic conditions on the parameters help classify the possible evolutionary outcomes (conditions A, B, and C in Fig. 1A). For certain parameter values, Realistic can be an evolutionarily stable strategy when the benefit-to-cost ratio is sufficiently low (conditions A and B in Fig. 1A). Also, for other parameter values, Manipulator can be evolutionary stable (condition C in Fig. 1A). Note that these conditions are not mutually exclusive, so both Realistic and Manipulator can be evolutionarily stable at the same time (Fig. 1B). When at least one of Realistic or Manipulator is not evolutionarily stable, then we obtain polymorphic equilibria. In such polymorphic equilibria, we have either three strategies (there is an equilibrium with Realistic, Other-regard, Selfish, and an equilibrium with Realistic, Manipulator, and Selfish) or two strategies (Realistic and Other-regard), the common feature of these being that Realistic is always present (Fig. 1B). We note here that Other-regard can only be present when Realistic is present. Moreover, according to condition A in Fig. 1A, Realistic should cooperate more often with Other-regard than with itself for the latter to make part of an equilibrium.

Table 1: Evolutionary fitness matrix amongst the 4 strategies considered in the analytical model.

	R	O	M	S
R	$q_{\mathrm{RR}}(b-c)$	$q_{\mathrm{RO}}(b-c)+(1-q_{\mathrm{RO}})b$	$q_{\mathrm{RM}}(b-c)+(1-q_{\mathrm{RM}})\left(b\left(\frac{u_{22}}{u_{12}+u_{22}}\right)\right)$	0
O	$q_{\mathrm{RO}}(b-c)+(1-q_{\mathrm{RO}})(-c)$	$b-c$	$q_{\mathrm{OM}}(b-c)+(1-q_{\mathrm{OM}})(-c)$	$-c$
M	$q_{\mathrm{RM}}(b-c)+(1-q_{\mathrm{RM}})\left((-c)\left(\frac{u_{22}}{u_{12}+u_{22}}\right)\right)$	$q_{\mathrm{OM}}(b-c)+(1-q_{\mathrm{OM}})b$	$b-c$	$-c\frac{u_{22}}{u_{12}+u_{22}}$
S	0	b	$b\frac{u_{22}}{u_{12}+u_{22}}$	0

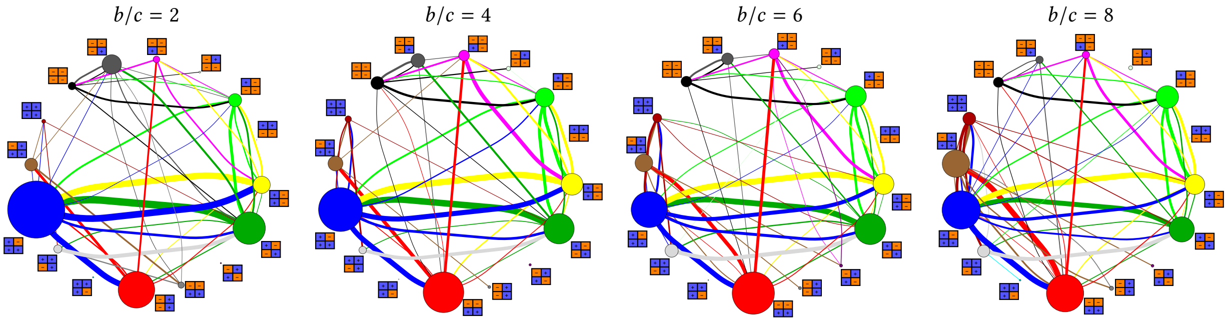
3.2 Simulations

3.2.1 Freely evolving utilities

In order to explore the entire set of possible strategies, we also performed individual-based simulations for various values of the benefit-to-cost ratio, b/c. In such simulations, the lock-in probabilities in behavioral equilibria, which played a critical role in determining the evolutionary outcome in the above 4-strategy model, will no longer be parameters but will have a value that depends on the utilities of the particular strategies involved in behavioral interactions. Our evolutionary simulations consist of the trait substitution sequence of adaptive dynamics. Namely, we assume that the genotype of an individual, $\mathbf{u} = (u_{11}, u_{12}, u_{21}, u_{22})$, is supported by one locus, and that the population is always monomorphic. At each iteration, we propose a mutation and determine whether the mutant invades the resident population using eqs. 41–42 of ref. (43), which is calculated for Wright's island model (in our case, the population is panmictic, or there is only one deme).

In order to represent the four utilities at the same time, we classified all strategies according to the sign of their utilities (as we demonstrated above, these signs provide necessary conditions on the possible behavioral equilibria), which results in $2^4 = 16$ classes of strategies (because each of the four utilities has two possible signs). We can first look at the proportion of time a simulation run spends in each of the 16 strategy classes, which is an approximation of the stationary distribution of the evolutionary dynamics. We find that 6 strategies are consistently represented more than 10% of the time in the occupation measure: Selfish, Avoid Sucker's Payoff, Manipulator, Matcher, Pareto, Anti-Cooperation. Avoid Sucker's Payoff (AS) is similar to Realistic except that it has a positive utility for mutual defection instead of a 0; AS produces the same behavioral equilibria as Realistic when paired with other strategies (Fig. S3). Matcher has positive utilities only for outcomes where its own action matches that of its opponent. Pareto has positive utility only for the outcome of mutual cooperation. Finally, Anti-Cooperation is the exact opposite of Pareto, as it has positive utilities for all outcomes except for mutual cooperation. In Fig. 2, we show results for various benefit-to-cost ratios, b/c, which leads to two main observations.

The first take-away is that our simulations confirm the overall pattern in the analysis of the



$u_{11} > 0$ 41%	$u_{12} > 0$ 84%
$u_{21} > 0$ 39%	$u_{22} > 0$ 22%

$u_{11} > 0$ 44%	$u_{12} > 0$ 83%
$u_{21} > 0$ 39%	$u_{22} > 0$ 31%

$u_{11} > 0$ 44%	$u_{12} > 0$ 79%
$u_{21} > 0$ 40%	$u_{22} > 0$ 30%

$u_{11} > 0$ 43%	$u_{12} > 0$ 72%
$u_{21} > 0$ 37%	$u_{22} > 0$ 29%

$\bar{u}_{11} = 1.28$	$\bar{u}_{12} = -4.76$
$\bar{u}_{21} = 1.47$	$\bar{u}_{22} = 3.81$

$\bar{u}_{11} = 0.58$	$\bar{u}_{12} = -4.73$
$\bar{u}_{21} = 1.50$	$\bar{u}_{22} = 2.59$

$\bar{u}_{11} = 0.72$	$\bar{u}_{12} = -4.06$
$\bar{u}_{21} = 1.44$	$\bar{u}_{22} = 2.96$

$\bar{u}_{11} = 0.88$	$\bar{u}_{12} = -2.96$
$\bar{u}_{21} = 1.96$	$\bar{u}_{22} = 3.17$

Figure 2: Simulation results for various benefit-to-cost ratios (b/c). The top row shows the invasion graph between the 16 classes of strategies defined by their utilities' sign (see main text), the second row shows the proportion of time each utility was positive in a simulation run, the third row shows the time average of utilities. In the invasion graph, the size of the nodes is proportional to the amount of time a simulation run spends in the corresponding strategy class. The edges are colored according to the invader strategy and thus indicate the direction of the edges. Edge thickness is proportional to the number of invasions that occured between a pair of strategies (and we do not show edges between pairs of strategies for which the number of invasions was less than 10). Simulation parameter values: $\min u = -10$; $\max u = 10$; $\xi = 2$; $c = 1$; $T = 150$.

replicator dynamics, where at low values of b/c the strategy AS (corresponding to the Realistic strategy in the analytical model) experiences few invasions. As b/c increases, more strategies are able to invade AS, and consequently the frequency of AS declines (Fig. 2 and Fig. 3). In particular, if we analyze the invasions between the 6 dominant strategies in our simulations (Fig. S4), we find that Manipulator, Matcher, and Pareto invade AS only for sufficiently high b/c. All these strategies have a positive utility for mutual cooperation; they also have a negative utility for the sucker's outcome ($u_{12} < 0$). The success of AS and of cooperative strategies more generally yield an average utility matrix of the AS type (Fig. 2), where average utilities are ordered as $\bar{u}(D,D) > \bar{u}(D,C) > \bar{u}(C,C) > \bar{u}(C,D)$, which is different than the ordering of the material payoffs, $\pi(D,C) > \pi(C,C) > \pi(D,D) > \pi(C,D)$. The strategy Pareto increases in frequency in the stationary distribution for increasing b/c (Fig. 3A), as the analysis shows that it invades AS for high enough b/c (Fig. S4). Strategies that are able to invade AS (Manipulator, Matcher, Pareto) can mutually invade one another and we indeed observe that an important number of invasions occur between AS, Manipulator, Matcher, Pareto (Fig. 2). As a consequence of the increasing success of strategies that positively value cooperation as a function of b/c, we observe that the overall cooperation frequency in the population increases for increasing b/c (Fig. 3B). Even though previous work has shown that cooperative strategies in the iterated Prisoner's Dilemma can be evolutionary robust (6), we could not expect this for the particular type of learning strategies that we have decided to study.

Contrasting the apparent success of conditionally cooperative strategies, a second major feature of our simulations is the success of Selfish. For all b/c, the simulation spends approximately 15-20% of the time in this strategy class, and for high b/c this makes Selfish the most represented strategy class in the stationary distribution (because of the decline of AS; Fig. 2 and Fig. 3A). Although this result could not be anticipated from our analysis of the replicator dynamics above, it is still consistent with the fact that Selfish was relatively stable (only Realistic could invade it). Analytically considering the invasion conditions between the 6 dominant strategy classes in the simulations (Fig. S4) reveals that Selfish is also relatively stable in this set, with only AS and Matcher being able to invade it. In our simulations, AS invades more frequently Selfish that Matcher does because, with our mutation scheme, an AS mutant is much more likely than a Matcher mutant to occur in a Selfish population given that we draw mutations from a doubly

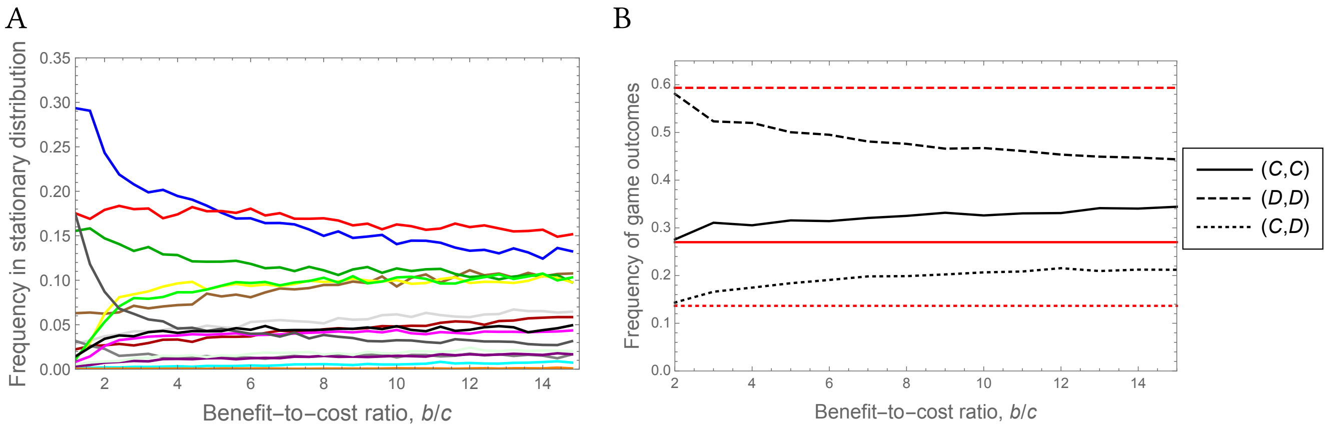


Figure 3: Effect of the benefit-to-cost ratio, b/c. (A) Proportion of time the simulation spends in each of the 16 strategy classes as a function of b/c, which is a measure of the stationary distribution. Colors of strategies are as in Fig. 2 (see also Fig. S5). (B) Time average of the frequency of game outcomes in a simulation run as a function of b/c (black lines); plain: (C,C); dashed: (D,D). The frequency of the third type of possible of outcome, (C,D), is the complementary ((C,D) and (D,C) are the same outcome). The red lines show the expected frequency of the corresponding game outcomes in the population if we draw strategies randomly from a uniform distribution. Parameter values are as in Fig. 2.

exponential distribution centered at the resident phenotype.

A final observation regarding strategy classes is that Anti-Cooperation is relatively successful for high b/c (Fig. 3A). Our invasion analysis in Fig. S4 shows that this strategy, despite having a positive utility for the sucker's outcome, compensates by exploiting certain cooperative strategies, such as Pareto. At low b/c, Anti-Cooperation gets exploited by strategies that have positive utilities for defection (such as AS or Selfish; Fig. S3) but as b/c increases, Anti-Cooperation becomes more stable against these strategies, which explains why it makes part of an important proportion of the stationary distribution of the evolutionary dynamics.

3.2.2 Utilities explicitly depending on material payoffs

In this section we perform additional simulations by constraining the utility function to be dependent on the material payoffs of the focal player and its opponent. This allows us to address more directly the question of whether (and, if any, what type of) other-regarding preferences evolve in our model. Specifically, for any game outcome $\mathbf{a} = (a_i, a_{-i})$, we consider utility functions of the form

$$u_i(\mathbf{a}) = \pi_i(\mathbf{a}) + \beta(\pi_i(\mathbf{a}) + c + k)(\pi_{-i}(\mathbf{a}) + c + k) + \alpha\pi_{-i}(\mathbf{a}) - \gamma|\pi_i(\mathbf{a}) - \pi_{-i}(\mathbf{a})|. \qquad (4)$$

where (α, β, γ) are player i's genetically determined parameters. In this section we will be interested in the evolution of these three parameters. In eq. 4, c is the negative of the sucker's payoff $(-c)$ and is added to the realized payoff to ensure that the term multiplied by β is always positive. The parameter k is here to allow the utility to vary as a function of β. Our utility function then measures the extent to which an individual is "additively" other-regarding ($\alpha \in [-1, 1]$), the extent to which he is "multiplicatively" other-regarding ($\beta \in [-1, 1]$), and inequity aversion ($\gamma \in [-1, 1]$). Even though this utility function can realize all of the 16 possible utility matrices discussed above, the structure of the phenotype space changes as compared to the above simulations where we let the utility matrix evolve in an unconstrained way (Fig. S5).

Our simulations with the utility function in eq. 4 show that the selection pressure on other-regarding preferences increases with b/c (Fig. 4A and Fig. S8). The average value of β is close to 0 for low enough b/c, but suddenly increases at a threshold value of b/c. For these higher b/c values, the average β is approximately 0.5, indicating the evolution of multiplicative other-regard. The average values of α and γ are negative for low b/c, indicating respectively a combination of competitive preferences (valuing negatively other's success) and inequity aversion. Both α and γ decrease in magnitude as b/c increases, but remain negative. This is a consequence of the fact that the selection pressure on α and γ decreases with increasing b/c, because the absolute difference between the temptation to defect, b and the sucker's payoff, $-c$, decreases. This pattern is accompanied by a general increase in the utility for mutual cooperation as a function of b/c (Fig. S7A in the SI). For high b/c, mutual cooperation becomes the preferred outcome of the evolutionarily stable utility function and mutual defection the least preferred outcome. In agreement with the above simulations for freely evolving utilities, AS is the dominant utility matrix for low b/c. The "Compliant" utility matrix (with all four utilities positive) becomes the dominant one for high b/c (Fig. S7B in the SI).

Even though other-regarding preferences evolve for sufficiently high b/c, this is not accompanied by the evolution of increased effective mutual cooperation, even though the frequency of mutual defection decreases. However, this decrease is due to an increase in the asymmetric (C, D)

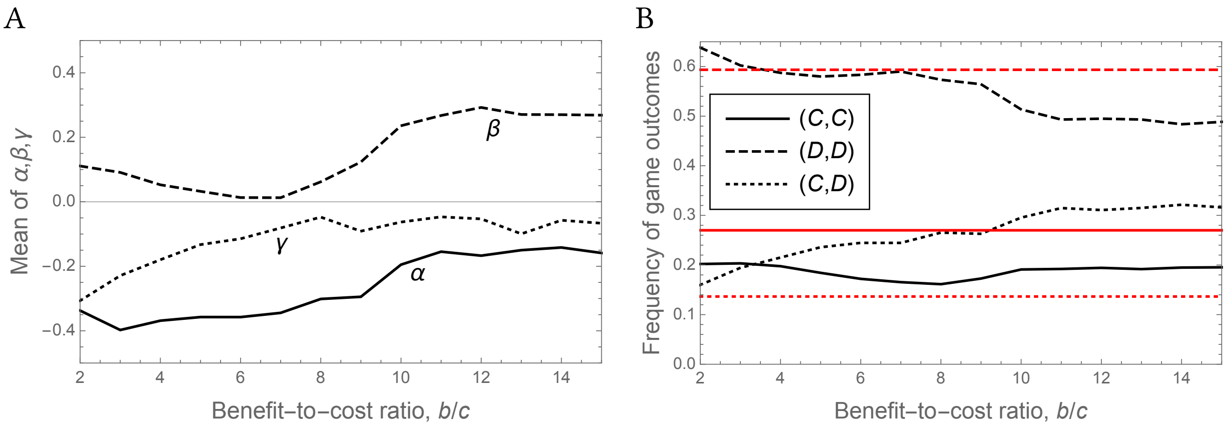


Figure 4: Results for the model where the utility depends explictly on material payoffs (eq. 4 with $k = 2$) as a function of the benefit-to-cost ratio, b/c. (A) Time average of α, β, γ in a simulation run. (B) Time average of the frequency of game outcomes in a simulation run (similar to Fig. 3).

outcome (Fig. 4B). Overall cooperation is thus increasing but individuals do not coordinate on cooperating at the same time. This can be explained by the fact that the Compliant utility matrix that evolves for high b/c can learn any outcome (all pure equilibria are stable when all utilities are positive). However, the fact that mutual defection is the least preferred outcome implies that the probability to learn this equilibrium will be the lowest of the four outcomes.

4 Discussion

We presented a model of how intrinsic rewards that drive learning in social interactions evolve. Rewards capture the intrinsic preferences of individuals over states of the world, and constitute the fundamental building block of reinforcement learning. Because all behaviors are in part influenced by learning, modeling the evolution of social behaviors in animals requires that we take into account how behavior is generated through learning within an individual's lifespan. Within this framework, we were able to address the question of whether other-regarding preferences support the evolution of cooperation, under the constraint of reinforcement learning. Hence, our work bridges a gap between models of the evolution of preferences in games (20–23, 25) and neuro-behavioral studies of social interactions in the laboratory (33, 44, 45). Overall, our results indicate that multiple preference functions can be evolutionarily stable when individuals interact repeatedly in the Prisoner's Dilemma. In particular, we find that evolutionarily successful preferences are of two general types: (1) those that have a positive utility for mutual cooperation but a negative utility for being exploited; (2) selfish preferences that associate positive utilities to outcomes where their carriers defect, and have negative utility for cooperation. This is true in both our analytical results in replicator dynamics and the numerical simulations in the whole strategy set. Further simulations show that other-regarding preferences evolve for suffficiently high benefit-to-cost ratio.

Our results bring together findings from economics, psychology, and basic brain physiology in a unifying evolutionary framework to account theoretically for the apparent other-regarding behavior of humans and other species. Economic theory relies on the concept of utility to capture behavior, but the utility function of an individual is by definition an internal construct that is difficult to access (44). In the context of learning, utility can be equated to reward, because rewards are at the core of repeated behaviors (29). Empirically, one way to try to access the utility or reward function is to observe the pattern of activation in the brain when individuals make decisions. Neuro-behavioral studies of social decision-making provide empirical support for our finding that positive preferences for cooperation are evolutionarily prevalent. These studies reveal that cooperation can generate rewards in the human brain, which is consistent with the positive utility of winning strategies for mutual cooperation found in our model (33, 44, 45).

While our model shows that evolution can lead to intrinsically rewarding mutual cooperation, such utilities do not necessarily correspond to pure other-regarding preferences. For low benefit-to-cost ratios, competitive preferences that value other’s payoff negatively tend to evolve. In contrast, for sufficiently high benefit-to-cost ratio we see the evolution of conditional (multiplicative) other-regarding preferences, in agreement with previous results that found these preferences to be evolutionarily stable in continuous social dilemmas (20). On the other hand, one can interpret our results for the freely evolving utilities as reflecting the evolution of the correct representation of real fitness effect of mutual cooperation, because mutual cooperation generates a positive effect on fitness. However, the “Realistic” utility function is not the only evolutionarily successful one in our model. For example, some evolutionarily successful preference functions value positively both mutual cooperation and mutual defection. These signs, together with a negative utility for the sucker’s outcome guarantee uninvadability by the Selfish preference function, because individuals with such preferences will learn to defect against Selfish. Therefore, our results do not show that natural selection leads to the correct representation of fitness effects in the brain in the context of learning. Another important distinction is that, even though the utility for the temptation to defect is the highest in the model with freely evolving utilities, this does not necessarily mean that there is no other-regard preferences. This fact is illustrated by our results for the constrained utility function, which capture cases where other-regard (e.g. a positive β) can evolve even if the values of other evolutionary parameters make the temptation outcome being more rewarding than mutual cooperation.

Taken together, these results suggest that learning agents can be selected to have some other-regarding utilities, but these are unlikely to be "pure" other-regard. Rather, any evolved rewards from cooperation is predicted to be conditional on both parties cooperating. This lends partial theoretical support to empirical studies of cooperation that indicate that humans and animals have prosocial preferences but are also averse to inequity (46, 47).

Our finding that a diversity of utility functions are favored by natural selection is also consistent with empirical findings that humans in behavioral experiments tend to act according to distinct behavioral types. In particular, strategies that value cooperation positively can produce similar behavior to that of reciprocating strategies (repeating the action of the partner in the previous round), and Selfish can produce the behavior of non-cooperators; these two behavioral types have recently been found to represent the action sequence of many human participants in laboratory experiments (15, 48) and have been considered as plausible evolutionarily significant behavioral rules in theoretical models (1, 2, 4, 6). Moreover, in addition to a diversity of preference types, our model also shows the potential for multiple behavioral outcomes in a population monomorphic for a given preference function. This is because of the fact that stochastic learning processes can converge to different equilibrium profiles, which provides another way for the behavioral variation observed in learning experiments (49).

In conclusion, our model articulates four levels of determinants of behavior: (1) the biological rewards at the core of brain functioning; (2) the psychological preferences that determine which states of the world are rewarding; (3) the social interactions that affect changes in the states of the world; (4) the biological process of natural selection determining which behavioral mechanisms prevail in an evolving population. We find that evolution of rewards for learning captures both the possibility of cooperation and a diversity of individual preferences that can be evolutionarily successful. These results show the promise of integrating learning based on evolving intrinsic rewards from social interactions as a proximate mechanism for understanding the nature of cooperation in humans and animals.

References

[1] R. L. Trivers, “The evolution of reciprocal altruism,” *The Quarterly Review of Biology*, vol. 46, no. 1, pp. 35–57, 1971.

[2] R. Axelrod and W. D. Hamilton, “The evolution of cooperation,” *Science*, vol. 211, pp. 1390–1396, 1981.

[3] G. S. Wilkinson, “Reciprocal altruism in bats and other mammals,” *Ethology and Sociobiology*, vol. 9, no. 2, pp. 85–100, 1988.

[4] L. Lehmann and L. Keller, “The evolution of cooperation and altruism – a general framework and a classification of models,” *Journal of Evolutionary Biology*, vol. 19, no. 5, pp. 1365–1376, 2006.

[5] K. Schneeberger, M. Dietz, and M. Taborsky, “Reciprocal cooperation between unrelated rats depends on cost to donor and benefit to recipient,” *BMC Evolutionary Biology*, vol. 12, no. 1, p. 41, 2012.

[6] A. J. Stewart and J. B. Plotkin, “From extortion to generosity, evolution in the Iterated Prisoner’s Dilemma,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 38, pp. 15348–15353, 2013.

[7] R. Kurzban, M. N. Burton-Chellew, and S. A. West, “The evolution of altruism in humans,” *Annual Review of Psychology*, vol. 66, no. 1, pp. 575–599, 2015.

[8] E. Fehr and S. Gächter, “Cooperation and Punishment in Public Goods Experiments,” *American Economic Review*, vol. 90, no. 4, pp. 980–994, 2000.

[9] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath, "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies," *American Economic Review*, vol. 91, no. 2, pp. 73–78, 2001.

[10] C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction.* Princeton, NJ: Princeton University Press, 2003.

[11] E. Fehr and U. Fischbacher, "The nature of human altruism," *Nature*, vol. 425, pp. 785–791, 2003.

[12] A. Chaudhuri, "Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature," *Experimental Economics*, vol. 14, no. 1, pp. 47–83, 2010.

[13] K. Binmore, "Economic man – or straw man?," *Behavioral and Brain Sciences*, vol. 28, no. 06, pp. 817–818, 2005.

[14] M. N. Burton-Chellew, H. H. Nax, and S. A. West, "Payoff-based learning explains the decline in cooperation in public goods games," *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 282, no. 1801, p. 20142678, 2015.

[15] M. N. Burton-Chellew, C. E. Mouden, and S. A. West, "Conditional cooperation and confusion in public-goods experiments," *Proceedings of the National Academy of Sciences*, p. 201509740, 2016.

[16] M. J. Osborne and A. Rubinstein, *A Course in Game Theory.* MIT Press, 1994.

[17] H. Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences.* Princeton University Press, 2009.

[18] P. Ockenfels, "Cooperation in prisoners' dilemma," *European Journal of Political Economy*, vol. 9, pp. 567–579, Nov. 1993.

[19] W. Güth, "An evolutionary approach to explaining cooperative behavior by reciprocal incentives," *International Journal of Game Theory*, vol. 24, no. 4, pp. 323–344, 1995.

[20] E. Akçay, J. Van Cleve, M. W. Feldman, and J. Roughgarden, "A theory for the evolution of

other-regard integrating proximate and ultimate perspectives," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 19061–19066, 2009.

[21] I. Alger and J. W. Weibull, "Homo moralis—preference evolution under incomplete information and assortative matching," *Econometrica*, vol. 81, no. 6, pp. 2269–2302, 2013.

[22] E. A. Ok and F. Vega-Redondo, "On the evolution of individualistic preferences: An incomplete information scenario," *Journal of Economic Theory*, vol. 97, no. 2, pp. 231–254, 2001.

[23] E. Dekel, J. C. Ely, and O. Yilankaya, "Evolution of preferences," *The Review of Economic Studies*, vol. 74, no. 3, pp. 685–704, 2007.

[24] A. J. Robson, "Efficiency in evolutionary games: Darwin, Nash and the secret handshake," *Journal of Theoretical Biology*, vol. 144, no. 3, pp. 379–396, 1990.

[25] E. Akçay and J. Van Cleve, "Behavioral responses in structured populations pave the way to group optimality," *American Naturalist*, vol. 179, no. 2, pp. 257–69, 2012.

[26] J. M. Pearce, *Animal Learning and Cognition: An Introduction*. Hove ; New York: Psychology Press, 3rd edition ed., 2008.

[27] S. J. Shettleworth, *Cognition, Evolution, and Behavior*. New York, NY: Oxford University Press, 2009.

[28] L. A. Dugatkin, *Principles of Animal Behavior*. New York, NY: WW Norton & Co, 2nd ed., 2010.

[29] W. Schultz, "Neuronal reward and decision signals: From theories to data," *Physiological Reviews*, vol. 95, no. 3, pp. 853–951, 2015.

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

[31] A. Dickinson and B. Balleine, "Motivational control of goal-directed action," *Animal Learning & Behavior*, vol. 22, no. 1, pp. 1–18, 1994.

[32] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139–154, 2009.

[33] C. H. Declerck, C. Boone, and G. Emonds, "When do people cooperate? The neuroeconomics of prosocial decision making," *Brain and Cognition*, vol. 81, no. 1, pp. 95–117, 2013.

[34] S. W. C. Chang, N. A. Fagan, K. Toda, A. V. Utevsky, J. M. Pearson, and M. L. Platt, "Neural mechanisms of social decision-making in the primate amygdala," *Proceedings of the National Academy of Sciences*, p. 201514761, 2015.

[35] R. Boyd and P. J. Richerson, *Culture and the Evolutionary Process.* Chicago, IL: University of Chicago Press, 1988.

[36] J. Josephson, "A numerical analysis of the evolutionary stability of learning rules," *Journal of Economic Dynamics and Control*, vol. 32, no. 5, pp. 1569–1599, 2008.

[37] S. Hamblin and L.-A. Giraldeau, "Finding the evolutionarily stable learning rule for frequency-dependent foraging," *Animal Behaviour*, vol. 78, no. 6, pp. 1343–1350, 2009.

[38] M. Arbilly, U. Motro, M. W. Feldman, and A. Lotem, "Co-evolution of learning complexity and social foraging strategies," *Journal of Theoretical Biology*, vol. 267, no. 4, pp. 573–581, 2010.

[39] E. Katsnelson, U. Motro, M. W. Feldman, and A. Lotem, "Evolution of learned strategy choice in a frequency-dependent game," *Proceedings of the Royal Society B: Biological Sciences*, p. rspb20111734, 2011.

[40] S. Dridi and L. Lehmann, "On learning dynamics underlying the evolution of learning rules," *Theoretical Population Biology*, vol. 91, pp. 20–36, 2014.

[41] T. Saah, "The evolutionary origins and significance of drug addiction," *Harm Reduction Journal*, vol. 2, p. 8, June 2005.

[42] P. D. Taylor and L. B. Jonker, "Evolutionary stable strategies and game dynamics," *Mathematical Biosciences*, vol. 40, no. 1–2, pp. 145–156, 1978.

[43] J. Van Cleve, "Social evolution and genetic interactions in the short and long term," *Theoretical Population Biology*, vol. 103, pp. 2–26, 2015.

[44] E. Fehr and C. F. Camerer, "Social neuroeconomics: The neural circuitry of social preferences," *Trends in Cognitive Sciences*, vol. 11, no. 10, pp. 419–427, 2007.

[45] C. C. Ruff and E. Fehr, "The neurobiology of rewards and values in social decision making," *Nature Reviews Neuroscience*, vol. 15, no. 8, pp. 549–562, 2014.

[46] E. Fehr and K. M. Schmidt, "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, vol. 114, no. 3, pp. 817–868, 1999.

[47] S. F. Brosnan and F. B. M. de Waal, "Evolution of responses to (un)fairness," *Science*, vol. 346, no. 6207, p. 1251776, 2014.

[48] U. Fischbacher, S. Gächter, and E. Fehr, "Are people conditionally cooperative? Evidence from a public goods experiment," *Economics Letters*, vol. 71, no. 3, pp. 397–404, 2001.

[49] T. Chmura, S. J. Goerg, and R. Selten, "Learning in experimental 2 × 2 games games," *Games and Economic Behavior*, vol. 76, no. 1, pp. 44–73, 2012.

[50] S. Dridi and L. Lehmann, "A model for the evolution of reinforcement learning in fluctuating games," *Animal Behaviour*, vol. 104, pp. 87–114, 2015.

[51] M. Benaïm, "Dynamics of stochastic approximation algorithms," in *Séminaire de Probabilités XXXIII*, vol. 1709, pp. 1–68, Berlin: Springer, J. Azéma et al. ed., 1999.

[52] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.

Supplementary Information

S1 Reinforcement learning

In our learning model, individuals hold action values, $V_{i,t+1}(a_i)$, for each action $a_i \in \{C, D\}$. Adapting the model of ref. (50), the learning rule of individual i in our model is to update action values according to

$$V_{i,t+1}(a_i) = V_{i,t}(a_i) + \gamma_t \mathbb{1}(a_i, a_{i,t}) u_i(a_i, a_{-i,t}), \tag{S1.1}$$

where $\mathbb{1}(a_i, a_{i,t})$ is an indicator variable that equals 1 if $a_i = a_{i,t}$, and 0 otherwise, and $\gamma_t \in (0, 1)$ is a dynamic learning rate. This learning rate is decreasing as the game proceeds, which implies that the initial rounds of interaction are critical in determining the stable outcome of the learning process. Such a condition ensures that learning converges during an individual's lifetime (40, 51). This assumption can be justified by the fact that the situation (game) faced by the individuals is constant. Finally $u_i(a_i, a_{-i,t})$ is the utility to i if he plays a_i given that his opponent plays $a_{-i,t}$ at time t. Importantly, under the rule in eq. S1.1 an action that is not played at time t keeps the same action value at time $t + 1$, in contrast to traditional reinforcement learning, where actions that are not played for a long period of time tend to be forgotten, and hence their values decay to 0. This feature, together with the fact that early experience is important, means that our model captures relatively fast-scale learning dynamics that converges to an equilibrium outcome rather than lifelong learning processes. Thus our model of repeated interactions need not be understood as implying that an individual is paired with a single opponent for its entire lifespan. Rather, an individual may play several repeated games with different partners, and each of these repeated interactions is assumed to last long enough for learning to converge.

We then assume that individuals want to choose the action with highest value $V_{i,t}(a_i)$, but also have some tendency to explore the action with smaller value. A widely used choice rule to capture this principle is the logit-choice function,

$$p_{i,t}(a_i) = \frac{\exp[\xi V_{i,t}(a_i)]}{\sum_{b_i \in \mathcal{A}_i} \exp[\xi V_{i,t}(b_i)]}, \tag{S1.2}$$

where $\xi > 0$ is the exploration parameter (the inverse $1/\xi$ can be seen as the noise level if we interpret this model as perturbed maximization of action values, 52) in choosing actions. In our case, there are two actions, C and D, hence eq. S1.2 is a sigmoid function, which can be thought as a generalization of the threshold rule that chooses the action with greater value $V_{i,t}(a_i)$. Eq. S1.2 approaches such threshold function when ξ gets larger.

S2 Behavioral analysis

In this section, we show how we analyze behavioral interactions by focusing on one particular example, Realistic vs. Other-regard (a Sagemath notebook that contains the analysis of all behavioral interactions is available on demand).

Treating Realistic as player u and Other-regard as player v, one starts by verifying what behavioral equilibria of eq. 1 (Fig. S1) exist for this particular interaction (see Fig. S2 for the vector field of this interaction). Recall that the utilities of Realistic have the signs $u_{11} > 0$, $u_{12} < 0$, $u_{21} > 0$, $u_{22} = 0$. The utilities of Other-regard have the signs $v_{11} > 0$, $v_{12} > 0$, $v_{21} < 0$, $v_{22} < 0$. Consequently, the equilibria that do not exist are the two interior equilibria as well as $\left(0, \frac{v_{22}}{v_{12}+v_{22}}\right)$, $\left(\frac{u_{22}}{u_{12}+u_{22}}, 0\right)$, and $\left(1, \frac{v_{21}}{v_{11}+v_{21}}\right)$. For example, the latter equilibrium does not exist because $\frac{v_{21}}{v_{11}+v_{21}}$ is either negative (when $|v_{11}| > |v_{21}|$) or greater than 1 (when $|v_{11}| < |v_{21}|$), which is impossible for a probability.

We then calculate the Jacobian matrix associated to eq. 1, evaluate it at each equilibrium, and calculate its eigenvalues. The pure equilibria $(0,0)$, $(1,0)$, $(0,1)$, and $(1,1)$ are straightforward to analyze because the sign of the eigenvalues are opposite to the sign of the utilities of the players. For example, the equilibrium $(1,1)$ has the associated eigenvalues $(-\xi u_{11}, -\xi v_{11})$, which makes it locally stable. The equilibrium $(0,1)$ is also locally stable because it has the associated eigenvalues $(-\xi u_{21}, -\xi v_{12})$, which are both negative. The equilibria where one player is mixing require a little more work. The equilibrium $\left(\frac{u_{21}}{u_{11}+u_{21}}, 1\right)$ has eigenvalues

$$\left(-\xi \frac{u_{21}v_{11} + u_{11}v_{12}}{u_{11} + u_{21}}, \xi \frac{u_{11}u_{21}}{u_{11} + u_{21}}\right). \tag{S2.1}$$

Solving the inequality

$$-\xi \frac{u_{21}v_{11} + u_{11}v_{12}}{u_{11} + u_{21}} < 0. \tag{S2.2}$$

shows that this is always true given the signs of $\mathbf{u}$ and $\mathbf{v}$. The second eigenvalue

$$\xi \frac{u_{11}u_{21}}{u_{11} + u_{21}} \tag{S2.3}$$

is always positive, making the equilibrium $\left(\frac{u_{21}}{u_{11}+u_{21}}, 1\right)$ a saddle.

S3 Fitness computation

In this Appendix, we delineate the logic behind the computation of the fitnesses in Table 1 for the interaction between Realistic and Other-regard. The other fitnesses are computed similarly.

As we have proven in Appendix S2, the interaction between Realistic and Other-regard can lead to two possible behavioral equilibria, $(1, 1)$ or $(0, 1)$. Because the original learning dynamics is stochastic, it may reach either equilibrium, but we cannot determine which one analytically. Indeed, the theory of stochastic approximations does not provide precise predictions about the lock-in probabilities in local attractors. Hence, in the analysis of the model we treat the probability q_{RO} to get attracted in the mutual cooperation equilibrium, $(1, 1)$, as a free parameter. Hence the probability to reach the equilibrium $(0, 1)$ is $1 - q_{\text{RO}}$. The payoff to the players at equilibrium $(1, 1)$ is $b - c$ for both players. At equilibrium $(0, 1)$, Realistic obtains b and Other-regard $-c$. Weighting these payoffs by the lock-in probabilities gives

$$(f_{\text{RO}}, f_{\text{OR}}) = (q_{\text{RO}}(b - c) + (1 - q_{\text{RO}})b, q_{\text{RO}}(b - c) + (1 - q_{\text{RO}})(-c)), \tag{S3.1}$$

which are the entries of the corresponding fitnesses in Table 1 of the main text.

S4 Replicator dynamics and evolutionary equilibria

We evaluate the evolutionary competition between Realistic, Other-regard, Selfish, and Manipulator using the 4-dimensional replicator dynamics. In order to do so, we need to construct the evolutionary fitness matrix, i.e., computing the fitness, f_{uv}, of a strategy u when matched with another strategy v in the repeated game. We need to do so for all strategies $u, v \in \{\text{R}, \text{O}, \text{M}, \text{S}\}$, where we denote a strategy by its initial (i.e., R denotes Realistic, etc.). Having obtained such fitnesses, one can then call $\mathbf{x} = (x_{\mathrm{R}}, x_{\mathrm{O}}, x_{\mathrm{M}}, x_{\mathrm{S}})$ the vector of frequencies in the population ($x_{\mathrm{R}} + x_{\mathrm{O}} + x_{\mathrm{M}} + x_{\mathrm{S}} = 1$) and define the average fitness of type u when the population is in state $\mathbf{x}$ as

$$f_u(\mathbf{x}) = \sum_{v \in \{\mathrm{R,O,M,S}\}} x_v f_{uv}, \qquad u \in \{\mathrm{R, O, M, S}\}. \tag{S4.1}$$

We then use the continuous-time replicator dynamics to assess the long-term frequencies of strategies in the population, i.e.,

$$\dot{x}_u = x_u \left(f_u(\mathbf{x}) - \bar{f}(\mathbf{x}) \right), \qquad u \in \{\mathrm{R, O, M, S}\}, \tag{S4.2}$$

where

$$\bar{f}(\mathbf{x}) = \sum_{u \in \{\mathrm{R,O,M,S}\}} x_u f_u(\mathbf{x}), \tag{S4.3}$$

is the average fitness in the population at state $\mathbf{x}$.

Solving for the equilibria of eq. S4.2, we find that there is an equilibrium with Realistic, Other-regard, and Selfish, given by

$$\begin{cases} x_{\mathrm{R}} = \frac{c}{b q_{\mathrm{RO}}}, \\ x_{\mathrm{O}} = \frac{(b-c) q_{\mathrm{RR}}}{b q_{\mathrm{RO}}^2}, \\ x_{\mathrm{S}} = 1 - x_{\mathrm{R}} - x_{\mathrm{O}}. \end{cases} \tag{S4.4}$$

We find another 3-strategy equilibrium with Realistic, Manipulator, and Selfish, with the frequencies

$$\begin{cases} x_{\mathrm{R}} = \frac{c q_{\mathrm{RM}} u_{22}(-b u_{12} + c[u_{12} + u_{22}])}{(b-c)^2 (q_{\mathrm{RM}}^2 - q_{\mathrm{RR}})(u_{12}[u_{12} + u_{22}]) - bc q_{\mathrm{RM}}^2 u_{22}^2}, \\ x_{\mathrm{M}} = \frac{(b-c) c q_{\mathrm{RR}} u_{22}(u_{12} + u_{22})}{(b-c)^2 (q_{\mathrm{RM}}^2 - q_{\mathrm{RR}})(u_{12}[u_{12} + u_{22}]) - bc q_{\mathrm{RM}}^2 u_{22}^2}, \\ x_{\mathrm{S}} = 1 - x_{\mathrm{R}} - x_{\mathrm{M}}. \end{cases} \tag{S4.5}$$

There exists a 2-strategy equilibrium with Realistic and Other-regard who are in frequencies

$$\begin{cases} x_{\mathrm{R}} = \frac{-c(1 - q_{\mathrm{RO}})}{(b-c)(q_{\mathrm{RR}} - q_{\mathrm{RO}})}, \\ x_{\mathrm{O}} = 1 - x_{\mathrm{R}}. \end{cases} \tag{S4.6}$$

Selfish invades the latter mix of Realistic and Other-regard, and makes part of the equilibrium when

$$q_{\mathrm{RR}} < q_{\mathrm{RO}}^2 \qquad \text{and} \qquad \frac{b}{c} > \frac{q_{\mathrm{RO}} - q_{\mathrm{RR}}}{q_{\mathrm{RO}}^2 - q_{\mathrm{RR}}}, \tag{S4.7}$$

which is obtained by finding the conditions under which the fitness of Selfish at the equilibrium of eq. S4.6, $f_{\mathrm{S}}(\mathbf{x}^*_{\mathrm{RO}})$, is higher than the average fitness in the population, $\bar{f}(\mathbf{x}^*_{\mathrm{RO}})$. When this occurs the equilibrium consisting of Realistic, Other-regard, and Selfish (eq. S4.4) becomes stable.

Supplementary tables

Table S1: Classification of behavioral outcomes amongst the 4 strategies considered.

Interaction	**Stable equilibria**
R vs. R	$(1, 1)$ and $(0, 0)$
R vs. O	$(0, 1)$ and $(1, 1)$
R vs. M	$(1, 1)$ and $(0, \frac{v_{22}}{v_{12}+v_{22}})$
R vs. S	$(0, 0)$
O vs. O	$(1, 1)$
O vs. M	$(1, 0)$ and $(1, 1)$
O vs. S	$(1, 0)$
M vs. M	$(1, 1)^*$
M vs. S	$(\frac{u_{22}}{u_{12}+u_{22}}, 0)$
S vs. S	$(0, 0)$

*for $|u_{12}| < |u_{21}|$, another equilibrium, $(\frac{u_{22}}{u_{12}+u_{22}}, 0)$, becomes stable but we place ourselves in the conditions where this condition does not hold.

Supplementary figures

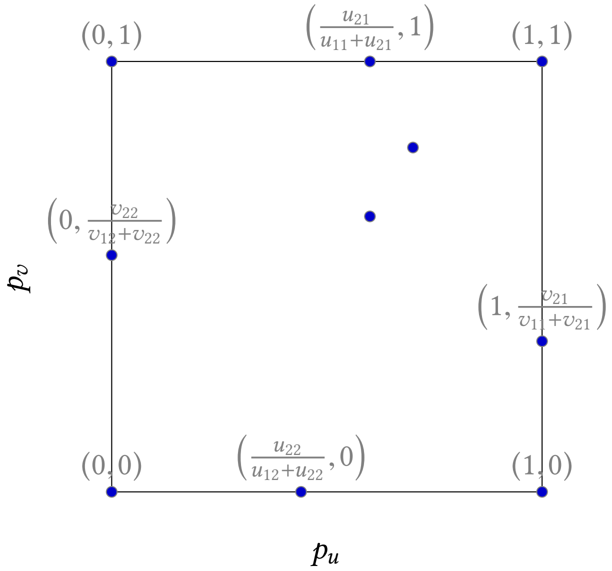

$(0,1)$
$\left(\frac{u_{21}}{u_{11}+u_{21}},1\right)$
$(1,1)$
$\left(0,\frac{v_{22}}{v_{12}+v_{22}}\right)$
p_v
$\left(1,\frac{v_{21}}{v_{11}+v_{21}}\right)$
$(0,0)$
$\left(\frac{u_{22}}{u_{12}+u_{22}},0\right)$
$(1,0)$
p_u

Figure S1: The ten generic behavioral equilibria in a 2×2 game between a player with utility function u, and his opponent with utility function v. The two interior equilibria have long expressions that are not shown here.

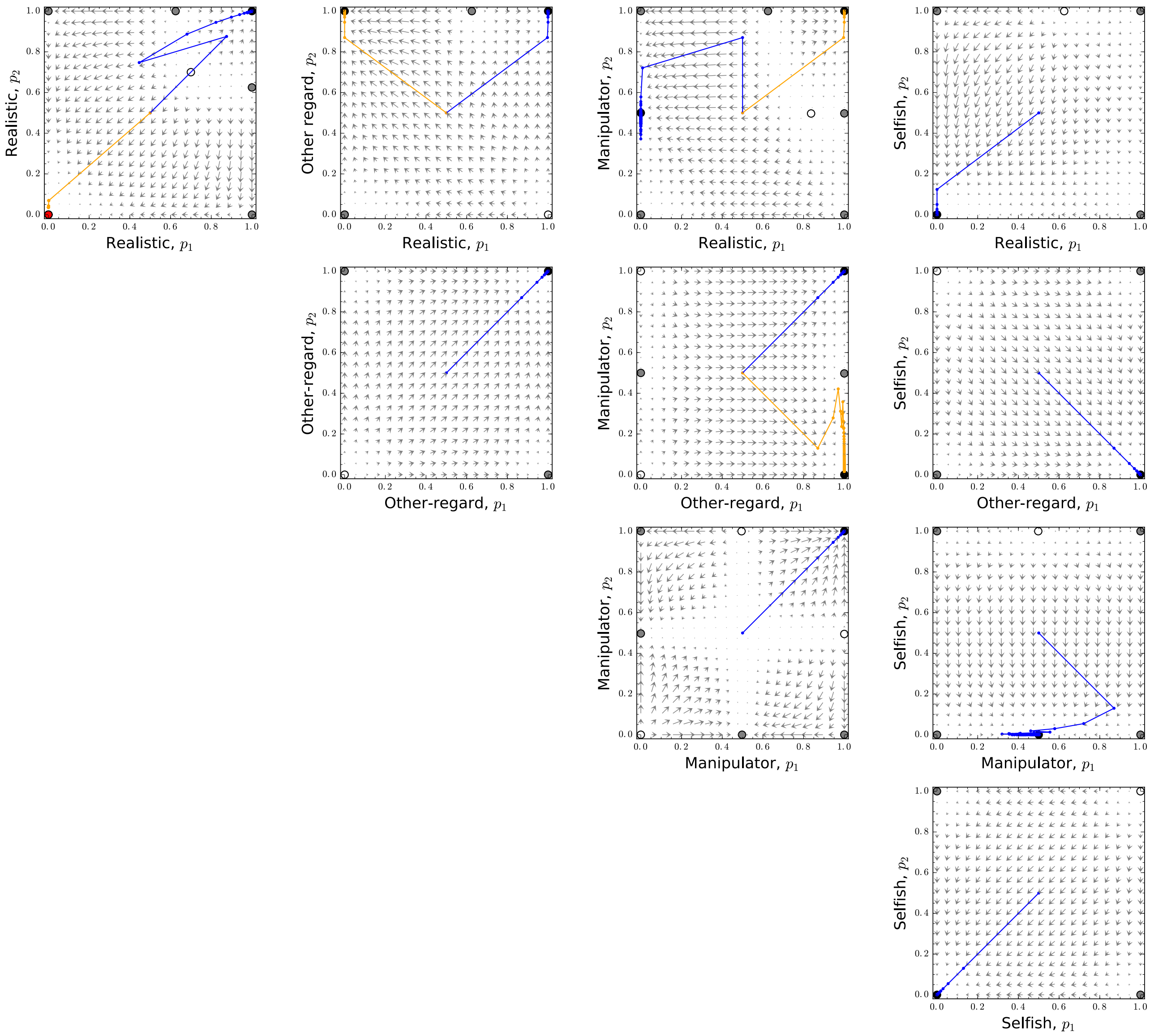


Figure S2: Vector fields (gray arrows) and stochastic trajectories (colored lines) for the 10 possible behavioral interactions between the 4 strategies Realistic, Other-regard, Manipulator, and Selfish. In each panel, on the x-axis is represented the probability that the row player cooperates (p_1), while on the y-axis, this is the probability that the column player cooperates (p_2). The stochastic trajectories are started from the center of the state space $(p_1,p_2) = (\frac{1}{2}, \frac{1}{2})$ and dots on it represent interaction rounds between the players. Circles represent equilibria: a white-filled circle is a source (both associated eigenvalues are positive); a gray-filled circle is a saddle (one positive and one negative associated eigenvalue); a black circle is a sink (both associated eigenvalues are negative). The red circle in the Realistic-Realistic interaction is a degenerate equilibrium with both zero eigenvalues, but it turns out to be locally stable. These plots were generated by setting positive utilities close to 1 and negative utilities close to -1 (which is why mixed equilibria always appear close to 0.5).

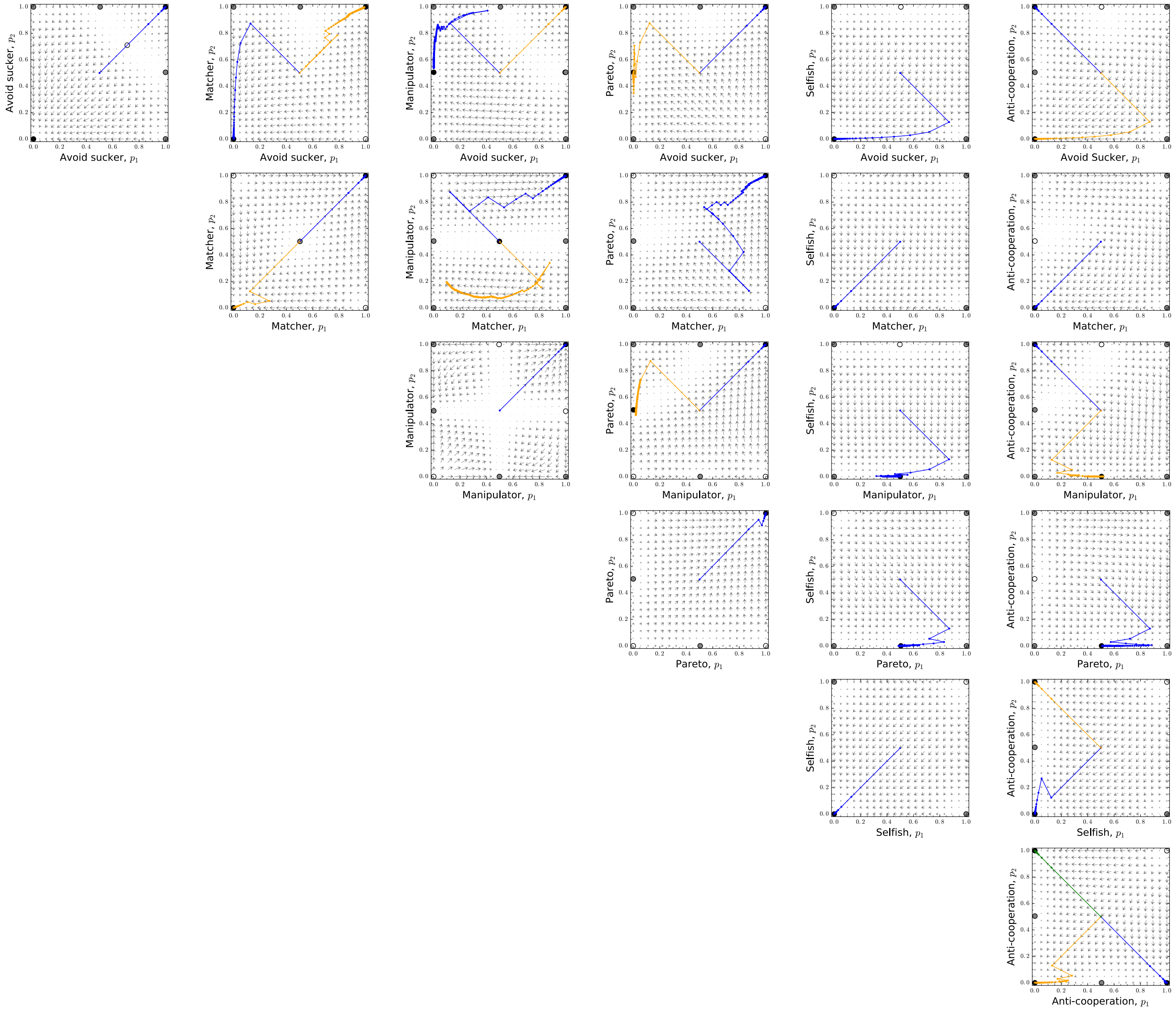
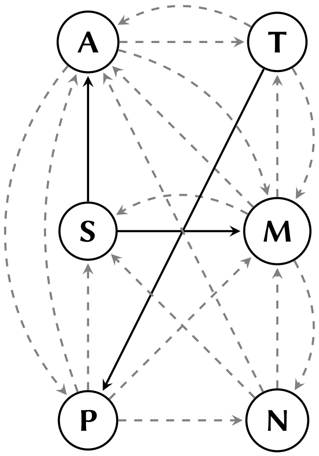


Figure S3: Vector fields (gray arrows) and stochastic trajectories (colored lines) for the 15 possible behavioral interactions between the 6 dominant strategies in the simulations: Avoid Sucker's Payoff, Matcher, Manipulator, Pareto, Selfish, and Anti-Cooperation. Otherwise similar to Fig. S2.



A → T:	$q_{\mathrm{TA}} > q_{\mathrm{AA}}$
A → P:	$q_{\mathrm{AA}} < q_{\mathrm{AP}}$ and $\left(\left(\frac{b}{c} \geq \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AP}}-q_{\mathrm{AA}}}\right) \text{ or } \left(\frac{b}{c} < \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AP}}-q_{\mathrm{AA}}} \text{ and } \frac{b}{c} + \frac{u^{\mathrm{P}}_{12}}{u^{\mathrm{P}}_{22}} > \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AP}}-q_{\mathrm{AA}}}\right)\right)$
A → M:	$q_{\mathrm{AA}} < q_{\mathrm{AM}}$ and $\left(\left(\frac{b}{c} \geq \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AM}}-q_{\mathrm{AA}}}\right) \text{ or } \left(\frac{b}{c} < \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AM}}-q_{\mathrm{AA}}} \text{ and } \frac{b}{c} + \frac{u^{\mathrm{M}}_{12}}{u^{\mathrm{M}}_{22}} > \frac{1-q_{\mathrm{AA}}}{q_{\mathrm{AM}}-q_{\mathrm{AA}}}\right)\right)$
T → A:	$q_{\mathrm{TA}} > q_{\mathrm{TT}}$
T → M:	True
M → A:	$\frac{u^{\mathrm{M}}_{22}}{u^{\mathrm{M}}_{12}+u^{\mathrm{M}}_{22}} b > b - c$
M → S:	$\frac{u^{\mathrm{M}}_{22}}{u^{\mathrm{M}}_{12}+u^{\mathrm{M}}_{22}} b > b - c$
M → T:	True
M → N:	True
P → A:	$\frac{u^{\mathrm{P}}_{22}}{u^{\mathrm{P}}_{12}+u^{\mathrm{P}}_{22}} b > b - c$
P → S:	$\frac{u^{\mathrm{P}}_{22}}{u^{\mathrm{P}}_{12}+u^{\mathrm{P}}_{22}} b > b - c$
P → M:	$\frac{u^{\mathrm{P}}_{22}}{u^{\mathrm{P}}_{12}+u^{\mathrm{P}}_{22}} b > b - c$
P → N:	$\frac{u^{\mathrm{P}}_{22}}{u^{\mathrm{P}}_{12}+u^{\mathrm{P}}_{22}} b > b - c$
N → A:	$\frac{b}{b-c} > \frac{1}{3q_{\mathrm{AN}}}$
N → S:	$\frac{b}{b-c} > \frac{1}{3}$
N → M:	True

Figure S4: Invasion diagram amongst the successful strategy classes in the evolutionary simulations with associated invasion conditions (similar to Fig. 1B). When we write "True", the invasion condition exists but is too long to be displayed here (a Mathematica notebook containing these conditions is available on demand). Name legend: A="Avoid Sucker's Payoff"; T="Matcher"; P="Pareto"; N="Anti-cooperation"; S="Selfish"; M="Manipulator".

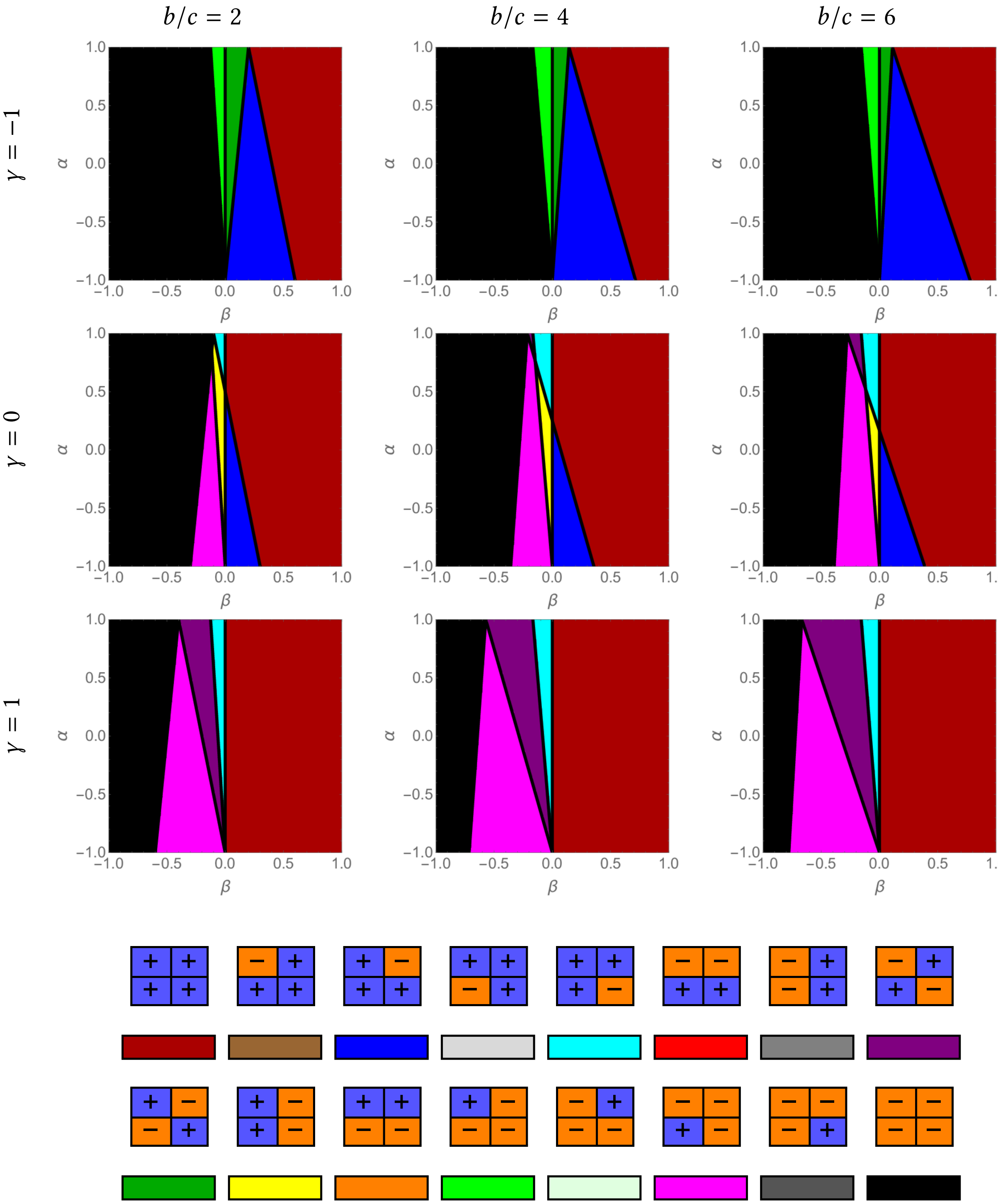


Figure S5: Phenotype space for the model where the utility depends explictly on material payoffs (eq. 4 with $k = 2$) as a function of β (x-axis), α (y-axis), γ (rows), and b/c (columns). Under the panels depicting the phenotype space, we show the color code of strategies (colors are under the corresponding utility matrix).

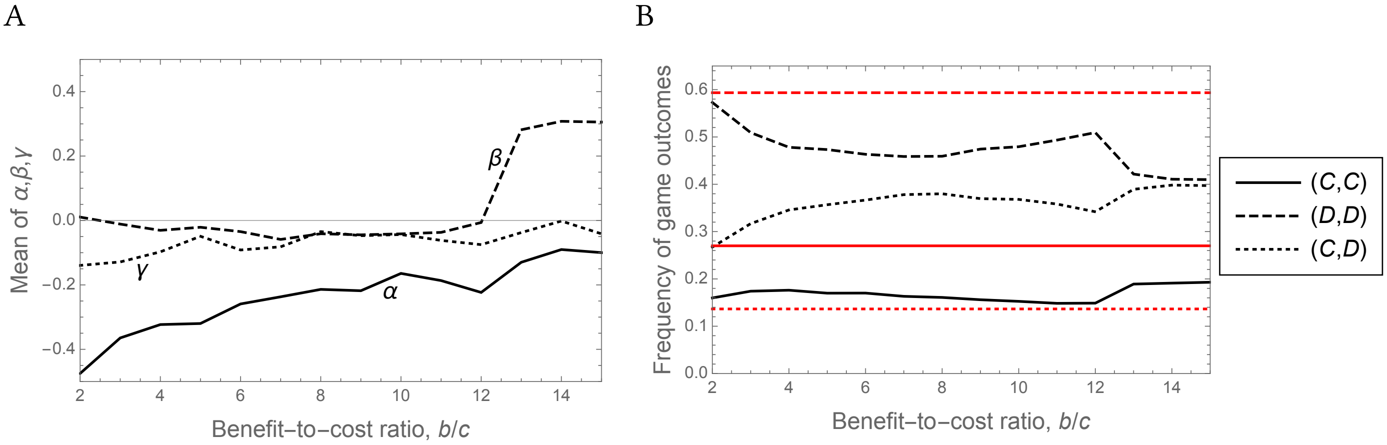


Figure S6: Results for the model where the utility depends explictly on material payoffs (eq. 4) as a function of the benefit-to-cost ratio, b/c (identical to Fig. 4 except that $k = 5$). (A) Time average of α, β, γ in a simulation run. (B) Time average of the frequency of game outcomes in a simulation run (similar to Fig. 3B and 4B).

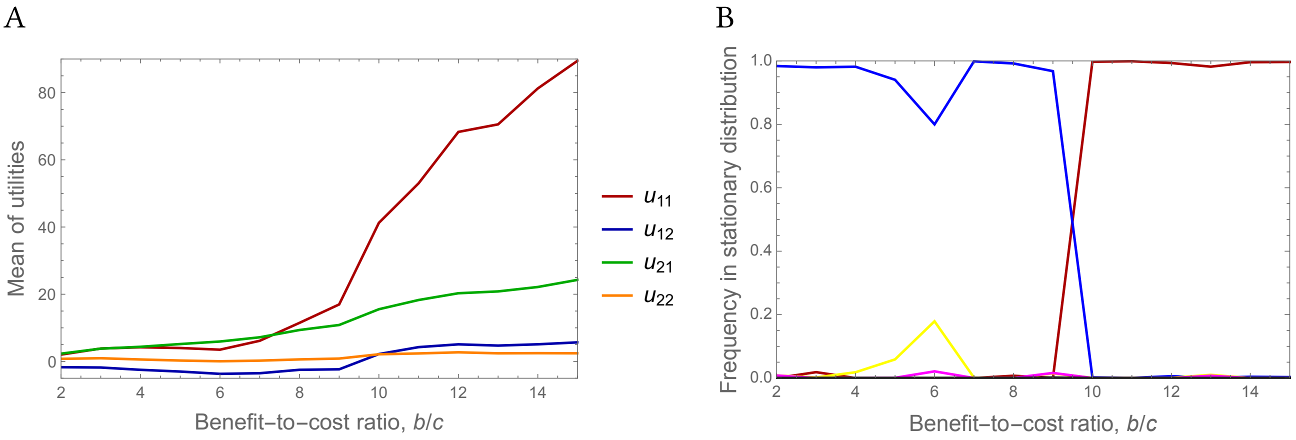


Figure S7: Additional results for the model where the utility depends explictly on material payoffs (eq. 4 with $k = 2$) as a function of the benefit-to-cost ratio, b/c. (A) Time average in a simulation run of the four utilities associated to each game outcome. (B) Proportion of time a simulation run spends in each strategy class (similar to Fig. 3A; see Fig. S5 for color code of strategies).

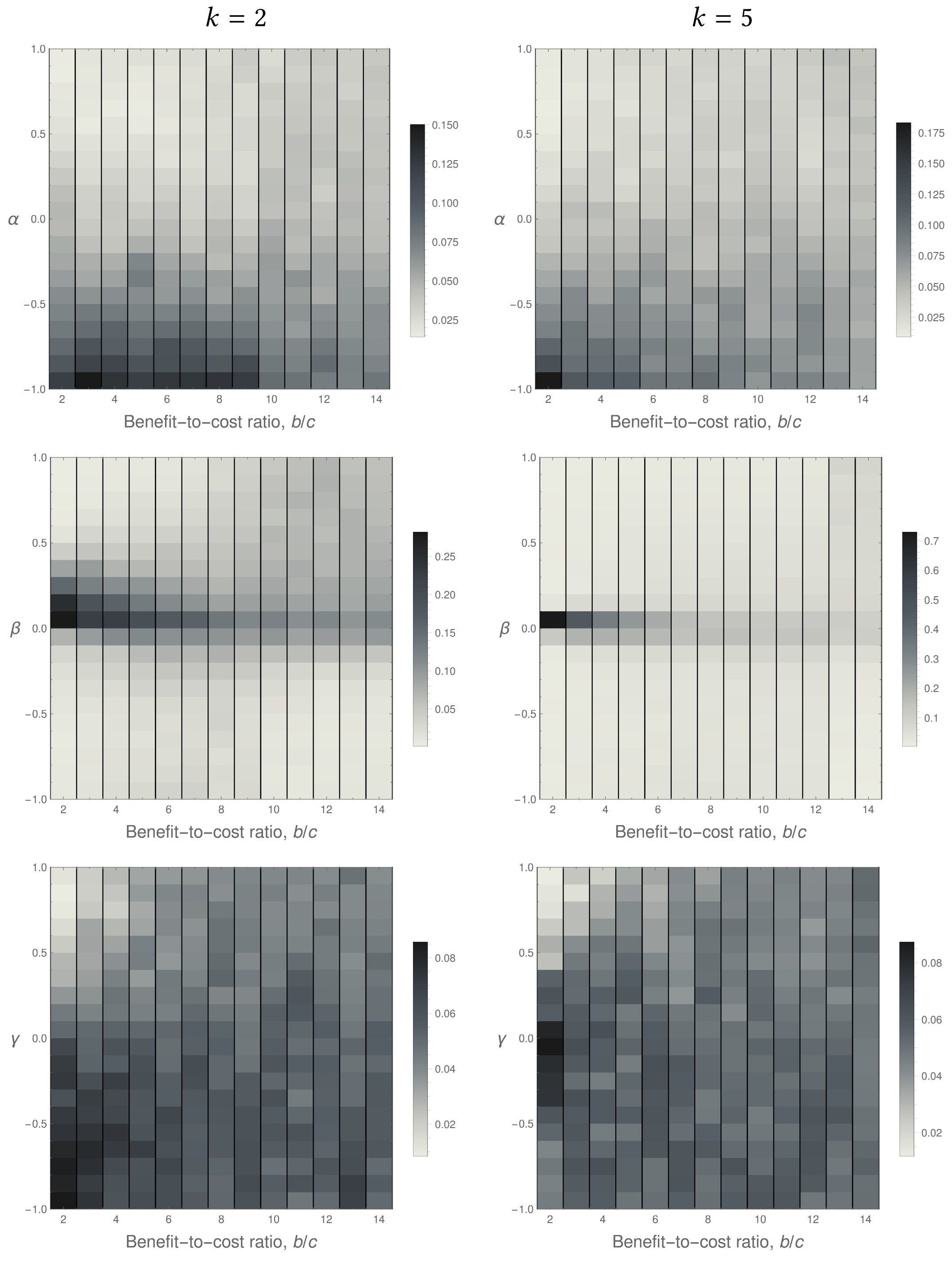


Figure S8: Stationary distributions of the traits measuring additive other-regard (α), multiplicative other-regard (β), and inequity aversion (γ) for various benefit-to-cost ratios (b/c) and values of k (see main text for definition, eq. 4) when constraining the utility function to explicitly depend on the material payoffs. In each subfigure, a column of values for a given benefit-to-cost ratio represents the stationary distribution of the trait, where darker shading corresponds to a higher frequency.